Hyperbolic Image-Text Representations

Karan Desai [1] **Maximilian Nickel** [2] **Tanmay Rajpurohit** [3] **Justin Johnson** [1 2] **Ramakrishna Vedantam** [4]

Abstract

Visual and linguistic concepts naturally organize themselves in a hierarchy, where a textual concept "dog" entails all images that contain dogs. Despite being intuitive, current large-scale vision and language models such as CLIP (Radford et al., 2021) do not explicitly capture such hierarchy. We propose MERU, a contrastive model that yields hyperbolic representations of images and text. Hyperbolic spaces have suitable geometric properties to embed tree-like data, so MERU can better capture the underlying hierarchy in image-text datasets. Our results show that MERU learns a highly interpretable and structured representation space while being competitive with CLIP's performance on standard multi-modal tasks like image classification and image-text retrieval.

Figure 1. **Hyperbolic image-text representations. Left:** Images and text depict *concepts* and can be jointly viewed in a *visual-semantic hierarchy*, wherein text *'exhausted doggo'* is more generic than an image (which might have more details like a cat or snow). Our method MERU embeds images and text in a hyperbolic space that is well-suited to embed tree-like data. **Right:** Representation manifolds of CLIP (*hypersphere*) and MERU (*hyperboloid*) illustrated in 3D. MERU assumes the origin to represent the *most generic concept*, and embeds text closer to the origin than images.

1. Introduction

Visual-semantic hierarchy. It is commonly said that *'an image is worth a thousand words'* – consequently, images contain a lot more information than the sentences which typically describe them. For example, given the middle image in Figure 1 one might describe it as *'a cat and a dog playing in the street'* or with a less specific sentence like *'exhausted doggo'* or *'so cute <3'*. These are not merely diverse descriptions but contain varying levels of detail about the underlying semantic contents of the image.

As humans, we can reason about the relative detail in each caption, and can organize such concepts into a meaningful visual-semantic hierarchy (Vendrov et al., 2016), namely, *'exhausted doggo'* $\rightarrow$ *'a cat and a dog playing in the street'* $\rightarrow$ (Figure 1 middle image). Providing multimodal models access to this inductive bias about vision and language has the potential to improve generalization (Radford et al., 2021), interpretability (Selvaraju et al., 2017) and enable better exploratory data analysis of large-scale datasets (Radford et al., 2021; Schuhmann et al., 2022).

Vision-language representation learning. Approaches such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have catalyzed a lot of recent progress in computer vision by showing that Transformer-based (Vaswani et al., 2017) models trained using large amounts of image-text data from the internet can yield transferable representations, and such models can perform *zero-shot* recognition and retrieval using natural language queries. All these models represent images and text as vectors in a high-dimensional Euclidean, affine space and normalize the embeddings to unit L^2 norm. However, such a choice of geometry can find it hard to capture the visual-semantic hierarchy.

An affine Euclidean space treats all embedded points in the same manner, with the same distance metric being applied to all points (Murphy, 2013). Conceptually, this can cause issues when modeling hierarchies – a *generic* concept (closer to the *root node* of the hierarchy) is close to many other concepts compared to a *specific* concept (which is only close to its immediate neighbors). Thus, a Euclidean space can find it hard to pack all the images that say a generic

KD and Rama did this work while at Meta. [1]University of Michigan [2]Meta AI [3]Independent Researcher [4]New York University. Correspondence to: Karan Desai <`kdexd@umich.edu`>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

concept *'curious kitty'* should be close to while also respecting the embedding structure for *'a cat and a dog playing on the street'*. Such issues are handled naturally by hyperbolic spaces – the volume increases exponentially as we move away from the origin (Lee, 2019), making them a continuous relaxation of trees. This allows a generic concept (*'cat'*) to have many neighbors by placing it close to the origin (Nickel & Kiela, 2017), and more specific concepts further away. Thus, distinct specific concepts like images in Figure 1 can be far away from each other while being close to some generic concept (*'animal'*).

Hyperbolic representations with MERU. In this work, we train the first large-scale contrastive image-text models that yield hyperbolic representations (Nickel & Kiela, 2017) – MERU [1] that captures the visual-semantic hierarchy (Figure 1). Our method conceptually resembles current state-of-the-art contrastive methods (Jia et al., 2021; Radford et al., 2021). Importantly the hierarchy *emerges* in the representation space, given access only to image-text pairs during training such models.

Practically, MERU confers multiple benefits such as (a) better performance on image retrieval and classification tasks, (b) more efficient usage of the embedding space, making it suited for resource-constrained, on-device scenarios, (c) an interpretable representation space that allows one to infer the relative semantic specificity of images and text. Overall, we summarize our contributions as follows:

- We introduce MERU, the first implementation of deep hyperbolic representations we are aware of, training ViTs (Dosovitskiy et al., 2021) with 12M image-text pairs.
- We provide a strong CLIP baseline that outperforms previous re-implementations (Mu et al., 2022) at comparable data scale, and systematically demonstrate the benefits of hyperbolic representations over this baseline on *zero-shot* retrieval and classification, and effectiveness for small embedding dimensions (Kusupati et al., 2022).
- We perform thorough qualitative analysis with MERU to demonstrate its potential for exploratory data analysis of large-scale multimodal datasets.

2. Preliminaries

We briefly review Riemannian manifolds (Section 2.1) and essential concepts of hyperbolic geometry (Section 2.2). For a more thorough treatment of the topic, we refer the reader to textbooks by Ratcliffe (2006) and Lee (2019).

[1]Meru is a mountain that symbolizes the *center of all physical, metaphysical, and spiritual universes* in Eastern religions like Hinduism and Buddhism. Our method is named MERU because the origin of the hyperboloid entails everything and plays a more vital role than in Euclidean (or generally, affine) spaces. See also: *Mount Semeru, Indonesia* (Sources – `wikipedia.org/wiki/Mount_Meru` and `wikipedia.org/wiki/Semeru`)

2.1. Riemannian manifolds

A *smooth surface* is a two-dimensional sheet which is *locally Euclidean* – every point on the surface has a local neighborhood which can be mapped to $\mathbb{R}^2$ via a differentiable and invertible function. *Smooth manifolds* extend the notion of smooth surfaces to higher dimensions.

A *Riemannian manifold* $(\mathcal{M}, g)$ is a smooth manifold $\mathcal{M}$ equipped with a *Riemannian metric* g. The metric g is a collection of inner product functions $g_{\mathbf{x}}$ for all points $\mathbf{x} \in \mathcal{M}$, and varies smoothly over the manifold. At any point $\mathbf{x}$, the inner product $g_{\mathbf{x}}$ is defined in the *tangent space* $\mathcal{T}_{\mathbf{x}}\mathcal{M}$, which is a Euclidean space that gives a linear approximation of $\mathcal{M}$ at $\mathbf{x}$. Euclidean space $\mathbb{R}^n$ is also a Riemannian manifold, where g is the standard Euclidean inner product.

Our main topic of interest is hyperbolic spaces, which are Riemannian manifolds with *constant negative curvature*. They are fundamentally different from Euclidean spaces that are *flat* (zero curvature). A hyperbolic manifold of n dimensions cannot be represented with $\mathbb{R}^n$ in a way that preserves both distances and angles. There are five popular models of hyperbolic geometry that either represent n-dimensional hyperbolic spaces either in $\mathbb{R}^n$ while distorting distances and/or angles (e.g. Poincaré ball model), or as a sub-manifold of $\mathbb{R}^{n+1}$ (e.g. the Lorentz model).

2.2. Lorentz model of hyperbolic geometry

We use the Lorentz model of hyperbolic geometry for developing MERU. This model represents a hyperbolic space of n dimensions on the upper half of a two-sheeted hyperboloid in $\mathbb{R}^{n+1}$. See Figure 1 for an illustration of $\mathcal{L}^2$ in $\mathbb{R}^3$. Hyperbolic geometry has a direct connection to the study of special relativity theory (Einstein, 1905; Einstein et al., 2015). We borrow some of its terminology in our discussion – we refer to the hyperboloid's axis of symmetry as *time dimension* and all other axes as *space dimensions* (Minkowski, 1908). Every vector $\mathbf{x} \in \mathbb{R}^{n+1}$ can be written as $[\mathbf{x}_{space}, x_{time}]$, where $\mathbf{x}_{space} \in \mathbb{R}^n$ and $x_{time} \in \mathbb{R}$.

Definition. Let $\langle \cdot, \cdot \rangle$ is Euclidean inner product and $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ denote the *Lorentzian inner product* that is induced by the Riemannian metric of the Lorentz model. For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}$, it is computed as follows:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = \langle \mathbf{x}_{space}, \mathbf{y}_{space} \rangle - x_{time}\, y_{time} \tag{1}$$

The induced *Lorentzian norm* is $\|\mathbf{x}\|_{\mathcal{L}} = \sqrt{|\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}}|}$. The Lorentz model possessing a constant curvature $-c$ is defined as a following set of vectors:

$$\mathcal{L}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -{}^1/_c\ ,\ c > 0\} \tag{2}$$

All vectors in this set satisfy the following constraint:

$$x_{time} = \sqrt{{}^1/_c + \|\mathbf{x}_{space}\|^2} \tag{3}$$

Geodesics. A *geodesic* is the shortest path between two points on the manifold. Geodesics in the Lorentz model are curves traced by the intersection of the hyperboloid with hyperplanes passing through the origin of $\mathbb{R}^{n+1}$. The *Lorentzian distance* between two points $\mathbf{x}, \mathbf{y} \in \mathcal{L}^n$ is:

$$d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \sqrt{1/c} \cdot \cosh^{-1}(-c \, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}) \tag{4}$$

Tangent space. The tangent space at some point $\mathbf{z} \in \mathcal{L}^n$ is a Euclidean space of vectors that are orthogonal to $\mathbf{z}$ according to the Lorentzian inner product:

$$\mathcal{T}_{\mathbf{z}}\mathcal{L}^n = \{\mathbf{v} \in \mathbb{R}^{n+1} : \langle \mathbf{z}, \mathbf{v} \rangle_{\mathcal{L}} = 0\} \tag{5}$$

Any vector in ambient space $\mathbf{u} \in \mathbb{R}^{n+1}$ can be projected to the tangent space $\mathcal{T}_{\mathbf{z}}\mathcal{L}^n$ via an orthogonal projection:

$$\mathbf{v} = \mathrm{proj}_{\mathbf{z}}(\mathbf{u}) = \mathbf{u} + c \, \mathbf{z} \, \langle \mathbf{z}, \mathbf{u} \rangle_{\mathcal{L}} \tag{6}$$

Exponential and logarithmic maps. The *exponential map* provides a way to map vectors from tangent spaces onto the manifold. For a point $\mathbf{z}$ on the hyperboloid, it is defined as $\mathrm{expm}_{\mathbf{z}} : \mathcal{T}_{\mathbf{z}}\mathcal{L}^n \to \mathcal{L}^n$ with the expression:

$$\mathbf{x} = \mathrm{expm}_{\mathbf{z}}(\mathbf{v}) = \cosh(\sqrt{c} \, \|\mathbf{v}\|_{\mathcal{L}}) \, \mathbf{z} + \frac{\sinh(\sqrt{c} \, \|\mathbf{v}\|_{\mathcal{L}})}{\sqrt{c} \, \|\mathbf{v}\|_{\mathcal{L}}} \, \mathbf{v} \tag{7}$$

Intuitively the exponential map shows how $\mathcal{T}_x\mathcal{L}^n$ *folds* on the manifold. Its inverse is the *logarithmic map* ($\mathrm{logm}_{\mathbf{z}} : \mathcal{L}^n \to \mathcal{T}_{\mathbf{z}}\mathcal{L}^n$), that maps $\mathbf{x}$ from the hyperboloid back to $\mathbf{v}$ in the tangent space:

$$\mathbf{v} = \mathrm{logm}_{\mathbf{z}}(\mathbf{x}) = \frac{\cosh^{-1}(-c \, \langle \mathbf{z}, \mathbf{x} \rangle_{\mathcal{L}})}{\sqrt{(c \, \langle \mathbf{z}, \mathbf{x} \rangle_{\mathcal{L}})^2 - 1}} \, \mathrm{proj}_{\mathbf{z}}(\mathbf{x}) \tag{8}$$

For our approach, we will only consider these maps where $\mathbf{z}$ is the origin of the hyperboloid ($\mathbf{O} = [\mathbf{0}, \sqrt{1/c}]$).

3. Approach

In this section, we discuss the modeling pipeline and learning objectives of MERU to learn hyperbolic representations of images and text. We use the tools of hyperbolic geometry introduced in Section 2 throughout our discussion.

Our model design is inspired by CLIP (Radford et al., 2021) due to its simplicity and scalability. As shown in Figure 2, we process images and text using two separate encoders, and obtain embedding vectors of a fixed dimension n. Beyond this, there are two crucial design choices: (1) transferring embeddings from Euclidean space to the Lorentz hyperboloid, and (2) designing suitable training objectives that induce semantics and structure in the representation space.

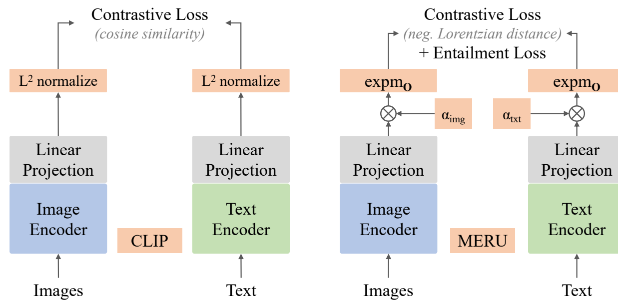


Figure 2. **MERU model design:** MERU comprises similar architectural components as standard image-text contrastive models like CLIP. While CLIP projects the embeddings to a unit hypersphere, MERU lifts them onto the Lorentz hyperboloid using the exponential map. The contrastive loss uses the negative of Lorentzian distance as a similarity metric, and a special entailment loss enforces *'text entails image'* partial order in the representation space.

Lifting embeddings onto the hyperboloid. Let the embedding vector from the image encoder or text encoder, after linear projection be $\mathbf{v}_{enc} \in \mathbb{R}^n$. We need to apply a transformation such that the resulting vector $\mathbf{x}$ lies on the Lorentz hyperboloid $\mathcal{L}^n$ in $\mathbb{R}^{n+1}$. Let the vector $\mathbf{v} = [\mathbf{v}_{enc}, 0] \in \mathbb{R}^{n+1}$. We observe that $\mathbf{v}$ belongs to the tangent space at the hyperboloid origin $\mathbf{O}$, as Eqn. 5 is satisfied: $\langle \mathbf{O}, \mathbf{v} \rangle_{\mathcal{L}} = 0$. Thus, we parameterize *only* the *space* components of the Lorentz model ($\mathbf{v}_{enc} = \mathbf{v}_{space}$). Due to such parameterization, we can simplify the exponential map from Eqn. 7 by writing only *space* components:

$$\mathbf{x}_{space} = \cosh(\sqrt{c} \, \|\mathbf{v}\|_{\mathcal{L}})\mathbf{0} + \frac{\sinh(\sqrt{c} \, \|\mathbf{v}\|_{\mathcal{L}})}{\sqrt{c} \, \|\mathbf{v}\|_{\mathcal{L}}} \mathbf{v}_{space}$$

The first term reduces to $\mathbf{0}$. Moreover, the Lorentzian norm of $\mathbf{v}$ simplifies to the Euclidean norm of *space* components: $\|\mathbf{v}\|^2_{\mathcal{L}} = \langle \mathbf{v}, \mathbf{v} \rangle_{\mathcal{L}} = \langle \mathbf{v}_{space}, \mathbf{v}_{space} \rangle - 0 = \|\mathbf{v}_{space}\|^2$. This substitution simplifies the above equation as follows:

$$\mathbf{x}_{space} = \frac{sinh(\sqrt{c} \, \|\mathbf{v}_{space}\|)}{\sqrt{c} \, \|\mathbf{v}_{space}\|} \mathbf{v}_{space} \tag{9}$$

The corresponding *time* component x_{time} can be computed from $\mathbf{x}_{space}$ using Eqn. 3, the resulting $\mathbf{x}$ *always* lies on the hyperboloid. This eliminates the need for an orthogonal projection (Eqn. 6) and simplifies the exponential map. Our parameterization is simpler than previous work which parameterizes vectors in full ambient space $\mathbb{R}^{n+1}$ (Law et al., 2019; Le et al., 2019; Nickel & Kiela, 2018).

Preventing numerical overflow. The exponential map scales $\mathbf{v}_{space}$ using an exponential operator. According to CLIP-style weight initialization, $\mathbf{v}_{space} \in \mathbb{R}^n$ would have an expected norm $= \sqrt{n}$. After exponential map, it becomes $e^{\sqrt{n}}$, which can be numerically large (*e.g.,* $n = 512$ and $c = 1$ gives $||\mathbf{x}_{space}|| \approx 6.7 \times 10^{10}$).

To fix this issue, we *scale* all vectors $\mathbf{v}_{space}$ in a batch before applying $\text{expm}_{\mathbf{O}}$ using two learnable scalars α_{img} and α_{txt}. These are initialized to $\sqrt{1/n}$ so that the Euclidean embeddings have an expected unit norm at initialization. We learn these scalars in logarithmic space to avoid collapsing all embeddings to zero. After training, they can be absorbed into the preceding projection layers.

Learning structured embeddings. Having lifted standard Euclidean embeddings onto the hyperboloid, we next discuss the losses used to enforce structure and semantics in representations learned by MERU. Recall that our motivation is to capture the visual-semantic hierarchy (Figure 1) to better inform the generalization capabilities of vision-language models. For this, an important desideratum is a meaningful notion of distance between semantically similar text and image pairs. We also want to induce a partial order between text and images as per the visual-semantic hierarchy to have better interpretability. We do this with a modified version of an entailment loss proposed by Le et al. (2019), that works for arbitrary hyperboloid curvatures $-c$.

3.1. Contrastive learning formulation

Given a batch of size B of image-text pairs and any j^{th} instance in batch, its image embedding $\mathbf{y}_j$ and text embedding $\mathbf{x}_j$ form a *positive* pair, whereas the remaining $B-1$ text embeddings in the batch $\mathbf{x}_i (i \neq j)$ form *negative* pairs.

In contrastive learning, we compute the negative Lorentzian distance as a similarity measure (Eqn. 4) for all B pairs in the batch. These logits are divided by a temperature τ and apply a softmax operator. Similarly, we also consider a contrastive loss for text, that treats images as negatives. The total loss $\mathcal{L}_{cont}$ is the average of these two losses computed for every image-text pair in the batch. Our implementation of the contrastive loss is the same as the multi-class N-pair loss from (Sohn, 2016) used in CLIP (Radford et al., 2021) with the crucial difference being that we compute distances on the hyperboloid instead of cosine similarity.

3.2. Entailment loss

In addition to the contrastive loss, we adapt an entailment loss (Ganea et al., 2018; Le et al., 2019) to enforce partial order relationships between paired text and images. Ganea et al. (2018) is more different from ours since they parameterize their representations according to the Poincaré ball model. Le et al. (2019) use this loss with a fixed $c = 1$, which we extend to handle arbitrary, learned curvatures.

Refer Figure 3 for an illustration in two dimensions. Let $\mathbf{x}$ and $\mathbf{y}$ denote the text and image embeddings of a single image-text pair. Note that the encoders only give $\mathbf{x}_{space}$ and $\mathbf{y}_{space}$ according to our parameterization. Corresponding x_{time} and y_{time} are calculated using Eqn. 3. We define an

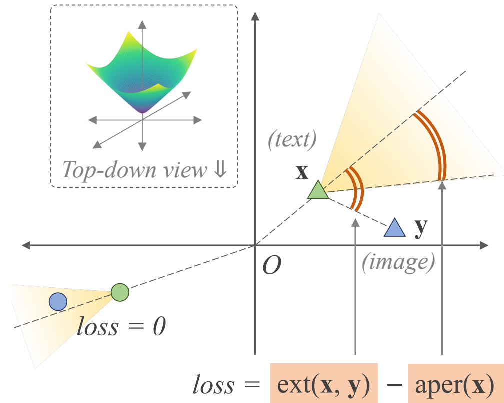


Figure 3. **Entailment loss (illustrated for $\mathcal{L}^2$):** This loss pushes image embedding $\mathbf{y}$ inside an imaginary cone projected by the paired text embedding $\mathbf{x}$, and is implemented as the difference of exterior angle $\angle O\mathbf{x}\mathbf{y}$ and half aperture of the cone. Loss is zero if the image embedding is already inside the cone *(left quadrant)*.

entailment cone for each $\mathbf{x}$, which narrows as we go farther from the origin. This cone is defined by the half-aperture:

$$\text{aper}(\mathbf{x}) = \sin^{-1}\left(\frac{2K}{\sqrt{c}\,\|\mathbf{x}_{space}\|}\right) \tag{10}$$

where a constant $K = 0.1$ is used for setting boundary conditions near the origin. We now aim to identify and penalize when the paired image embedding $\mathbf{y}$ lies outside the entailment cone. For this, we measure the exterior angle $\text{ext}(\mathbf{x}, \mathbf{y}) = \pi - \angle \mathbf{O}\mathbf{x}\mathbf{y}$ as shown in Figure 3:

$$\text{ext}(\mathbf{x}, \mathbf{y}) = \cos^{-1}\left(\frac{y_{time} + x_{time}\; c\,\langle \mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}}}{\|\mathbf{x}_{space}\|\sqrt{(c\,\langle \mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}})^2 - 1}}\right) \tag{11}$$

If the exterior angle is smaller than the aperture, then the partial order relation between $\mathbf{x}$ and $\mathbf{y}$ is already satisfied and we need not penalize anything, while if the angle is greater, we need to reduce it. This is captured by the following loss function (written below for a single $\mathbf{x}$, $\mathbf{y}$ pair):

$$\mathcal{L}_{entail}(\mathbf{x}, \mathbf{y}) = \max(0,\ \text{ext}(\mathbf{x}, \mathbf{y}) - \text{aper}(\mathbf{x})) \tag{12}$$

We provide exact derivations of the above equations for half-aperture and exterior angle in Appendix A. Overall, our total loss is $\mathcal{L}_{cont} + \lambda\mathcal{L}_{entail}$ averaged over each minibatch.

4. Experiments

Our main objective in the experiments is to establish the competitiveness of hyperbolic representations of MERU as compared to Euclidean representations obtained from CLIP-style models. To this end, we train models using large amounts of image-text pairs and transfer them to a variety of image classification and retrieval tasks.

4.1. Training details

Baselines. We primarily compare with CLIP (Radford et al., 2021), that embeds images and text on a unit hypersphere in a Euclidean space. CLIP was trained using a private dataset of 400M image-text pairs. Several follow-up works re-implement CLIP and use publicly accessible datasets like YFCC (Thomee et al., 2016), Conceptual Captions (Changpinyo et al., 2021; Sharma et al., 2018), and LAION (Schuhmann et al., 2021; 2022); notable examples are OpenCLIP (Ilharco et al., 2021), SLIP (Mu et al., 2022), DeCLIP (Li et al., 2022), and FILIP (Yao et al., 2022). We develop our CLIP baseline and train it using a *single* public dataset – RedCaps (Desai et al., 2021) – for easier reproducibility. Our smallest model trains using $8\times$ V100 GPUs in *less than one day* and significantly outperforms recent CLIP re-implementations that use YFCC (Mu et al., 2022).

Refer Appendix B for details about our CLIP baseline. Our implementation is based on PyTorch (Paszke et al., 2019) and `timm` (Wightman, 2019) libraries.

Models. We use the Vision Transformer (Dosovitskiy et al., 2021) as image encoder, considering three models of varying capacity – ViT-S (Chen et al., 2021; Touvron et al., 2021), ViT-B, and ViT-L. All use a patch size of 16. The text encoder is same as CLIP – a 12-layer, 512 dimensions wide Transformer (Vaswani et al., 2017) language model. We use the same byte-pair encoding tokenizer (Sennrich et al., 2016) as CLIP, and truncate input text at maximum 77 tokens.

Data augmentation. We randomly crop 50–100% area of images and resize them to 224×224, following (Mu et al., 2022). For text augmentation, we randomly *prefix* the subreddit names to captions as '`{subreddit} : {caption}`'.

Initialization. We initialize image/text encoders in the same style as CLIP, except for one change: we use a *sine-cosine* position embedding in ViT, like (Chen et al., 2021; He et al., 2022), and keep it frozen while training. We initialize the softmax temperature as $\tau = 0.07$ and clamp it to a minimum value of 0.01. For MERU, we initialize the learnable projection scalars $\alpha_{img} = \alpha_{txt} = {}^{1}/\sqrt{512}$, the curvature parameter $c = 1.0$ and clamp it in $[0.1, 10.0]$ to prevent training instability. All scalars are learned in logarithmic space as $log(1/\tau)$, $log(c)$, and $log(\alpha)$.

Optimization. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.2 and $(\beta_1, \beta_2) = (0.9, 0.98)$. We disable weight decay for all gains, biases, and learnable scalars. All models are trained for $120K$ iterations with batch size 2048 (≈ 20 epochs). The maximum learning rate is 5×10^{-4}, increased linearly for the first $4K$ iterations, followed by cosine decay to zero (Loshchilov & Hutter, 2016). We use mixed precision (Micikevicius et al., 2018) to accelerate training, except computing exponential map and losses for MERU in FP32 precision for numerical stability.

Table 1. **Zero-shot image and text retrieval.** Best performance in every column is highlighted in green. MERU performs better than CLIP for both datasets and across all model sizes.

		text → image				*image → text*			
		COCO		Flickr		COCO		Flickr	
		R5	R10	R5	R10	R5	R10	R5	R10
ViT S/16	CLIP	29.9	40.1	35.3	46.1	37.5	48.1	42.1	54.7
	MERU	**30.5**	**40.9**	**37.1**	**47.4**	**39.0**	**50.5**	**43.5**	**55.2**
ViT B/16	CLIP	32.9	43.3	40.3	51.0	41.4	52.7	**50.2**	**60.2**
	MERU	**33.2**	**44.0**	**41.1**	**51.6**	**41.8**	**52.9**	48.1	58.9
ViT L/16	CLIP	31.7	42.2	39.0	49.3	40.6	51.3	47.8	58.5
	MERU	**32.6**	**43.0**	**39.6**	**50.3**	**41.9**	**53.3**	**50.3**	**60.6**

Loss multiplier (λ) for MERU. We set $\lambda = 0.2$ by running a hyperparameter sweep with ViT-B/16 models for one epoch. Some $\lambda > 0$ is necessary to induce partial order structure, however, quantitative performance is less sensitive to the choice of $\lambda \in [0.01, 0.3]$; Higher values of λ strongly regularize against the contrastive loss and hurt performance.

4.2. Image and text retrieval

CLIP-style contrastive models perform image and text retrieval within batch during training, making them ideal for retrieval-related downstream applications. We evaluate the retrieval capabilities of MERU as compared to CLIP on two established benchmarks: COCO and Flickr30K (Chen et al., 2015; Young et al., 2014), that comprise 5000 and 1000 images respectively and five captions per image. COCO evaluation uses the `val2017` split while Flickr30K uses the `test` split defined by Karpathy & Fei-Fei (2015). We perform *zero-shot transfer*, without any additional training using these datasets. We *squeeze* images to 224×224 pixels before processing them through the image encoder.

Inference with MERU. We rank a pool of candidate image/-text embeddings for retrieval in decreasing order of their Lorentzian inner product (Eqn. 1) with a text/image query embedding. Some transfer tasks like *open-vocabulary detection* (Gu et al., 2022; Zareian et al., 2021) may require calibrated scores, for them we recommend using the training procedure – compute the negative of distance (Eqn. 4), divide by temperature and apply a softmax classifier.

Results. Table 1 reports recall@{5,10} of MERU and the reproduced CLIP baselines on these benchmarks. Hyperbolic representations of MERU mostly perform best for all tasks and models (except Flickr30K text retrieval with ViT-B/16). This is encouraging evidence that hyperbolic spaces have suitable geometric properties to learn strong representations for retrieval applications. Surprisingly, increasing model size (ViT-B/16 → ViT-L/16) does not improve image retrieval for both, MERU and CLIP. We believe that better quality of text queries is important for image retrieval – increasing the size of text encoder can alleviate this issue.

Table 2. **Zero-shot image classification.** We train MERU and CLIP models with varying parameter counts and transfer them *zero-shot* to 20 image classification datasets. Best performance in every column is highlighted in green. Hyperbolic representations from MERU match or outperform CLIP on 13 out of the first 16 datasets. On the last four datasets (gray columns), both MERU and CLIP have *near-random* performance, as concepts in these datasets are not adequately covered in the training data.

		ImageNet	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Country211	MNIST	CLEVR	PCAM	SST2
ViT S/16	CLIP	34.3	74.5	**60.1**	24.4	**33.8**	27.5	**11.3**	**1.4**	15.0	**73.7**	63.9	47.0	88.2	18.6	31.4	**5.2**	10.0	19.4	50.2	50.1
	MERU	**34.4**	**75.6**	52.0	**24.7**	33.7	**28.0**	11.1	1.3	**16.2**	72.3	**64.1**	**49.2**	**91.1**	**30.4**	**32.0**	4.8	7.5	14.5	51.0	50.0
ViT B/16	CLIP	**37.9**	**78.9**	65.5	**33.4**	33.3	29.8	**14.4**	1.4	17.0	77.9	**68.5**	50.9	92.2	25.6	31.0	**5.8**	10.4	14.3	54.1	51.5
	MERU	37.5	78.8	**67.7**	32.7	**34.8**	**30.9**	14.0	**1.7**	**17.2**	**79.3**	**68.5**	**52.1**	**92.5**	**30.2**	**34.5**	5.6	13.0	13.5	49.8	49.9
ViT L/16	CLIP	38.4	80.3	**72.0**	**36.4**	36.3	32.0	**18.0**	1.1	16.5	78.8	**68.3**	48.6	**93.7**	26.7	35.4	6.1	14.8	13.6	51.2	51.1
	MERU	**38.8**	**80.6**	68.7	35.5	**37.2**	**33.0**	16.6	**2.2**	**17.2**	**80.0**	67.5	**52.1**	**93.7**	**28.1**	**36.5**	**6.2**	11.8	13.1	52.7	49.3

4.3. Image classification

Learning from language supervision allows CLIP to perform *zero-shot* image classification, wherein one may specify label sets as text queries (Elhoseiny et al., 2013) instead of using pre-defined ontologies (Deng et al., 2009; Miller, 1992). Classifier weights are obtained by embedding label-based queries (also called *prompts*) using the text encoder.

In this section, we evaluate MERU on 20 image classification benchmarks covering a wide variety of visual concepts. These are used by Radford et al. (2021) and several follow-up works (Li et al., 2022; Mu et al., 2022; Yao et al., 2022), and available with open-source libraries like `tensorflow-datasets` and `torchvision` [2]. We report top-1 mean per-class accuracy for all datasets to account for any label imbalance. We use multiple prompts per dataset, most of which follow Radford et al. (2021). We *ensemble* these multiple prompts by averaging their embeddings before lifting them onto the hyperboloid (Eqn. 9). See Tables 6 and 8 in Appendix for details about datasets and prompts.

Results. Table 2 shows strong transfer performance of MERU, matching or outperforming CLIP on 13 out of 16 standard datasets. While MERU is effective on recall-based measures (Table 1), it does not come at the expense of precision (Murphy, 2013). Overall, hyperbolic representations from MERU are competitive with their Euclidean counterparts across varying model architectures (ViT-S/B/L).

All models have *near-random* performance on four benchmarks. Concepts in these datasets have low coverage in RedCaps, like PCAM (Veeling et al., 2018) containing medical scans, or SST2 (Socher et al., 2013) containing movie reviews rendered as images. Performance on these benchmarks does not indicate the efficacy of our RedCaps-trained models; using larger training datasets like LAION (Schuhmann et al., 2022) may yield meaningful trends.

[2] `tensorflow.org/datasets` and `pytorch.org/vision`

		Embedding width				
		512	256	128	96	64
COCO *text→image*	CLIP	31.7	31.8	31.4	29.6	25.7
	MERU	**32.6**	**32.7**	**32.7**	**31.0**	**26.5**
COCO *image→text*	CLIP	40.6	41.0	40.4	37.9	33.3
	MERU	**41.9**	**42.5**	**42.6**	**40.5**	**34.2**
ImageNet	CLIP	38.4	38.3	37.9	35.2	30.2
	MERU	**38.8**	**38.8**	**38.8**	**37.3**	**32.3**

Table 3. **MERU and CLIP with different embedding widths.** We report *zero-shot* COCO recall@5 and ImageNet top-1 accuracy. MERU outperforms CLIP at lower embedding widths.

4.4. Resource-constrained deployment

We hypothesize that embeddings that capture a rich visual-semantic hierarchy can use the volume in the representation space more efficiently. This is useful for on-device deployments with runtime or memory constraints that necessitate low-dimensional embeddings (Kusupati et al., 2022).

To verify this hypothesis, we train MERU and CLIP models that output 64–512 dimensions wide embeddings. We initialize the encoders from ViT-L/16 models (Table 2, last two rows) to reduce compute requirements, keep them frozen, and re-initialize projection layers and learnable scalars. We train for $30K$ iterations and evaluate on *zero-shot* COCO retrieval and ImageNet (Russakovsky et al., 2014) classification. Results in Table 3 show that MERU consistently performs better at low embedding widths. This indicates that hyperbolic embeddings may be an appealing solution for resource-constrained on-device applications.

4.5. Ablations

In this section, we ablate our MERU models to observe the impact of our design choices. We experiment with two image encoders, ViT-B/16 and ViT-L/16, and evaluate for zero-shot COCO retrieval and ImageNet classification.

Table 4. **MERU ablations.** We ablate three design choices of MERU and report *zero-shot* COCO recall@5 and ImageNet top-1 accuracy. Our design choices are crucial for training stability when using a larger model (ViT-L/16) with MERU.

	COCO *text→image*	COCO *image→text*	ImageNet
MERU ViT-B/16	33.2	41.8	37.5
1. *no entailment loss*	33.7	43.5	36.2
2. *fixed* $c = 1$	33.2	42.1	37.9
3. $\langle\cdot,\cdot\rangle_{\mathcal{L}}$ *in contrastive*	32.6	42.3	37.3
MERU ViT-L/16	32.6	41.9	38.8
1. *no entailment loss*	32.7	42.2	33.8
2. *fixed* $c = 1$	0.9	0.9	0.7
3. $\langle\cdot,\cdot\rangle_{\mathcal{L}}$ *in contrastive*	–	*did not converge*	–

Specifically, we train three ablations with the default hyperparameters (Section 4.1), except having one difference each. Results are shown in Table 4 above.

No entailment loss: We only use the contrastive loss for training this ablation. This effectively means setting $\lambda = 0$. Note that this ablation is mathematically impossible for CLIP as there is no obvious notion of entailment that can be defined when all the embeddings have a unit norm. Disabling the entailment loss is mostly inconsequential to MERU's performance. This shows that choosing a hyperbolic space is sufficient to improve *quantitative* performance over CLIP. Entailment loss is crucial for better structure and interpretability, as will be discussed in Section 5.

Fixed curvature parameter: Recall that our models treat the hyperboloid curvature as a learnable parameter during training. Here we train an ablation using a fixed curvature $c = 1$. This has negligible impact on MERU ViT-B/16, but learning curvature is crucial when scaling model size – MERU ViT-L/16 model with fixed $c = 1$ is difficult to optimize and performs poorly on convergence. As far as we are aware, no prior work learns the curvature (Atigh et al., 2022; Khrulkov et al., 2020; Nickel & Kiela, 2018).

Lorentzian inner product in contrastive loss: CLIP-style contrastive loss uses the inner product defined on the hypersphere (cosine similarity). Similarly, we consider the *Lorentzian inner product* (Eqn. 1) in the contrastive loss instead of negative Lorentzian distance. With this, MERU ViT-L/16 is difficult to train. Loss diverges due to numerical overflow, as Lorentzian inner product is numerically large and unbounded in $(-\infty, -{}^1\!/\!_c]$, unlike cosine similarity $\in [-1, 1]$. Lorentzian distance applies a logarithmic operator ($cosh^{-1}$) on the Lorentzian inner product, slowing down its growth and hence improving numerical stability.

We hope these ablations serve as guidelines for work in other domains that study hyperbolic representation learning.

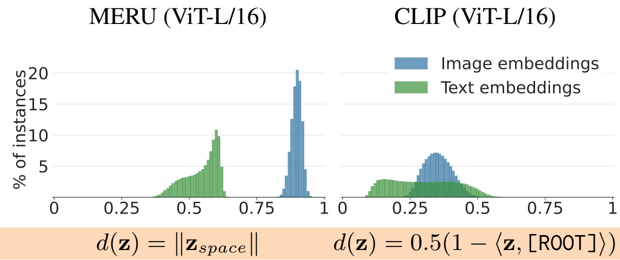


Figure 4. **Distribution of embedding distances from `[ROOT]`:** We embed all 12M training images and text using trained MERU and CLIP. Note that precise distance is not necessary for this analysis, so we compute simple monotonic transformations of distances, $d(\mathbf{z})$. MERU embeds text closer to `[ROOT]` than images.

5. Qualitative analysis

In this section, we probe our trained models to infer the visual-semantic hierarchy captured by MERU and CLIP. Apriori we hypothesize that MERU is better equipped to capture this hierarchy due to the geometric properties of hyperbolic spaces and an entailment loss that enforces the partial-order relationship *'text entails image'*. All our analysis in this section uses ViT-L/16 models.

Preliminary: `[ROOT]` embedding. Recall Figure 1 – if we think of the visual-semantic hierarchy as a tree, then its *leaf nodes* are images and the *intermediate nodes* are text descriptions with varying *semantic specificity*. Naturally, the *root node* should represent the *most generic concept*. We denote its embedding in the representation space as `[ROOT]`.

For MERU, `[ROOT]` is the origin of the Lorentz hyperboloid as it entails the entire representation space. The location of `[ROOT]` for CLIP is not as intuitive – the notion of entailment is mathematically not defined, and the origin does not lie on the hypersphere. We empirically estimate CLIP's `[ROOT]` as an embedding vector that has the least distance from all embeddings of the training dataset. Hence, we average all 2×12M embeddings of images and text in RedCaps, followed by L^2 normalization. `[ROOT]` will be different for different CLIP models, whereas it is fixed for MERU.

Embedding distances from `[ROOT]`. In a representation space that effectively captures the visual-semantic hierarchy, text embeddings should lie closer to `[ROOT]` than image embeddings, since text is more *generic* than images (Figure 1). Figure 4 shows the distribution of embedding distances from `[ROOT]` – these distributions overlap for CLIP but are separated for MERU. The range of distributions in Figure 4 (left) hints that MERU embeds text and images in two *concentric, high-dimensional rings* around `[ROOT]`. The *ring* of text is more *spread out*, whereas the ring of images is relatively *thin*. This resembles the structure of the visual-semantic hierarchy – images only occupy *leaf nodes* whereas text occupies many intermediate nodes.

MERU	CLIP
a bengal cat sitting beside wheatgrass on a white surface	*a bengal cat sitting beside wheatgrass on a white surface*
bengal	↓
cat	↓
domestic	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
white horse	*white horse*
equine	↓
equestrian	↓
beauty	↓
female	↓
fluffy	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
photography of rainbow during cloudy sky	*phenomenon*
rainbow	↓
phenomenon	↓
rural	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
retro photo camera on table	↓
fujinomiya	↓
vintage	↓
style	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
avocado toast	*avocado toast*
healthy breakfast	*delicious*
delicious	↓
homemade	↓
fresh	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
brooklyn bridge	*photo of brooklyn bridge, new york*
new york city	*new york city*
city	*new york*
outdoors	↓
day	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
taj mahal	*taj mahal through an arch*
monument	*travel*
architecture	*inspiration*
travel	↓
day	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
sydney opera house	*sydney opera house*
opera house	*opera house*
holiday	*gift*
day	*beauty*
`[ROOT]`	`[ROOT]`

Figure 5. **Image traversals with MERU and CLIP.** We perform text retrieval at multiple steps while traversing from an image embedding to `[ROOT]`. Overall, CLIP retrieves fewer textual concepts (top row), but in some cases it reveals a coarse hierarchy (bottom row). MERU captures hierarchy with significantly greater detail, we observe that: (1) Text becomes more *generic* we move towards `[ROOT]`, *e.g.*, *white horse → equestrian* and *retro photo camera → vintage*. (2) MERU has higher recall of concepts than CLIP, like words in bottom row: *homemade, city, monument*. (3) MERU also shows systematic text→image entailment, *e.g.*, *day* entails many images captured in daylight.

Image traversals. In a discrete tree, one can discover the *ancestors* of any node by performing shortest-path traversal to the *root node* (Dijkstra, 1959). We perform such traversals for images with MERU and CLIP. If the representation space has captured the visual-semantic hierarchy, then a shortest-path traversal from an image to `[ROOT]` should let us infer textual concepts that describe the image with varying levels of abstraction. We briefly describe this analysis here, refer Appendix D for more details.

We traverse from an image and `[ROOT]` by interpolating 50 equally spaced steps along the geodesic connecting their embedding vectors. We use every interpolated step embedding as a query to perform retrieve the nearest neighbor from a set of text embeddings $\mathcal{X}$, that also include `[ROOT]`.

We display results with 60 randomly selected images collected from `pexels.com`, a website that offers freely usable stock photos. We use two different sets $\mathcal{X}$ having text sourced from: (1) 750 captions obtained using the image metadata from `pexels.com`, and (2) 8.7M captions from the YFCC dataset (Thomee et al., 2016).

Figure 5 shows results with 8 selected images and captions from `pexels.com`. Appendix D includes results with 52 other images and with YFCC captions *without cherry-picking*. CLIP seems to capture hierarchy to some extent, often retrieving very few (or zero) captions between image and `[ROOT]`. MERU captures it with much finer granularity, retrieving concepts that gradually become more *generic* as we move closer to `[ROOT]`.

6. Related work

Visual-language representation learning. Soon after the initial success of deep learning on ImageNet (Krizhevsky et al., 2012), deep metric learning (Sohn, 2016; Song et al., 2015) was used to learn vision-language representations in a shared semantic space (Frome et al., 2013; Karpathy & Fei-Fei, 2015). The motivations at the time included the possibility of improving vision models (Frome et al., 2013), enabling zero-shot learning by expressing novel categories as sentences (Elhoseiny et al., 2013; Frome et al., 2013), and better image-text retrieval (Karpathy & Fei-Fei, 2015; Young et al., 2014). Another line of work proposed learning visual models from language supervision via objectives like textual n-gram prediction (Li et al., 2017), or *generative* objectives like masked language modeling (Bulent Sariyildiz et al., 2020) or image captioning (Desai & Johnson, 2021).

More recent approaches like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) use contrastive metric learning to pre-train Vision Transformers (Dosovitskiy et al., 2021) and have helped to better realize the motivations of the earlier works in practice. While all prior works learn Euclidean embeddings, MERU explicitly works in the hyperbolic space that is conceptually better for embedding the visual-semantic hierarchy (Figure 1) underlying images and text. Our results (Section 4) demonstrate that MERU yields strong performance as prior works, and also offers better interpretability to the representation space.

Entailment embeddings. In a vision and language context, Order Embeddings (Vendrov et al., 2016) propose capturing the partial order between language and vision by enforcing that text embeddings $\mathbf{x}$ and image embeddings $\mathbf{y}$, should satisfy $\mathbf{y} \leq \mathbf{x}$ for all dimensions i. While enforcing order is useful for retrieval, in our initial experiments, we found that distance-based contrastive learning to be crucial for better performance on classification and retrieval. Thus, we focus on adapting the currently successful contrastive learning and add our entailment objective in conjunction, to obtain the desired structure in the representation space.

For NLP and knowledge graph embedding applications, several approaches embed partially ordered data (Bai et al., 2021; Dasgupta et al., 2020; Ganea et al., 2018; Nguyen et al., 2017; Vilnis et al., 2018) or discover ordering from pairwise similarities (Le et al., 2019; Nickel & Kiela, 2017; Tifrea et al., 2018). Our work has a flavor of both these lines of work, since we impose structure *across* modalities, but order also emerges *within* modality (Figure 5).

Hyperbolic representations in computer vision. Khrulkov et al. (2020) learn hyperbolic image embeddings using image-label pairs, while Atigh et al. (2022) study image segmentation by utilizing hyperbolic geometry. More recently, Ermolov et al. (2022) and Ge et al. (2023) extend standard contrastive self-supervised learning framework (He et al., 2020; Wu et al., 2018) in vision to learn hyperbolic representations. In contrast to all these works, MERU learns multimodal representations with an order of magnitude more data and shows strong *zero-shot* transfer abilities across generic artificial intelligence tasks (Radford et al., 2021).

7. Conclusion

In this paper, we learn large-scale image-text representations (MERU) to capture the visual-semantic hierarchy underlying images and text. Our key innovation is to bring advances in learning hyperbolic representations to practical, large-scale deep learning applications. MERU is competitive or more performant than approaches that learn Euclidean representations (like CLIP). It does so along with capturing hierarchical knowledge which allows one to make powerful inferences such as reasoning about images at different levels of abstraction. Beyond this, our model also provides clear performance gains for small embedding dimensions (which are useful in resource-constrained settings). We hope this work catalyzes progress in learning useful representations from large amounts of unstructured data.

Future work. In this *scaling* era, we are seeing rapid progress with large multi-modal models trained using millions (or even billions) of image-text pairs. The quality and concept distribution of training data plays a vital role in the efficacy of these models. Such training data is becoming increasingly opaque and black-box due to its unprecedented scale. We believe that the time is ripe to revisit the unreasonable effectiveness of data in deep learning (Halevy et al., 2009; Sun et al., 2017). Modeling hierarchies can help uncover higher-order relationships beyond basic data statistics. As a concrete example, Figure 1 “so cute <3” is an extremely generic caption and does not the *precise* details in images. Such captions add noisy supervision in contrastive loss by making false negative pairs with many images in the batch. Image traversals with MERU Figure 5 can discover such noisy captions. ML practitioners can filter or re-caption such training images to improve dataset quality and train subsequent models for improved performance.

Limitations. Our work is not without limitations. While MERU yields hyperbolic representations that excel at zero-shot retrieval and image classification tasks, the linear probe evaluations in Table 7 show that the underlying Euclidean representations from the image encoder of MERU underperform CLIP. Exploring MERU’s transferability to other tasks that involve few-shot learning or full-model fine-tuning is also beyond scope of this paper. Finally, while we provide ample qualitative analysis of image traversals, future work can propose more systematic ways to evaluate the hierarchical knowledge captured by vision-language models.

References

Atigh, M. G., Schoep, J., Acar, E., van Noord, N., and Mettes, P. Hyperbolic Image Segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 9

Bai, Y., Ying, R., Ren, H., and Leskovec, J. Modeling Heterogeneous Hierarchies with Relation-specific Hyperbolic Cones. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 9

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. The Pushshift Reddit Dataset. *arXiv preprint arXiv:2001.08435*, 2020. 16

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – Mining Discriminative Components with Random Forests. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014. 16, 17

Bulent Sariyildiz, M., Perez, J., and Larlus, D. Learning visual representations with caption annotations. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 9

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

Chen, T., Xu, B., Zhang, C., and Guestrin, C. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016. 16

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2021. 5

Cheng, G., Han, J., and Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 2017. 17

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing Textures in the Wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 17

Coates, A., Ng, A., and Lee, H. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. 17

Dasgupta, S. S., Boratko, M., Zhang, D., Vilnis, L., Li, X. L., and McCallum, A. Improving Local Identifiability in Probabilistic Box Embeddings. *arXiv preprint arXiv:2010.04831*, 2020. 9

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6

Desai, K. and Johnson, J. VirTex: Learning Visual Representations from Textual Annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 9

Desai, K., Kaul, G., Aysola, Z., and Johnson, J. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021. 5, 16, 19

Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische mathematik*, 1959. 8

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised Visual Representation Learning by Context Prediction. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015. 17

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2, 5, 9

Einstein, A. Zur Elektrodynamik bewegter Körper. *Annalen der physik*, 1905. 2

Einstein, A., Lorentz, H. A., Minkowski, H., and Weyl, H. *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity*. Martino Fine Books, 2nd edition, 2015. 2

El Banani, M., Desai, K., and Johnson, J. Learning Visual Representations via Language-Guided Sampling. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 17

Elhoseiny, M., Saleh, B., and Elgammal, A. Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2013. doi: 10.1109/ICCV.2013.321. 6, 9, 17

Ermolov, A., Mirvakhabova, L., Khrulkov, V., Sebe, N., and Oseledets, I. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 9

Fei-Fei, L., Fergus, R., and Perona, P. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *CVPR Workshop*, 2004. 17

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., and Mikolov, T. Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2013. 9

Fürst, A., Rumetshofer, E., Lehner, J., Tran, V. T., Tang, F., Ramsauer, H., Kreil, D., Kopp, M., Klambauer, G., Bitto, A., et al. Cloob: Modern hopfield networks with infoloob outperform clip. 2022. 17

Ganea, O.-E., Bécigneul, G., and Hofmann, T. Hyperbolic Entailment Cones for Learning Hierarchical Embeddings. *arXiv preprint arXiv:1804.01882*, 2018. 4, 9, 15

Ge, S., Mishra, S., Kornblith, S., Li, C.-L., and Jacobs, D. Hyperbolic Contrastive Learning for Visual Representations beyond Objects. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 9

Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 5

Halevy, A., Norvig, P., and Pereira, F. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 2009. URL http://www.computer.org/portal/cms_docs_intelligent/intelligent/homepage/2009/x2exp.pdf. 9

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 9

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

Helber, P., Bischke, B., Dengel, A. R., and Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 17

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303. 19

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. OpenCLIP, 2021. URL https://doi.org/10.5281/zenodo.5143773. 5

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1, 2, 9

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. B. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 17

Karpathy, A. and Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5, 9

Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., and Lempitsky, V. S. Hyperbolic Image Embeddings. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7, 9

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 17

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3D Object Representations for Fine-Grained Categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, 2013. 17

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. 17

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 9

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S. M., Jain, P., and Farhadi, A. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 6

Law, M. T., Liao, R., Snell, J., and Zemel, R. S. Lorentzian Distance Learning for Hyperbolic Representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. 3

Le, M., Roller, S., Papaxanthos, L., Kiela, D., and Nickel, M. Inferring Concept Hierarchies from Text Corpora via Hyperbolic Embeddings. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2019. 3, 4, 9

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2010. 17

Lee, J. M. *Introduction to Riemannian Manifolds*. Graduate Texts in Mathematics. Springer International Publishing, 2019. ISBN 9783319917542. URL https://books.google.com/books?id=UIPltQEACAAJ. 2, 15

Li, A., Jabri, A., Joulin, A., and van der Maaten, L. Learning visual n-grams from web data. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 9

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 5, 6, 17

Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 1989. 17

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 19

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 5

Maji, S., Rahtu, E., Kannala, J., Blaschko, M. B., and Vedaldi, A. Fine-Grained Visual Classification of Aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 17

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. Mixed precision training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 5, 16

Miller, G. A. WordNet: A Lexical Database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York*, 1992. 6

Minkowski, H. Raum und Zeit. *Physikalische Zeitschrift*, 1908. 2

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 6, 16

Murphy, K. P. *Machine learning : a probabilistic perspective*. MIT Press, 2013. 1, 6

Nguyen, K. A., Köper, M., im Walde, S. S., and Vu, N. T. Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 9

Nickel, M. and Kiela, D. Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 9

Nickel, M. and Kiela, D. Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 3, 7

Nilsback, M.-E. and Zisserman, A. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*, 2008. 17

Noroozi, M. and Favaro, P. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 17

Parkhi, O., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and Dogs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 16, 17

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5, 19

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011. 17

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1, 2, 3, 4, 5, 6, 9, 16, 17, 18, 19

Ratcliffe, J. G. *Foundations of Hyperbolic Manifolds*. Graduate Texts in Mathematics. Springer New York, 2006. ISBN 9780387331973. URL `https://books.google.com/books?id=JV9m8o-ok6YC`. 2

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2014. 6

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 1, 5, 6

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

Sennrich, R., Haddow, B., and Birch, A. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2016. 5

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2018. 5

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013. 6, 17

Sohn, K. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 4, 9

Song, H. O., Xiang, Y., Jegelka, S., and Savarese, S. Deep Metric Learning via Lifted Structured Feature Embedding. *arXiv preprint arXiv:1511.06452*, 2015. 9

Speer, R. ftfy, 2019. URL `https://doi.org/10.5281/zenodo.2591652`. 19

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 9

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 2016. 5, 8, 16

Tifrea, A., Bécigneul, G., and Ganea, O. Poincaré GloVe: Hyperbolic Word Embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 9

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 5

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. CNNs for Digital Pathology. *arXiv preprint arXiv:1806.03962*, 2018. 6, 17

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. Order-embeddings of images and language. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 1, 9

Vilnis, L., Li, X. L., Murty, S., and McCallum, A. Probabilistic Embedding of Knowledge Graphs with Box Lattice Measures. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2018. 9

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. J. The Caltech-UCSD Birds-200-2011 Dataset. 2011. 17

Wightman, R. PyTorch Image Models. `https://github.com/rwightman/pytorch-image-models`, 2019. 5

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 9

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 17

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 5, 6

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2014. 5, 9

Zareian, A., Rosa, K. D., Hu, D. H., and Chang, S.-F. Open-vocabulary object detection using captions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., and Houlsby, N. A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 17

Zhang, R., Isola, P., and Efros, A. A. Colorful Image Colorization. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 17

Acknowledgments

We thank our wonderful colleagues for helpful discussions and feedback on the paper (*in alphabetical order*, grouped by their affiliation during the undertaking of this project):

– *Meta:* Léon Bottou, Kamalika Chaudhuri, Ricky Chen, Piotr Dollár, Surya Ganguli, Rohit Girdhar, Naman Goyal, Wei-Ning Hsu, Mark Ibrahim, Ishan Misra, Ari Morcos, Devi Parikh, David Schwab, Mannat Singh, Shubham Toshniwal, and Karen Ullrich.
– *University of Michigan:* Mohamed El Banani, Ang Cao, Daniel Geng, Richard Higgins, Gaurav Kaul, Nilesh Kulkarni, Andrew Lee, Tiange Luo, Jeongsoo Park, Chris Rockwell, Dandan Shan, and Ayush Shrivastava.
– *Meta (intern cohort):* Desi Ivanova, Lyle Kim, Andre Niyongabo Rubungo, Elizabeth Salesky, and Sagar Vaze.

KD thanks Julius Berner and Steffen Schneider for fruitful discussions over fruity cocktails. KD also thanks *Cannelle* and many other cafés in Ann Arbor for their high-quality espresso shots and many hours of free Wi-Fi.

Appendix

A. Entailment loss derivations

We derive the entailment loss components (Eqn. 12) used in our approach. Note that for $c > 0$, the curvature of the hyperboloid is $-c$.

Half-aperture. To derive the entailment loss for arbitrary curvatures $c > 0$, we start with the expression of half-aperture for the Poincaré ball, introduced by Ganea et al. (2018). Let $\mathbf{x}_b$ be a point on the Poincaré ball, the cone half-aperture is defined as follows:

$$\text{aper}_b(\mathbf{x}_b) = \sin^{-1}\left(K \frac{1 - c\,\|\mathbf{x}_b\|^2}{\sqrt{c}\,\|\mathbf{x}_b\|}\right) \tag{13}$$

The Poincaré ball model and Lorentz hyperboloid model are isometric to each other – one can map any point $\mathbf{x}_b$ from the Poincaré ball to another point $\mathbf{x}_h$ on the hyperboloid using the following differentiable transformation:

$$\mathbf{x}_h = \frac{2\mathbf{x}_b}{1 - c\,\|\mathbf{x}_b\|^2} \tag{14}$$

The half-aperture of a cone should be invariant to the exact hyperbolic model we use, hence $\text{aper}_h(\mathbf{x}_h) = \text{aper}_b(\mathbf{x}_b)$. Substituting Eqn. 14 in Eqn. 13, we get the expression:

$$\text{aper}_h(\mathbf{x}_h) = \sin^{-1}\left(\frac{2K}{\sqrt{c}\,\|\mathbf{x}_h\|}\right)$$

Exterior angle. Consider three points $\mathbf{O}$ (the origin), $\mathbf{x}$ (text embedding) and $\mathbf{y}$ (image embedding). Then, a hyperbolic triangle is a closed shape formed by the geodesics connecting each pair of points. Similar to the Euclidean plane, the hyperbolic plane also has its law of cosines that allows us to talk about the angles in the triangle (Lee, 2019). Let the Lorentzian distances (Eqn. 4) be $x = d(\mathbf{O}, \mathbf{y})$, $y = d(\mathbf{O}, \mathbf{x})$, and $z = d(\mathbf{x}, \mathbf{y})$. We can write the expression of *exterior angle* as follows:

$$\begin{aligned}\text{ext}(\mathbf{x}, \mathbf{y}) &= \pi - \angle\mathbf{Oxy} \\ &= \pi - \cos^{-1}\left[\frac{\cosh(z\sqrt{c})\cosh(y\sqrt{c}) - \cosh(x\sqrt{c})}{\sinh(z\sqrt{c})\sinh(y\sqrt{c})}\right]\end{aligned}$$

We use the relation $\pi - cos^{-1}(t) = cos^{-1}(-t)$ in the above equation. Then, let us define a function $g(t) = cosh(t\sqrt{c})$ for brevity, and substitute in the above equation. We also substitute $\sinh(t) = \sqrt{\cosh^2(t) - 1}$ as per the hyperbolic trigonometric identity. Putting it all together, we get:

$$\text{ext}(\mathbf{x}, \mathbf{y}) = \cos^{-1}\left[\frac{g(x) - g(z)g(y)}{\sqrt{g(z)^2 - 1}\sqrt{g(y)^2 - 1}}\right] \tag{15}$$

Now all we need is to compute $g(x)$, $g(y)$, and $g(z)$. We substitute the $z = d(\mathbf{x}, \mathbf{y})$ in $g(z)$ below:

$$\begin{aligned}g(z) &= \cosh\left(d(\mathbf{x}, \mathbf{y})\sqrt{c}\right) \\ &= \cosh\left(\frac{1}{\sqrt{c}}\cosh^{-1}(-c\,\langle\mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}}) \cdot \sqrt{c}\right) \\ &= -c\,\langle\mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}}\end{aligned}$$

Similarly, $g(x) = -c\langle\mathbf{O}, \mathbf{y}\rangle_{\mathcal{L}}$ and $g(y) = -c\langle\mathbf{O}, \mathbf{x}\rangle_{\mathcal{L}}$. The Lorentzian inner product (Eqn. 1) with origin $\mathbf{O}$ simplifies:

$$\langle\mathbf{O}, \mathbf{x}\rangle_{\mathcal{L}} = -\frac{x_{time}}{\sqrt{c}} \quad \text{and} \quad \langle\mathbf{O}, \mathbf{y}\rangle_{\mathcal{L}} = -\frac{y_{time}}{\sqrt{c}}$$

Through this, we get $g(x) = x_{time}\sqrt{c}$ and $g(y) = y_{time}\sqrt{c}$. Finally, we can substitute $g(x)$, $g(y)$, and $g(z)$ to re-write Eqn. 15 to give the final expression as follows:

$$\text{ext}(\mathbf{x}, \mathbf{y}) = cos^{-1}\left(\frac{y_{time} + x_{time}\,c\,\langle\mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}}}{\sqrt{x_{time}^2 c - 1}\sqrt{(c\,\langle\mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}})^2 - 1}}\right)$$

Finally, we use the relation between x_{time} and $\mathbf{x}_{space}$ (Eqn. 3) to simplify the denominator, giving the final expression of exterior angle as follows:

$$\text{ext}(\mathbf{x}, \mathbf{y}) = cos^{-1}\left(\frac{y_{time} + x_{time}\,c\,\langle\mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}}}{\|\mathbf{x}_{space}\|\sqrt{(c\,\langle\mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}})^2 - 1}}\right)$$

Table 5. **CLIP baseline.** We develop a strong CLIP baseline that trains on an 8-GPU machine in less than one day (ViT-S image encoder), starting with SLIP (Mu et al., 2022) as a reference. We benchmark improvements on zero-shot image classification across 16 datasets. Our RedCaps-trained CLIP baseline (last row) is a significantly stronger baseline than its YFCC-trained counterparts.

	Images Seen	ImageNet	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Country211	Average
YFCC15M-trained models																		
SLIP's CLIP (Mu et al., 2022)	368M	32.0	43.7	61.9	30.2	30.9	41.3	3.5	3.9	18.1	26.1	51.4	48.7	87.3	17.5	16.8	8.7	32.6
Our implementation	368M	33.1	42.3	64.9	34.4	33.7	43.8	2.9	5.1	19.1	25.0	49.8	47.2	87.4	26.8	21.6	9.0	34.1
+ BS 4096→2048	184M	28.2	34.2	58.7	29.4	27.4	39.4	2.9	4.3	16.5	20.1	43.8	42.2	85.4	20.2	19.0	8.5	30.0
+ *sin-cos pos embed*	184M	28.7	34.2	67.3	33.6	25.4	41.1	3.1	4.2	17.8	21.0	44.3	43.6	86.4	18.6	19.6	8.3	31.1
RedCaps-trained models																		
+ YFCC→RedCaps	184M	32.6	71.5	61.4	25.6	29.9	27.5	10.1	1.5	14.3	72.7	62.8	42.2	88.0	18.1	30.5	4.9	37.1
+ 90K→120K iters.	246M	33.9	72.5	60.1	24.4	30.0	27.5	11.3	1.4	13.1	73.7	63.9	44.4	88.2	18.6	31.4	5.2	37.5
+ our zero-shot prompts	246M	34.3	74.5	60.1	24.4	33.8	27.5	11.3	1.4	15.0	73.7	63.9	47.0	88.2	18.6	31.4	5.2	38.1

B. Developing a strong CLIP baseline

One of our contributions is to establish a lightweight, yet strong CLIP baseline. The original CLIP models (Radford et al., 2021) are trained using a private dataset of 400M image-text pairs across 128 GPUs for more than 10 days. We aim to maximize accessibility for future works, hence we decide our hyperparameters such that our smallest model can train on a single 8-GPU machine in less than one day.

We start with a reference CLIP ViT-S/16 baseline from SLIP (Mu et al., 2022) and carefully introduce one modification at a time. We benchmark improvements on zero-shot image classification across 16 datasets used in our main experiments, using text prompts used by (Radford et al., 2021). Results are shown in Table 5.

CLIP baseline by SLIP. This re-implemented baseline was trained using a 15M subset of the YFCC dataset (Thomee et al., 2016). We re-evaluate the publicly released ViT-S/16 checkpoint [3] using our evaluation code; it obtains 32.6% average accuracy across all datasets.

Our re-implementation. We attempt a faithful replication of CLIP by following hyperparameters in SLIP. Our implementation obtains slightly higher average performance (34.1%) with three minor changes:

- We use an *undetached* gather operation to collect all image/text features across all GPUs for contrastive loss. This ensures proper gradient flow across devices.
- The above change allows using weight decay $= = 0.2$ like OpenAI's CLIP, unlike 0.5 used by SLIP's CLIP.
- During training and inference, we resize input images using *bicubic* interpolation like original CLIP, instead of bilinear interpolation in SLIP's CLIP.

Fitting the model on 8-GPUs. This CLIP model requires 16× V100 32GB GPUs with a batch size of 4096 and automatic mixed precision (Micikevicius et al., 2018). Techniques like gradient checkpointing (Chen et al., 2016) can reduce memory requirements, but it comes at a cost of reduced training speed. Hence we avoid making it a requirement and simply reduce the batch size to 2048. This incurs a performance drop as the effective images seen by the model are halved. We offset the effective shortening of the training schedule by using fixed *sine-cosine* position embeddings in ViT, so learning position-related inductive biases is not required. This change slightly improves average accuracy ($30.0\% \rightarrow 31.1\%$ average accuracy).

Training with RedCaps dataset. RedCaps dataset (Desai et al., 2021) comprises 12M image-text pairs from Reddit, sourced from Pushshift (Baumgartner et al., 2020). Training with RedCaps significantly improves performance over YFCC-trained models ($31.1\% \rightarrow 37.1\%$ average accuracy), especially on datasets whose concepts have high coverage in RedCaps, *e.g.,* Food-101 (Bossard et al., 2014) and Pets (Parkhi et al., 2012).

To account for the smaller size of RedCaps, we increase the training iterations from 90K up to 120K. Finally, we modify zero-shot prompts for some datasets to match the linguistic style of RedCaps. For example, many captions in `r/food` simply mention the name of the dish in the corresponding image, hence we use the prompt '`food : {}`'. See Table 8 for the list of prompts for all datasets. We did not extensively tune these prompts, but we checked performance on the held-out validation sets to avoid cheating on the test splits.

Finally, our CLIP ViT-S/16 baseline trains on 8× V100 32 GB GPUs within ≈14 hours and achieves 38.1% average performance across 16 datasets. We use these hyperparameters for all MERU and CLIP models in our experiments.

[3] `github.com/facebookresearch/slip`

Table 6. **Datasets used for image classification evaluation.** Datasets in highlighted rows do not have an official validation split – we use a random held-out subset of the training split. EuroSAT and RESISC do not define any splits; we randomly sample non-overlapping splits. CLEVR Counts is derived from CLEVR (Johnson et al., 2017) and SST2 was introduced as an NLP dataset by (Socher et al., 2013).

Dataset	Classes	Train	Val	Test
Food-101 (Bossard et al., 2014)	101	68175	7575	25250
CIFAR-10 (Krizhevsky, 2009)	10	45000	5000	10000
CIFAR-100 (Krizhevsky, 2009)	100	45000	5000	10000
CUB-2011 (Wah et al., 2011)	200	4795	1199	5794
SUN397 (Xiao et al., 2010)	397	15880	3970	19849
Stanford Cars (Krause et al., 2013)	196	6515	1629	8041
FGVC Aircraft (Maji et al., 2013)	100	3334	3333	3333
DTD (Cimpoi et al., 2014)	47	1880	1880	1880
Oxf-IIIT Pets (Parkhi et al., 2012)	37	2944	736	3669
Caltech-101 (Fei-Fei et al., 2004)	102	2448	612	6084
Flowers (Nilsback & Zisserman, 2008)	102	1020	1020	6149
STL-10 (Coates et al., 2011)	10	4000	1000	8000
EuroSAT (Helber et al., 2019)	10	5000	5000	5000
RESISC (Cheng et al., 2017)	45	3150	3150	25200
Country211 (Radford et al., 2021)	211	31650	10550	21100
MNIST (LeCun et al., 2010)	10	48000	12000	10000
CLEVR Counts (Zhai et al., 2019)	8	4500	500	5000
PCAM (Veeling et al., 2018)	2	262144	32768	32768
SST2 (Radford et al., 2021)	2	6920	872	1821

Table 7. **Linear probe evaluation.** We train a logistic regression classifier on embeddings extracted from the image encoders of CLIP and MERU (before projection layers). Note that embeddings from MERU are *not* lifted onto the hyperboloid.

		Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Country211	MNIST	CLEVR	PCAM	SST2
ViT S/16	CLIP	85.3	89.6	72.3	68.8	61.1	60.5	42.2	71.2	87.9	88.4	96.2	95.5	95.7	88.1	15.0	98.5	57.5	84.6	54.9
	MERU	85.2	89.7	70.9	69.2	59.6	58.0	43.1	70.2	87.5	85.6	95.5	95.5	95.8	87.0	14.8	98.2	56.8	84.1	54.5
ViT B/16	CLIP	88.4	92.2	76.5	73.2	64.7	71.1	50.4	72.6	90.2	89.6	97.3	97.1	96.9	90.0	16.7	98.9	52.7	84.4	57.6
	MERU	88.2	92.3	74.6	70.9	63.4	68.4	48.2	70.7	90.3	88.6	96.6	96.7	96.5	89.0	16.5	98.7	56.0	85.5	56.2
ViT L/16	CLIP	89.6	95.3	80.5	75.7	66.0	75.7	54.5	75.7	92.0	92.0	97.4	97.6	96.9	90.5	17.8	99.2	55.6	87.5	56.1
	MERU	89.0	94.1	77.3	74.2	63.7	71.9	51.2	70.9	90.1	87.5	96.7	97.3	96.8	89.1	17.0	98.9	55.4	86.0	55.8

C. Linear probe evaluation

Our experimental evaluations (Section 4) focus on *zero-shot* transfer (Elhoseiny et al., 2013; Radford et al., 2021). Another established protocol to evaluate visual representations is *linear probe evaluation*, which involves training linear models on *frozen* image embeddings. This protocol is popular in self-supervised representation learning literature, with Doersch et al. (2015), Zhang et al. (2016), and Noroozi & Favaro (2016) being notable early works. We follow the implementation of Kornblith et al. (2019) as it is simple and less sensitive to choice of evaluation hyperparameters. This setup is also followed by CLIP (Radford et al., 2021) and many recent works on representation learning (El Banani et al., 2023; Fürst et al., 2022; Li et al., 2022).

We evaluate using datasets listed in Table 6. We train a logistic regression classifier on embeddings extracted from the image encoder (before projection layer) of MERU and CLIP. For MERU, these underlying representations belong to a Euclidean space. We use the implementation from `scikit-learn` (Pedregosa et al., 2011) library, with L-BFGS (Liu & Nocedal, 1989) optimizer and search the regularization cost per dataset, $C \in [10^{-6}, 10^{6}]$, performing two-step search on `val` split like Radford et al. (2021). Then we train a final classifier on combined `train` and `val` splits for a maximum of 1000 iterations, then report top-1 mean per-class accuracy on the `test` split.

Results in Table 7 show that MERU mostly matches or underperforms CLIP. Our main focus is not on improving the underlying Euclidean representations from the encoders, but to demonstrate strong *zero-shot* transfer and interpretability benefits. Future work can focus on improving MERU's capabilities on other transfer applications.

Table 8. **Prompts used for zero-shot classification (Section 4.3).** Most of these prompts are same as (Radford et al., 2021). We modify prompts for some datasets, that significantly improved performance for both MERU and CLIP – We did not perform extensive prompt tuning, we simply checked the performance on `val` splits for our CLIP baseline (Appendix B). **NOTE:** Some prompts use the word '`porn`' as it is included in the subreddit name. It does not indicate pornographic content but simply high-quality photographs.

ImageNet (our prompts)		
i took a picture : itap of a {}.	pics : a bad photo of the {}.	pics : a origami {}.
pics : a photo of the large {}.	pics : a {} in a video game.	pics : art of the {}.
pics : a photo of the small {}.		

Food-101 (our prompts)
food : {}.
food porn : {}.
CIFAR-10 and CIFAR-100
a photo of a {}.
a blurry photo of a {}.
a black and white photo of a {}.
a low contrast photo of a {}.
a high contrast photo of a {}.
a bad photo of a {}.
a good photo of a {}.
a photo of a small {}.
a photo of a big {}.
a photo of the {}.
a blurry photo of the {}.
a black and white photo of the {}.
a low contrast photo of the {}.
a high contrast photo of the {}.
a bad photo of the {}.
a good photo of the {}.
a photo of the small {}.
a photo of the big {}.
CUB-2011 (our prompts)
bird pics : {}.
birding : {}.
birds : {}.
bird photography : {}.
SUN397
a photo of a {}.
a photo of the {}.
Stanford Cars
a photo of a {}.
a photo of the {}.
a photo of my {}.
i love my {}!
a photo of my dirty {}.
a photo of my clean {}.
a photo of my new {}.
a photo of my old {}.
FGVC Aircraft
a photo of a {}, a type of aircraft.
a photo of the {}, a type of aircraft.

DTD (our prompts)
pics : {} texture.
pics : {} pattern.
pics : {} thing.
pics : this {} texture.
pics : this {} pattern.
pics : this {} thing.
Oxford-IIIT Pets
a photo of a {}, a type of pet.
Caltech-101
a photo of a {}.
a painting of a {}.
a plastic {}.
a sculpture of a {}.
a sketch of a {}.
a tattoo of a {}.
a toy {}.
a rendition of a {}.
a embroidered {}.
a cartoon {}.
a {} in a video game.
a plushie {}.
a origami {}.
art of a {}.
graffiti of a {}.
a drawing of a {}.
a doodle of a {}.
a photo of the {}.
a painting of the {}.
the plastic {}.
a sculpture of the {}.
a sketch of the {}.
a tattoo of the {}.
the toy {}.
a rendition of the {}.
the embroidered {}.
the cartoon {}.
the {} in a video game.
the plushie {}.
the origami {}.
art of the {}.
graffiti of the {}.
a drawing of the {}.
a doodle of the {}.

Oxford Flowers (our prompts)
flowers : {}.
STL10
a photo of a {}.
a photo of the {}.
EuroSAT
a centered satellite photo of {}.
a centered satellite photo of a {}.
a centered satellite photo of the {}.
RESISC
satellite imagery of {}.
aerial imagery of {}.
satellite photo of {}.
aerial photo of {}.
satellite view of {}.
aerial view of {}.
satellite imagery of a {}.
aerial imagery of a {}.
satellite photo of a {}.
aerial photo of a {}.
satellite view of a {}.
aerial view of a {}.
satellite imagery of the {}.
aerial imagery of the {}.
satellite photo of the {}.
aerial photo of the {}.
satellite view of the {}.
aerial view of the {}.
Country211
a photo i took in {}.
a photo i took while visiting {}.
a photo from my home country of {}.
a photo from my visit to {}.
a photo showing the country of {}.
MNIST
a photo of the number: "{}".
CLEVR
a photo of {} objects.
Patch Camelyon
this is a photo of {}.
Rendered SST2
a {} review of a movie.

D. Image traversals: more details and results

Our qualitative analysis in Section 5 involves inferring the learned visual-semantic hierarchy in the representation space through image traversals. We performed shortest-path traversal from a given image embedding $\mathbf{y}$ to the `[ROOT]` embedding by interpolating 50 equally spaced steps. At each step, we retrieve text from a set $\mathcal{X}$ of text embeddings (including `[ROOT]`). Here we include the precise methodology details to perform image traversals.

MERU and CLIP have different methods for interpolation and nearest-neighbor retrieval due to the difference in geometric properties of Euclidean and hyperbolic spaces.

Interpolating steps:

- **CLIP:** We linearly interpolate between L^2 normalized embeddings of $\mathbf{y}$ and `[ROOT]`, and then L^2 normalize all *step* embeddings. In PyTorch (Paszke et al., 2019), `torch.lerp` can perform this linear interpolation.
- **MERU:** We linearly interpolate in the tangent space, between $\mathbf{v} = \mathrm{logm}_{\mathbf{O}}(\mathbf{y})$ (Eqn. 8) and $\mathbf{O}$ (origin is `[ROOT]`), then lift all *step* embeddings onto the hyperboloid.

Nearest-neighbor text retrieval:

- **CLIP:** We select $\mathbf{x} \in \mathcal{X}$ having the highest *cosine similarity* with the *step* embedding.
- **MERU:** First we create a subset $\mathcal{X}_e \subset \mathcal{X}$ of text embeddings that *entail* the given *step* embedding, *i.e.*, Eqn. 12 evaluates to 0 (note that `[ROOT]` entails everything). Then we select $\mathbf{x} \in \mathcal{X}_e$ having the highest *Lorentzian inner product* with the *step* embedding.

At any given step, the caption associated with the retrieved texct embedding $\mathbf{x}$ (or `[ROOT]`) is the retrieved nearest neighbor. We observed that multiple consecutive *steps* retrieve the same caption, so our results only display *unique* captions encountered during the traversal.

Caption sources: We create the set of text embeddings $\mathcal{X}$ using captions collected from two different sources.

- **`pexels.com` metadata:** We manually collect metadata (Figure 6), then filter tags to only keep nouns and adjectives (total 750 captions and tags). We filter by performing parts-of-speech using the RoBERTa (Liu et al., 2019) model (`en-core-web-trf`) from SpaCy (Honnibal et al., 2020) library. Finally, we convert tags to captions by filling prompts – `'a photo of {}.'` for nouns, and `'this photo is {}.'` for adjectives.
- **YFCC dataset:** We use the `text descriptions` of the YFCC-15M subset (Radford et al., 2021). We perform minimal text processing of these captions according to RedCaps (Desai et al., 2021), to match the training data distribution. This involves converting to lowercase, using `ftfy` (Speer, 2019) to strips accents and non-latin characters, removing all sub-strings enclosed in brackets (`(.*)`, `[.*]`), and replacing social media handles (words starting with '@') with a <usr>. We also remove captions having more than 20 tokens (for ease of visualization). Finally, we obtain ≈8.7M captions.

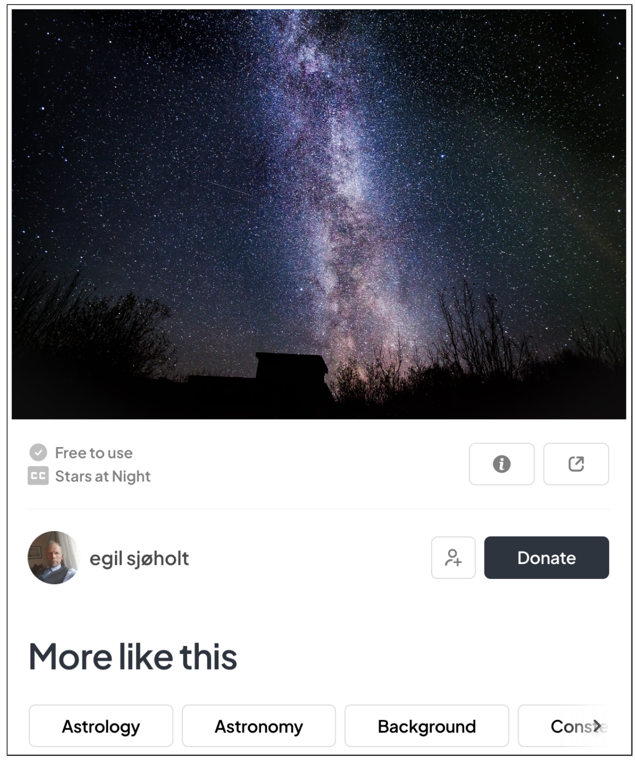


Figure 6. **`pexels.com` webpage** of an image used in our results. We manually collect the closed caption (*CC*) and *'More like this'* tags for all images to create the retrieval set for image traversals.

Results: Figure 5 shows selected qualitative examples with 8 out of 60 images. On the next pages, Figures 7 to 11 include results with other 52 images. After the image credits (Appendix E), we display results with YFCC captions.

MERU	CLIP
golden gate	*golden gate bridge, san francisco, california*
san francisco	*famous landmark*
tourist spot	↓
photo	↓
power	↓
[ROOT]	[ROOT]

MERU	CLIP
white cliffs of dover in england	*white cliffs of dover*
white cliffs of dover	*cliffs*
white	*rocky*
coast	↓
country	↓
[ROOT]	[ROOT]

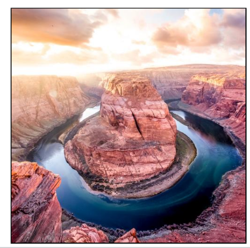
MERU	CLIP
the famous fountain paint pots in yellowstone national park	*yellowstone*
yellowstone	*beauty*
national park	↓
power	↓
[ROOT]	[ROOT]

MERU	CLIP
the parthenon temple ruins in athens greece	*the parthenon temple ruins in athens greece*
historical site	*famous landmark*
architecture	*low angle shot*
domestic	↓
[ROOT]	[ROOT]

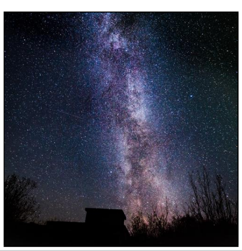
MERU	CLIP
big ben	*big ben*
holiday	↓
day	↓
[ROOT]	[ROOT]

MERU	CLIP
karlskirche	*karlskirche church*
architecture	*church*
style	↓
[ROOT]	[ROOT]

MERU	CLIP
fuji	*fuji*
japan	*cozy*
holiday	↓
[ROOT]	[ROOT]



MERU	CLIP
horseshoe bend	*horseshoe bend*
outdoors	*national park*
↓	*credit*
[ROOT]	[ROOT]



MERU	CLIP
milky way	↓
rural	↓
[ROOT]	[ROOT]

MERU	CLIP
volcano erupting at night under starry sky	*volcano erupting at night under starry sky*
active volcano	*volcanic*
outdoors	↓
[ROOT]	[ROOT]

MERU	CLIP
northern lights norway	*northern lights norway*
aurora	*aurora*
scenic	*outdoors*
outdoor	↓
[ROOT]	[ROOT]

MERU	CLIP
california	*welcome to fabulous las vegas nevada signage*
↓	*famous landmark*
[ROOT]	[ROOT]

Figure 7. **Image traversals with MERU and CLIP (locations and landmarks).** Retrieved captions are sourced from `pexels.com` metadata. MERU captures a more systematic and fine-grained visual-semantic hierarchy than CLIP – trends are same as Figure 5.

MERU	CLIP
squirrel up on the snow covered tree	*squirrel up on the snow covered tree*
squirrel	*squirrel*
wildlife	↓
fluffy	↓
[ROOT]	[ROOT]

MERU	CLIP
seagull	*seagull*
bird	*bird*
air	↓
coast	↓
day	↓
[ROOT]	[ROOT]

MERU	CLIP
cute pug sitting on floor in white kitchen	*cute pug sitting on floor in white kitchen*
pug	↓
domestic	↓
little	↓
[ROOT]	[ROOT]

MERU	CLIP
three zebras	*three zebras*
zebras	*wild animals*
safari	↓
animal photography	↓
wild	↓
[ROOT]	[ROOT]

MERU	CLIP
monarch butterfly perching on red flower	*monarch butterfly*
monarch butterfly	↓
butterfly	↓
beauty	↓
day	↓
[ROOT]	[ROOT]

MERU	CLIP
red hibiscus in bloom	*red hibiscus in bloom*
hibiscus	*hibiscus*
bloom	*blooming flowers*
style	↓
[ROOT]	[ROOT]

MERU	CLIP
white chicken on green grass field	*white chicken on green grass field*
cockerel	↓
chicken	↓
style	↓
[ROOT]	[ROOT]

MERU	CLIP
yellow blue and white macaw perched on brown tree branch	*yellow blue and white macaw perched on brown tree branch*
parrot	*parrot*
hungry	*animal*
female	↓
[ROOT]	[ROOT]

MERU	CLIP
edible agaric	*edible agaric*
mushroom	*mushroom*
beauty	*beauty*
little	↓
[ROOT]	[ROOT]

MERU	CLIP
aquatic animals	*aquatic animals*
sea life	*sea life*
style	*calamity*
[ROOT]	[ROOT]

MERU	CLIP
financial	*adorable*
cute	↓
[ROOT]	[ROOT]

MERU	CLIP
an orca whale jumping out of the water	*an orca whale jumping out of the water*
whale	*whale*
[ROOT]	[ROOT]

Figure 8. **Image traversals with MERU and CLIP (flora and fauna).** Retrieved captions are sourced from `pexels.com` metadata. MERU captures a more systematic and fine-grained visual-semantic hierarchy than CLIP – trends are same as Figure 5.

MERU	CLIP
bread and coffee for breakfast	bread and coffee for breakfast
pastry	↓
art	↓
[ROOT]	[ROOT]

MERU	CLIP
grilled cheese	grilled cheese
lunch	↓
delicious	↓
classic	↓
[ROOT]	[ROOT]

MERU	CLIP
bowl of ramen	ramen
local food	↓
tasty	↓
[ROOT]	[ROOT]

MERU	CLIP
green chili peppers and a knife	green chili peppers and a knife
spicy food	↓
spicy	↓
[ROOT]	[ROOT]

MERU	CLIP
spinach caprese salad	spinach caprese salad
lunch	lunch
homemade	↓
style	↓
[ROOT]	[ROOT]

MERU	CLIP
cupcakes	cupcakes
chocolate cupcakes	↓
delicious	↓
homemade	↓
clean	↓
day	↓
[ROOT]	[ROOT]

MERU	CLIP
pav bhaji	pav bhaji dish on a bowl
indian food	indian food
traditional food	meal
local food	dinner
spicy	↓
[ROOT]	[ROOT]

MERU	CLIP
clear glass bottle filled with broccoli shake	smoothie
smoothie	homemade
local food	vegetable
homemade	↓
spicy	↓
[ROOT]	[ROOT]

MERU	CLIP
vada pav	cheese
traditional food	↓
[ROOT]	[ROOT]

MERU	CLIP
old fashioned	nutrition
spicy	↓
style	↓
[ROOT]	[ROOT]

MERU	CLIP
latte	latte
design	↓
style	↓
[ROOT]	[ROOT]

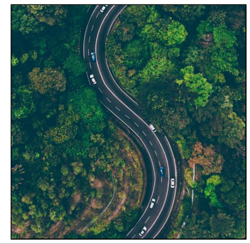
MERU	CLIP
espresso martini	↓
cocktail	↓
dessert	↓
hot	↓
[ROOT]	[ROOT]

Figure 9. **Image traversals with MERU and CLIP (food and drinks).** Retrieved captions are sourced from `pexels.com` metadata. MERU captures a more systematic and fine-grained visual-semantic hierarchy than CLIP – trends are same as Figure 5.

MERU	CLIP
campfire	*inferno*
fire	↓
blaze	↓
hot	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
cumulus	*cumulus*
white clouds	↓
clouds	↓
health	↓
fluffy	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
raining in the city	*raining in the city*
weather	*downtown*
simple	↓
day	↓
`[ROOT]`	`[ROOT]`



MERU	CLIP
road	*aerial view of road in the middle of trees*
travel	*aerial shot*
style	*rural*
↓	*clean*
`[ROOT]`	`[ROOT]`

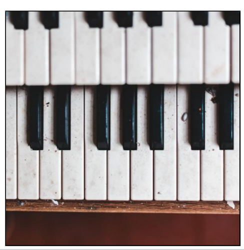
MERU	CLIP
mountain bike on the beach	*mountain bike on the beach*
analog	*bicycle*
retro	↓
style	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
lights	*white heart shaped candle on dried leaves*
evening	*holiday*
day	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
bedroom	↓
clean	↓
`[ROOT]`	`[ROOT]`

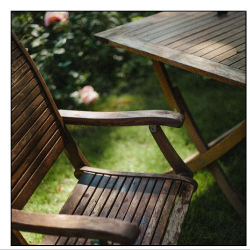
MERU	CLIP
clean bathroom	*stainless steel faucet on white ceramic sink*
investment	↓
clean	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
jack o lantern with light	*jack o lantern with light*
carved pumpkin	↓
halloween	↓
hot	↓
`[ROOT]`	`[ROOT]`



MERU	CLIP
piano keys	*musical instrument*
keyboard	*music*
analog	↓
vintage	↓
style	↓
`[ROOT]`	`[ROOT]`

MERU	CLIP
assorted gift boxes on floor near christmas tree	*christmas presents*
christmas presents	*christmas gifts*
christmas gifts	↓
`[ROOT]`	`[ROOT]`



MERU	CLIP
garden table and chair	*seat*
table	↓
design	↓
comfort	↓
`[ROOT]`	`[ROOT]`

Figure 10. **Image traversals with MERU and CLIP (objects and scenes).** Retrieved captions are sourced from `pexels.com` metadata. MERU captures a more systematic and fine-grained visual-semantic hierarchy than CLIP – trends are same as Figure 5.

MERU	CLIP
turned on floor lamp near sofa on a library room	*bookshelves*
books	↓
bookshelves	↓
comfort room	↓
cozy	↓
style	↓
[ROOT]	[ROOT]

MERU	CLIP
pineapple	*ripe pineapple on gray rock beside body of water*
inspiration	*pineapple*
health	*calamity*
little	↓
[ROOT]	[ROOT]

MERU	CLIP
cockatiel	*cockatiel*
female	↓
[ROOT]	[ROOT]

MERU	CLIP
currency	*euro*
simple	*revenue*
[ROOT]	[ROOT]

Figure 11. **Image traversals (objects and scenes).** Retrieved captions are sourced from `pexels.com` metadata. MERU captures a more systematic and fine-grained visual-semantic hierarchy than CLIP – trends are same as Figure 5.

E. Image credits

All images displayed in this paper are collected from `pexels.com`, a photography website that offers images with permissible usage licenses. Below is the list of the image source URLs listed in order of their appearance in the paper. We thank all the photographers for generously sharing these images.

Illustration of the visual-semantic hierarchy (Figure 1).

– `www.pexels.com/photo/adult-yellow-labrador-retriever-standing-on-snow-field-1696589`
– `www.pexels.com/photo/homeless-cat-fighting-with-dog-on-street-6601811`
– `www.pexels.com/photo/short-coated-gray-cat-20787`

Image traversals – results in the main paper (Figure 5).

(1) `www.pexels.com/photo/a-bengal-cat-sitting-beside-wheatgrass-on-a-white-surface-7123957`
(2) `www.pexels.com/photo/white-horse-running-on-green-field-1996337`
(3) `www.pexels.com/photo/photography-of-rainbow-during-cloudy-sky-757239`
(4) `www.pexels.com/photo/retro-photo-camera-on-table-7162551`
(5) `www.pexels.com/photo/avocado-toast-served-on-white-plate-10464867`
(6) `www.pexels.com/photo/photo-of-brooklyn-bridge-new-york-2260783`
(7) `www.pexels.com/photo/taj-mahal-through-an-arch-2413613`
(8) `www.pexels.com/photo/sydney-opera-house-7088958`

Image traversals – locations and landmarks (Figure 7).

(9) `www.pexels.com/photo/golden-gate-bridge-san-francisco-california-1141853`
(10) `www.pexels.com/photo/white-cliffs-of-dover-in-england-9692909`
(11) `www.pexels.com/photo/the-famous-fountain-paint-pots-in-yellowstone-national-park-12767016`
(12) `www.pexels.com/photo/the-parthenon-temple-ruins-in-athens-greece-14446783`
(13) `www.pexels.com/photo/famous-big-ben-under-cloudy-sky-14434677`
(14) `www.pexels.com/photo/karlskirche-church-7018621`
(15) `www.pexels.com/photo/mt-fuji-3408353`
(16) `www.pexels.com/photo/horseshoe-bend-arizona-2563733`
(17) `www.pexels.com/photo/stars-at-night-1906667`
(18) `www.pexels.com/photo/volcano-erupting-at-night-under-starry-sky-4220967`

(19) www.pexels.com/photo/northern-lights-1933319
(20) www.pexels.com/photo/attraction-building-city-hotel-415999

Image traversals – flora and fauna (Figure 8).

(21) www.pexels.com/photo/squirrel-up-on-the-snow-covered-tree-15306429
(22) www.pexels.com/photo/a-seagull-flying-under-blue-sky-12509256
(23) www.pexels.com/photo/cute-pug-sitting-on-floor-in-white-kitchen-11199295
(24) www.pexels.com/photo/three-zebras-2118645
(25) www.pexels.com/photo/monarch-butterfly-perching-on-red-flower-1557208
(26) www.pexels.com/photo/red-hibiscus-in-bloom-5801054
(27) www.pexels.com/photo/white-chicken-on-green-grass-field-58902
(28) www.pexels.com/photo/yellow-blue-and-white-macaw-perched-on-brown-tree-branch-12715261
(29) www.pexels.com/photo/closeup-photo-of-red-and-white-mushroom-757292
(30) www.pexels.com/photo/photo-of-jellyfish-lot-underwater-3616240
(31) www.pexels.com/photo/yellow-labrador-retriever-wearing-red-cap-4588002
(32) www.pexels.com/photo/an-orca-whale-jumping-out-of-the-water-7767974

Image traversals – food and drinks (Figure 9).

(33) www.pexels.com/photo/bread-and-coffee-for-breakfast-15891938
(34) www.pexels.com/photo/grilled-cheese-on-a-plate-14941252
(35) www.pexels.com/photo/bowl-of-ramen-12984979
(36) www.pexels.com/photo/green-chili-peppers-and-a-knife-5792428
(37) www.pexels.com/photo/spinach-caprese-salad-on-white-ceramic-plate-4768996
(38) www.pexels.com/photo/chocolate-cupcakes-635409
(39) www.pexels.com/photo/pav-bhaji-dish-on-a-bowl-5410400
(40) www.pexels.com/photo/clear-glass-bottle-filled-with-broccoli-shake-1346347
(41) www.pexels.com/photo/vada-pav-15017417
(42) www.pexels.com/photo/old-fashioned-cocktail-drink-4762719
(43) www.pexels.com/photo/coffee-in-white-ceramic-teacup-on-white-ceramic-suacer-894696
(44) www.pexels.com/photo/espresso-martini-in-close-up-photography-15082368

Image traversals – objects and scenes (Figure 10).

(45) www.pexels.com/photo/photograph-of-a-burning-fire-672636
(46) www.pexels.com/photo/white-clouds-in-blue-sky-8354530
(47) www.pexels.com/photo/raining-in-the-city-2448749
(48) www.pexels.com/photo/aerial-view-of-road-in-the-middle-of-trees-1173777
(49) www.pexels.com/photo/mountain-bike-on-the-beach-10542237
(50) www.pexels.com/photo/wax-candles-burning-on-ground-14184952
(51) www.pexels.com/photo/white-wooden-shelf-beside-bed-2062431
(52) www.pexels.com/photo/stainless-steel-faucet-on-white-ceramic-sink-3761560
(53) www.pexels.com/photo/jack-o-lantern-with-light-5659699
(54) www.pexels.com/photo/black-and-white-piano-keys-4077310
(55) www.pexels.com/photo/assorted-gift-boxes-on-floor-near-christmas-tree-3394779
(56) www.pexels.com/photo/garden-table-and-chair-14831985

Image traversals – objects and scenes (Figure 11).

(57) www.pexels.com/photo/turned-on-floor-lamp-near-sofa-on-a-library-room-1907784
(58) www.pexels.com/photo/ripe-pineapple-on-gray-rock-beside-body-of-water-29555
(59) www.pexels.com/photo/close-up-shot-of-a-cockatiel-13511241
(60) www.pexels.com/photo/antique-bills-business-cash-210600

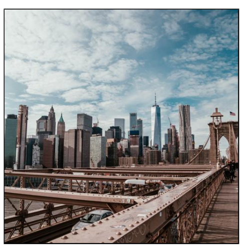
Image traversals with YFCC captions.

(1)

MERU	CLIP
leopard and stig have a beautiful piano at their home.	*loki is a 1 year old bengal cat.*
merlin wasn't impressed to leave the last house and his precious cat grass	↓
my parents cat 'barry' loves being photographed!	↓
house cat posing	↓
mr . bo-majed	↓
our cat, our love our third member of our family. :)	↓
why are you taking pictures? it's dilo don't fill it up. :)	↓
[ROOT]	[ROOT]

(2)

MERU	CLIP
caught my attention by the beautiful light cascading on a grass behind this fellow.	*pity about the camera shake in the evening light*
the focus is all wrong , but the white on the tail and the tongue are pretty cool.	↓
just a goofy white guy.	↓
he was an active one, running to and fro.	↓
but then, she was happy to pose for me	↓
if she were a race horse her name would be poopbiscuit.	↓
dorky photo is dorky	↓
she looks so leery of the camera in this photo.	↓
this is only luky.	↓
[ROOT]	[ROOT]



(3)

MERU	CLIP
going across brooklyn bridge on the way to brooklyn 3 likes on instagram	*shot from the manhattan end of the brooklyn bridge*
manhattan depuis le brooklyn bridge park, a brooklyn.	*much more scenic to walk on than the brooklyn bridge*
bridge, manhattan skyline	*new york new york!*
shot from near the middle of the brooklyn bridge.	↓
this city goes on forever	↓
the city that never sleeps	↓
the city that never sleeps...it can't.	↓
it can be seen from most places in the city	↓
[ROOT]	[ROOT]

(4)

MERU	CLIP
avocado, roasted garlic and sriracha on light rye & raisins sourdough.	*la tartine*
with avocado slices - yum!	*a toast to the new place*
vaguely-healthy	↓
on bread, probably not what you should do with it, but it was a good meal	↓
nice if you like this sort of thing.	↓
[ROOT]	[ROOT]

(5)

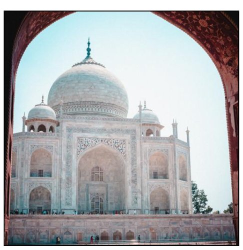
MERU	CLIP
avenida paulista. fisheye 2 kodak elite chrome 100	*leica m7 with voigtlander zoom finder and dsptch camera strap*
rolleiflex kodak portra 160 epson v500 scanner	*rollei minidigi*
...of my brand new shiny 7.1mp camera.	*zeiss ikon icarex, 02/2010*
camera de 5mp	*faux lomo from www.dumpr.net - photo fun*
i'm a lumix camera fan now.	*zeiss-ikon*
[ROOT]	[ROOT]

(6)

MERU	CLIP
double rainbows in our field were too good to pass up photographing	*double rainbows in our field were too good to pass up photographing*
whoa... double rainbow	*is that... a double rainbow? ;-)*
is that... a double rainbow? ;-)	*what does it mean!*
what does it mean of this picture?	↓
only god could create something so beautiful.	↓
this is a good one to end with. reminds me of the woman in this picture.	↓
look out for that right one.	↓
[ROOT]	[ROOT]

(7)

MERU	CLIP
sydney opera house, october 2012.	*gros plan sur l'opera de sydney*
sydney opera house see where this picture was taken.	*you can just make out the opera house in the far left.*
from the new opera house	*my sydney*
i think this is the last one i have of the opera house.	↓
oh, and some opera house, too.	↓
just next to the famous opera house	↓
from horseshoe bay.	↓
taken from the donau.	↓
[ROOT]	[ROOT]



(8)

MERU	CLIP
captured this during my visit to taj mahal, seems like it still inspires young hearts..........	*a weekend adventure to agra to see the taj mahal, see also afternoon and night. luxury.*
the royal mausoleum on the grounds of 'iolani palace	*this pre-dates the taj mahal*
<usr> taj	*you couldn't photograph inside the tombs, so this is all i can show*
rotunda at nmai	*the beauty of age, the mark of wisdom*
outside of yet another palace	*don't remember where.*
taken from city palace.	↓
photography ii	↓
it can be seen from most places in the city	↓
kla photography	↓
[ROOT]	[ROOT]

(9)

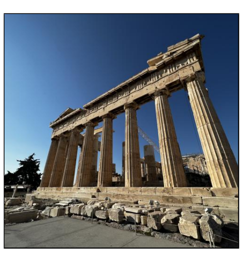
MERU	CLIP
a high dynamic range shot of the golden gate bridge on a foggy afternoon	*golden gate bridge thru photoshop lightroom*
working on the perfect golden gate shot.	*golden gates*
golden gate iii	*no...it's not the golden gate ;)*
everyone who's been to sf has to take this photo at least once in their lives, right?	*by gusf bit.ly/17hga6r*
just got back from sf. will post more on my photoblog: ohad.me	*the independent sf*
¡3 sf	*thinking about painting this makes my shoulder hurt.*
come out and play sf	*still searching for the shot around here.*
back from golden bay	↓
without fog	↓
`[ROOT]`	`[ROOT]`

(10)

MERU	CLIP
white cliffs of dover. august, maybe 2004?	*coloured sand cliffs of alum bay, isle of wight, 1 may 2012.*
calcite cumbria england	*the cliffs are made of limestone.*
poland rocks.	*falkenberg from the south*
white point natural area	*at juta village*
balderstone close	*it's pretty rocky there.*
nepomuk rocks...	↓
l'eglise de giverny	↓
one point if you can tell me where this was taken.	↓
also some kind of guenon, methinks.	↓
`[ROOT]`	`[ROOT]`

(11)

MERU	CLIP
at yellowstone national park there are geyser pools called painted pots because of the colors they exude.	*yellowstone - noth entrance*
yellowstone - noth entrance	*...like i was, how yellowstone got its name*
no trip to yellowstone is complete without it	*i don't remember where this one was. it was striking.*
people enjoy the hot spring, even at this time.	*ain't he a beaut?*
there are hot springs around here somewhere...	↓
wy'east	↓
so many places that were stunning to look at.	↓
there are some special places in the earth. this is one !	↓
with this photo... it's almost like taking a vacation just looking at this.	↓
`[ROOT]`	`[ROOT]`



(12)

MERU	CLIP
the new parthenon museum, next to the acropolis.	*athens archaeological site of the acropolis the parthenon*
this is the magnificent temple of zeus, located in the center of athens	*temple of jupiter and ruins - selinunte*
lil-bit bigger than the athens arch	*ruins from roman time*
roman building , later used as royal residence .	*can't remember where this was*
roman fort/settlement.	↓
built in roman times	↓
we will miss our old place.	↓
at quimbledon.	↓
`[ROOT]`	`[ROOT]`


(13)

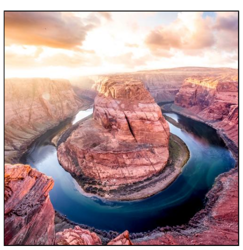
MERU	CLIP
obligatory big ben shot	*big ben i el parlament*
it's what they call a 'big clock' could be thought of as 'big ben'	*or big ben, to his friends*
yeah, i know: big ben is the bell, not the clock tower...	↓
the famous tower	↓
i pretty much only took a photo of this because it was in english	↓
guess what time i took this picture .	↓
`[ROOT]`	`[ROOT]`


(14)

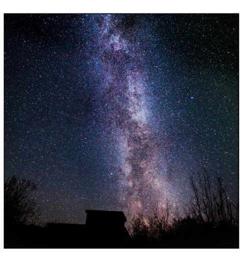
MERU	CLIP
karlskirche, just on the outer ring in vienna .	*kazan, kremin, annunciation cathedral 02/25/2007*
vienna, austria - st. charles church	*bulgaria, sofia, st. nikolai russian church*
wawel cathedral	*almost like a cathedral. a cathedral to transit.*
near st. george's cathedral	*as beautiful as any cathedral*
from my old pda.	*not far from the cathedral*
`[ROOT]`	`[ROOT]`


(15)

MERU	CLIP
who needs mt.fuji	*fuji provia 100*
fuji	*fuji-q highlands*
mt haba	*fuji f30*
mount.	*fuji-san in the background...*
at quimbledon.	*fairmount, in*
`[ROOT]`	`[ROOT]`


(16)

MERU	CLIP
a single exposure of horseshoe bend at sunrise. certainly one of my favorites from the trip.	*yes, there is that backdrop of horseshoe bend :)*
horseshoe bend	*searching for the one ring*
bend over	*<usr>,usa*
looking back at horseshoe canyon	↓
horseshoe bay... this is as close to paradise as you can get!!!!	↓
canyon country, specifically	↓
if you use my photo please post a link and let me know.	↓
`[ROOT]`	`[ROOT]`


(17)

MERU	CLIP
the milk way over bleriot ferry provincial park near drumheller, alberta.	*the milky way as it appeared above the farmhouse in grey county - 30 sec exposure*
the south-western part of the milky way	*outside a house from the austmarka region*
we were in quite a rural place, although there were still lights on the horizon.	*this was just a couple of miles from the farmhouse we stayed in.*
keeping the peace while bush was in town	↓
`[ROOT]`	`[ROOT]`

(18)

MERU	CLIP
lava as seen through night shot of volcan arenal	*lava as seen through night shot of volcan arenal*
the majestic villarica volcano	*volcan osorno*
nice photo of us in front of an active volcano.	*volcanic origin*
still an active volcano	*the volcano!*
around the khorgo volcano.	*with volcan lanin in the background*
mt stromlo	↓
i'm not sure where we were for this shot.	↓
for some reason, i think this photo is great!	↓
`[ROOT]`	`[ROOT]`

(19)

MERU	CLIP
arcs of the northen lights over the mountains neat troms, norway,	*northern lights on the otherside of patreksfjorur*
northern lights, norway	*aurora boreale a kangerlussuaq*
in the night see where this picture was taken.	*to a cinema the aurora*
from the north. see where this picture was taken.	*the village at the end of the world*
sometimes something just looks out of place !	*over a year's worth of photos here.*
`[ROOT]`	`[ROOT]`

(20)

MERU	CLIP
cozy cone motel sign with tower of terror in background. california's adventure park at disneyland resort.	*adam taylor ollie over sign kodak: iso 200*
my favorite tourist attraction in la.	*enjoying jason scott's talk.*
if this place didn't scream la, i don't know what does.	*lost in las vegas- max ruckman*
funny, this place was empty.	*hollywood rip, ride, rockit*
photo : l.g.	↓
`[ROOT]`	`[ROOT]`

(21)

MERU	CLIP
a squirrel enjoying the snow on a not-very cold day.	*a squirrel enjoying the snow on a not-very cold day.*
winter male still coming to food after the snow.	*winter male still coming to food after the snow.*
i don't usually see these type of squirrels down here. loved this little guy. :)	↓
the last nut, my dear!	↓
it's kinda fuzzy. but i love this picture for some reason.	↓
i had to take one picture, okay?	↓
`[ROOT]`	`[ROOT]`

(22)

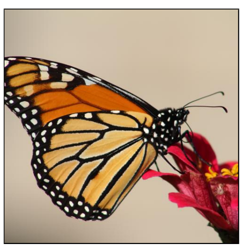
MERU	CLIP
a gull in flight in stratford	*a common or arctic tern flying above the scottish peninsula of kintyre.*
gull on the wing	*more bird shots at dyrholaey.*
we also saw gull-billed but never close enough to photo.	*wouldn't be a trip without at least one picture of a seagull*
taken at little gull islands - b, little gull islands	*seagull!*
taken with the seagull	*some bird thing.*
i was running after the seagull as i took this photo.	*it took patience to get this shot since the stupid bird kept looking away*
not a very nice bird, but still interesting to take pictures of...	↓
i took one westward shot, just to see it	↓
i keep taking this photograph	↓
`[ROOT]`	`[ROOT]`

(23)

MERU	CLIP
margaret willie sanborn-sebo harvey henry sanborn pug dog framed	*sacha my friend and companion patiently waiting for dad to finish taking photos !!!*
everyday bear takes up this position as he waits for his mom to make his food.	*eli begged mommy to take some photos*
patiently waiting for the photographer to get his "shot".	$\downarrow$
"this picture isn't going to work, and i'm going to show you why..."	$\downarrow$
thinking to himself, "what's missing?"	$\downarrow$
this is us... trying to be sultry.	$\downarrow$
to do nothing or to do something.	$\downarrow$
`[ROOT]`	`[ROOT]`

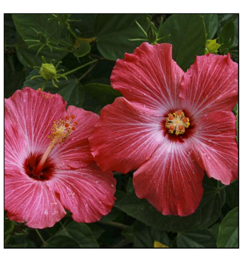
(24)

MERU	CLIP
zebras, ruaha national park	*zebras at kidepo national park, uganda.*
zebra fouls with their mohicans, south africa	*zebra fouls with their mohicans, south africa*
the zebras	*zebras at the watering hole*
photographer: simone kuipers	*three zebras*
photographer: mark antos	*good things come in threes. apparently, so do zebras.*
from photo safari, take 2 .	*the group comes around the bend.*
at oudja	*wild animals*
at quimbledon.	*more straglers*
$\downarrow$	*of the group*
$\downarrow$	*just a little to the left of the middle*
`[ROOT]`	`[ROOT]`



(25)

MERU	CLIP
monarch, danaus plexippus . shot in waimanalo, hawaii.	*a monarch pauses for a drink at a butterfly bush.*
taken at the desert botanical garden in phoenix, arizona, during its seasonal monarch butterfly exhibit.	*monarch in a standard profile*
i'm exercising to capture butterflies.	*the monarch*
visitor to the butterfly tree,	$\downarrow$
monarch	$\downarrow$
butterflies are always free, so enjoy as many of them as you want.	$\downarrow$
butterfly photography	$\downarrow$
i tried to get more shots but it flew away.	$\downarrow$
insect porn	$\downarrow$
`[ROOT]`	`[ROOT]`



(26)

MERU	CLIP
i love the big blooms of the hibiscus with their bold colour.	*some hibiscus-like blossoms beside the visitor center at moody gardens in galveston, tx*
these looked like hibiscus, but i think they are something else.	*after many years nurturing this back to life, the recent heat and rainfall have produced more spectacular blooms.*
only a few blooms this time of year....	*this beautiful species may be the hibiscus according to my wife but i am not so sure,*
much better bloom this time...	*these looked like hibiscus, but i think they are something else.*
i usually dispise flower photos, but i actually like this one.	*looked like they had a lot of nice camera's and video gear. :-)*
this is one is good.	*flourishing.*
`[ROOT]`	`[ROOT]`

(27)

MERU	CLIP
crele old english game bantam cockerel	*read "the beautiful yellow bird - a cautionary tale for new mps" only at acid rabbi.*
sabrina is running and twirling to the bachanalia song from the samson and delila opera.	*flaunt your assets!*
really interesting gadgetar! posted by second life resident ina centaur. visit kaneohe.	*most of the girls had gone to college with mary, whose husband was competing.*
harvey henry sanborn marie elizabeth campbell	↓
sam is showing good form and follow through here.	↓
this is only luky.	↓
`[ROOT]`	`[ROOT]`

(28)

MERU	CLIP
blue-and-yellow macaw at the zooparque itatiba.	*i like this portrait of this blue and yellow macaw, because of the black background...*
one of my fav birds to shoot at the zoo	*from parrot island*
i like these birds, especially when you can see the yellow of their eyes!	*<usr> bird sanctuary.*
there are more colorful birds, but there are few birds with as much character.	*zero post processing*
a beautiful bird, and was quite happy to pose for me.	*one of a bunch*
never had a bird pose for well before!	↓
hi!some more photos...ana	↓
`[ROOT]`	`[ROOT]`

(29)

MERU	CLIP
fly agaric in the forest with a little spider.	*freshly popped amanita muscaria in the forest.*
all alone on the forest floor	*a fly agaric on the rise.*
from a recent new forest trip	*or "fly agaric". "if your viking gets to choose."*
i like this photo, so here it is too.	*reminded me a bit of alice in wonderland*
↓	*reminded me of alice in wonderland*
↓	*this one goes out to forest love.*
`[ROOT]`	`[ROOT]`

(30)

MERU	CLIP
taken at mystic aquarium, ct	*taken by michael i love watching jellyfish. wish these pics had turned out better.*
i don't really like jellyfish, but they are beautiful.	*jelly.*
something about this reminds me of every photo i've ever taken of jellyfish	↓
it is really very cool to be able to see them...	↓
it did not feel like an aquarium when i took the picture	↓
really not much they let you see here.	↓
that i met.	↓
`[ROOT]`	`[ROOT]`

(31)

MERU	CLIP
in new hat from adji	*includes my twit hat- wink*
oui, girl friend.	*in new hat from adji*
the hat actually belongs to honey :)	*:) he's still in japan right now... i miss him.*
fashion photo session	↓
with new hat	↓
a very fit lady.	↓
the boy has style!	↓
`[ROOT]`	`[ROOT]`

(32)

MERU	CLIP
humpback coming up for air near juneau, alaska	*orca, craig, alaska, tongass nf. usfs francisco sanchez*
here, this orca swims about before the show starts.	*named for its shape. not for the occasional whales found in its waters.*
an orca during the show "believe" in sea world .	*a bit shakey!*
humpback	*takin' off.*
a shamu in action	↓
pretend you see a whale and i'll take your photo.	↓
cropitornot shot this one	↓
[ROOT]	[ROOT]

(33)

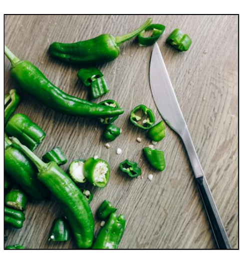
MERU	CLIP
a pastry & a coffee for breakfast yummy	*today's breakfast: toast with honey, egg and black coffee*
a right and proper afternoon tea #1	*afternoon cream tea*
made to perfection. much needed morning tea!	*have a nice cup of tea and a sandwich.*
some photos from me when i working at home	↓
sobrepeso en la proa	↓
[ROOT]	[ROOT]

(34)

MERU	CLIP
'tooey's delight' gc16aga	*dominic samonte & stephanie estabillo*
30 seconds into gina's five-minute toast.	*chicken and cheese sandwich.*
the plain toast was, um, plain.	↓
with grilled cheese sandwich.	↓
i don't normally like toasted sandwiches, but this one was delicious!	↓
on bread, probably not what you should do with it, but it was a good meal	↓
call center del club cantv	↓
so. good.	↓
[ROOT]	[ROOT]

(35)

MERU	CLIP
at momofuku	*where the magic happens at momofuku noodle bar*
the ramen got pwned.	*@ terakawa ramen 4sq.com/ekseer*
siawase ramen	*ramen exploration - still looking*
this i ate, and it was great!	*with michael cotta*
i ate some of this.	↓
[ROOT]	[ROOT]



(36)

MERU	CLIP
tiny hot peppers from the freezer	*poblanos to be roasted*
the ones with jalapenos are the ones that are ruling	*they're best as peppers from the garden!*
add the chillies and cook for another minute.	*do nothing gardening in action!*
getting ready for some chili-dippin'	↓
yo tengo el poder!	↓
toying around with a f/1.8	↓
you canno tmake this stuff up	↓
[ROOT]	[ROOT]

(37)

MERU	CLIP
last nights' salad, cropped	*the first of many caprese salads while we traveled.*
nice salad, summery, fresh.	*for a homegrown salad*
contemp salad	*another salad :)*
i made the salad	↓
it's a verso/kveton thing.	↓
`[ROOT]`	`[ROOT]`

(38)

MERU	CLIP
awesome guinness & chocolate cupcakes by kari stewart our wonderful studio manager	*i made cupcakes and took a million photos...*
peanut butter cupcakes with whipped chocolate ganache.	*from the chocolate lady*
vegan cupcakes. chocolate, coffee and cinnamon.	↓
coffee and chocolate: a "can't miss" cupcake...	↓
including a flourless chocolate cupcake!	↓
chocolate makes everything better. thanks <usr>	↓
chocolate-y goodness!	↓
this week's take, brought over by a friend.	↓
`[ROOT]`	`[ROOT]`

(39)

MERU	CLIP
new delhi's best cholle bhature	*pav bhaji appetizer, $5 dinner time only tirupathi bhimas, milpitas*
en tlaquepaque, jal.	*dal monte la motta .*
west indian food. you can't help but smile when you eat it .	*the famous curry mile, taken on saturday 5th march*
<usr> buse dal lof	*sloppy everything*
paraje guer aike	*the midas, great food*
manoush	↓
with sambuca	↓
`[ROOT]`	`[ROOT]`

(40)

MERU	CLIP
matcha green tea ice blended	*chimichurri sauce recipe*
yes i had a pesto drink	*trust me, i tried to make this less green.*
hot/fermented	*make some good ones!*
that's right, i've been experimenting. trying to keep things fresh.	↓
ondel-ondel	↓
one love hi pawa	↓
`[ROOT]`	`[ROOT]`

(41)

MERU	CLIP
bread with rosemary and garlic infused olive oil at jaleo, a tapas restaurant in my neighborhood.	*sliders served at lee roy selmon's restaurant.*
with rosemary and parmigiano... our mellow new year's menu	*biergarten*
bread with olive oil and vinegar	*sliders & greens*
my welcome brunch to vienna..a cheese party	↓
they serve it with some sort of sauce . this is their version of "bread".	↓
a good shot of how the bread should look.	↓
the bread is real, i think. at least, not glass.	↓
trying out some bread	↓
`[ROOT]`	`[ROOT]`

(42)

MERU	CLIP
adam johnston	*sparkling apple juice my blog: mikaeladanvers.com*
smoked salmon vodka.	*nombre sells the accesories*
it's so cold in here! look at the frosty vodka	*taste like it's fermenting into alcohol.*
warm vodka. i'm still cringing.	↓
tasty, tasty cocktails	↓
made with hugin	↓
off for cocktails	↓
[ROOT]	[ROOT]

(43)

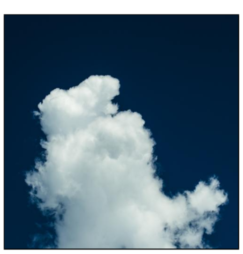
MERU	CLIP
it feels like winter again #coffeemornings	*a latte from blue bottle coffee in oakland, ca.*
by phil o'kane aka icedcoffee	*sweet cold coffee of destiny especially for <usr> :-) 1 likes on instagram*
1 comments on instagram: sorelle_de_latte: you looked great!	↓
heh, maybe my latte art chops aren't so bad after all :-)	↓
cheers <usr> ;)	↓
i likes this one more, i think.	↓
[ROOT]	[ROOT]

(44)

MERU	CLIP
ceu do mapia	*caffe ladro at 5 corners in edmonds. americanvirus.com*
with a leaf, at the caffe espresso.	*chocolate catalan donkey with dinosaur eggs, at eastern time*
bourbon & branch	↓
the braan	↓
enjoying the riff raff	↓
b b: the braes	↓
[ROOT]	[ROOT]

(45)

MERU	CLIP
i was well pleased that i managed to capture the flambe moment ;o)	*usfws photo/heather webb*
everyone was taking pictures of the fire	↓
note to self: need to make more photos of fire again.	↓
this was one of the other interesting things. fire is always fun ;)	↓
i so wanted to photograph this.	↓
[ROOT]	[ROOT]

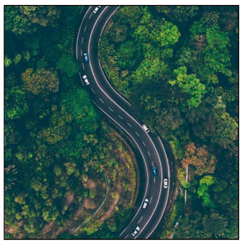


(46)

MERU	CLIP
*i caught a cloud! i caught a cloud! *grin**	*textures courtesy of shadowhousecreations.blogspot.com/*
it's the cloud, baby!	*my first cloud photo ever. :)*
i liked this cloud.	*blowing cloud!*
i took this one because of that cloud. seriously.	↓
i like this. i think it liked me.	↓
[ROOT]	[ROOT]

(47)

MERU	CLIP
a drizzly february day in vancouver.	*taken with a disposable camera on a rainy seattle day.*
weather gloomy all the way to chicago	*a cold rainy day in chicago*
summer rain in the city.	*junechicago*
street<usr>	↓
<usr> street	↓
urban as usual	↓
street pics	↓
[ROOT]	[ROOT]



(48)

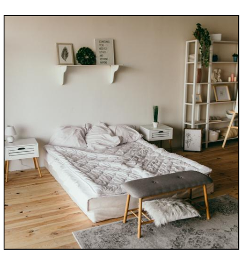
MERU	CLIP
this beautiful track goes through deep gorges of nilgiris towards mangalore.	*sepang international circuit malaysia*
cipularang highway	*very well finished road for indonesia!! ... looking down the road*
<usr> road	*nature highway..*
kinokuniya just down the road	*if you don't know where you are going, any road will take you there.*
crookhaven	*look, there's a road and everything.*
[ROOT]	[ROOT]

(49)

MERU	CLIP
lebei is driving the quadricycle we rented on the toronto islands.	*my cruiser on 80 mile beach. messing with my 15mm zeiss*
this moto is still the best anything i've ever owned.	*biked to the nearest beach and took some pictures.*
you have no idea how much i love this thing.	*i was traveling alone, so i instead of me i had to take pictures of my bike*
↓	*...and she rides like a dream. initial impressions ride report*
[ROOT]	[ROOT]

(50)

MERU	CLIP
in a few of the branches of some trees were little tea-light lanterns	*in a few of the branches of some trees were little tea-light lanterns*
earth hour candle light	*they're only lanterns*
my first try making this photo with a mini lantern	*2nd shoot coming out too*
quant little light.	*not lit*
little lights.	↓
i actually detest gold, but i liked the material contrast here.	↓
i just really liked the light, alright?	↓
a pain to get photos of these.	↓
if you use my photo please post a link and let me know.	↓
[ROOT]	[ROOT]



(51)

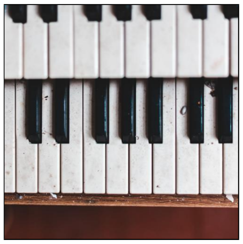
MERU	CLIP
simple scandinavian style decor. clifton, bristol	*ikea/helsinki design week party*
rental apartment in berlin	*with jeff from simple plan in denmark*
my flat getting more and more comfortable with more furniture coming in - and notice i now have fans!	↓
our first apartment	↓
ready made living space	↓
[ROOT]	[ROOT]

(52)

MERU	CLIP
studio di personaggio. character design.	*model: yasemin snoek stylist: melanie vink*
decorated with silver and nickle plated	*we bought one of these for a friend's wedding. no, not you julee.*
new faucet set.	↓
elegant bath items, though	↓
oh to have that kind of luxury	↓
a little luxury	↓
rich sigfrit kicks things off.	↓
the things i'll do for a shot	↓
can you guess what club this is for?!	↓
[ROOT]	[ROOT]

(53)

MERU	CLIP
*pentax *ist ds/ iso 1600 —— happy halloween!*	*jack-o-lantern with other light up decorations. jack-o-lantern.*
we wish you a happy all hallows night!	*i am so spoooooky*
happy hallowe'en, everyone!	↓
happy halloween, yo.	↓
no be long now jack	↓
can you feel the spirit of e-xtrategy?	↓
[ROOT]	[ROOT]

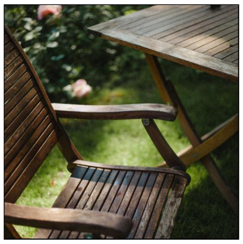


(54)

MERU	CLIP
piano keyboard	*vintage typewriter photo by rusty blazenhoff*
musical ben	*the typewriter is the best dead thing i ever found*
she was really good, great voice, excellent guitar playing, and really nice to chat to.	*musical harmony*
the highly-touted prospect, not the guitar player	*where some parts of me came of age*
l.a.r.g.e	↓
[ROOT]	[ROOT]

(55)

MERU	CLIP
my christmas shopping is done, and what's more, my presents are all wrapped!	*my homemade christmas book - dec. 5th and 6th.*
this year's wrapping job.	*gift-wrapped for any occasion.*
and gift wrapped!	*a wrapped xmas present*
21 presents for my 21st for msh may	*our christmas present put to good use.*
christmas is coming. won't someone think of my needs?	↓
the only good thing the guys did was dropped off the gifts.	↓
[ROOT]	[ROOT]



(56)

MERU	CLIP
table in the backyard of the summer house in melby, denmark.	*chair detail*
kitch at airbnb at corte del correggio - note window behind chairs	↓
photo by laura nawrocik some patio furniture that needs a little cleaning.	↓
from a bench on the north side.	↓
the backyard of the b and b we stayed at	↓
another from this shoot in a more traditional style.	↓
traditional place to take a picture.	↓
a nice place to take pictures!	↓
[ROOT]	[ROOT]

(57)

MERU	CLIP
found at city lights book store, sf www.citylights.com	*ikea catalog waiting for pickup, fairborn, ohio.*
in powells rare book room	*special collections - amsterdam, netherlands*
in the rare book reading room	*one needs to get there to read the it full.*
book heaven.	↓
not a book in sight	↓
even more books	↓
sorry... i really like this space.	↓
with something like this, you would have to get a few	↓
[ROOT]	[ROOT]

(58)

MERU	CLIP
a pineapple grows in the wild between goa gajah and yeh pulu, bali.	*so much organic burden on its way to the sea.*
the pineapple, dunmore park. n	*everything is so organic on lamu island.*
it started with more fruits but i didn't take this picture till late	*it's been that long since i took this that the next lot is already growing.*
shot <usr> beach	*the one that didn't get picked yet*
a regular sight from our coast	↓
la nature au carre.	↓
i would be happy if all my photos turned out like this one.	↓
this is one is good.	↓
[ROOT]	[ROOT]

(59)

MERU	CLIP
sulphur crested cockatoos are great characters	*soleil when she was a baby with her green feathers*
au pied du ciel	*lenny white)*
pale male's mate #5	*white tee's - photos for everyone*
i think this is a nice photo of sean, though i doubt he will think so.	↓
photo: george struikelblok	↓
this is birdy.	↓
q's and a's	↓
[ROOT]	[ROOT]

(60)

MERU	CLIP
uzi usb drive by dan helmick. brass, wood, riveted.	*i swear, this is how the coins landed. so i had to take a photo.*
the coin toss	*money was a fun picture to take.*
inset a coin...it moves...	*i really should have pictures of all the money, but after awhile one loses interest.*
ver. 2	↓
made in a post secret kinda a way to tell something.	↓
no-one will ever guess	↓
[ROOT]	[ROOT]