ContextCLIP: Contextual Alignment of Image-Text pairs on CLIP visual representations

Chanda Grover
Ashoka University
Sonipat, Haryana, India

Indra Deep Mastan
LNMIT
Jaipur, Rajasthan, India

Debayan Gupta
Ashoka University
Sonipat, Haryana, India

ABSTRACT

State-of-the-art empirical work has shown that visual representations learned by deep neural networks are robust in nature and capable of performing classification tasks on diverse datasets. For example, CLIP demonstrated zero-shot transfer performance on multiple datasets for classification tasks in a joint embedding space of image and text pairs. However, it showed negative transfer performance on standard datasets, e.g., BirdsNAP, RESISC45, and MNIST. In this paper, we propose ContextCLIP, a contextual and contrastive learning framework for the contextual alignment of image-text pairs by learning robust visual representations on Conceptual Captions dataset. Our framework was observed to improve the image-text alignment by aligning text and image representations contextually in the joint embedding space. ContextCLIP showed good qualitative performance for text-to-image retrieval tasks and enhanced classification accuracy. We evaluated our model quantitatively with zero-shot transfer and fine-tuning experiments on CIFAR-10, CIFAR-100, Birdsnap, RESISC45, and MNIST datasets for classification task.

KEYWORDS

Contrastive, Contextual, Alignment, Retrieval

1 INTRODUCTION

Learning robust image representations is an enduring problem in computer vision, as is the related question of making those representations transferable to some other dataset. Image representations also contain linguistic context, making a transfer to other domains a reasonable problem to pursue. No other benchmarks have been pursued as religiously as the ImageNet [10, 40] network architectures for a broad array of computer vision problems like transferring to new datasets [14, 39], object detection [20], image segmentation [4] etc. Recently, natural language-based supervision has been extensively researched for learning robust image representations, e.g., CLIP [37], ALIGN [21], and BASIC [36] have enjoyed great success in extending contrastive learning to paired image-text data, with impressive zero-shot classification and robustness. Visual self-supervision [35] and use of unpaired data [45] have been included in achieving such robust visual representations. Although these methods have achieved good performance in terms of positive zero-shot transfer on multiple datasets, they have shown negative zero-shot transfer on some datasets like MNIST, RESISC45, and BirdsNap.

Contrastive learning only exploits the high-level image-text pair similarity information but does not exploit semantic-level image-text pair similarity; it does not capture the contextual alignment of image and text pairs. Our work identifies contextual learning complements the contrastive learning objective for image-text alignment. Fig. 2 depicts the addition of contextual information to image-text pairs. As shown in Fig. 2a, CLIP enables the alignment of image-text pairs through contrastive learning. It only aligns I_1 with T_1 and not with T_2 and so on. Somewhat more formally, the image-text alignment problem can be described as follows: given a set of images and captions, determine the relation between image-text pairs based on the *contextual similarity*. The standard contrastive learning objective aims to identify the matched image text pairs ("positives") against the mismatched ("negatives") image-text pairs. [5, 46]. This, however, does not make the alignment of I_1 and T_1 at the semantic level. For example, Fig. 2b illustrates words such as orange and mango of captions T_1 are matched with regions corresponding to mango and orange in image I_1.

In this paper, we aim to learn robust image representation by studying image-text contextual alignment, i.e., given valid image-text pairs, we compute the image and text embedding that can be contextually aligned in the embedding space. Such an objective can be fulfilled when one can align the learned image and text embeddings at the context level. We hypothesize that because of the contextual non-alignment of the embeddings (of true image-text pairs), CLIP showed negative zero-shot transfer on standard datasets, e.g., BirdsNAP, RESISC45, and MNIST. To fix this, we propose ContextCLIP, a framework for the contextual alignment of image-text pairs based on the robust visual representation of CLIP using contextual loss [33]. Contextual loss aligns image-text pairs at the semantic level; specifically, given an image-text pair learned through contrastive learning, we augment the contrastive learning objective with a contextual learning objective at the feature level. The latter uses the nearest neighbor field to measure the similarity between images and texts. It considers a particular feature of one modality and identifies the "most similar" of the other. These features are matched in a projected space of image and text modalities (low-dimensional) and introduces contextual alignment at the feature level by reducing the distance between the image and text embeddings. We show that explicit formulation of contextual alignment objective improves the overall alignment between image-text pairs, as shown in Tables 1 and 2. In all cases, we pre-train our models on a small fraction of the Conceptual Captions 3M dataset [41] (around 6K image-text pairs) and evaluate them on CIFAR-10, CIFAR-100, MNIST, RESISC45, BirdsNap with zero-shot transfer and fine-tuning experiments on

Figure 1: The figure shows the output comparison between (a) CLIP [37] and (b) ContextCLIP (ours) for text-to-image retrieval task.

standard datasets. For qualitative performance evaluation, we performed text-to-image retrieval experiments; Fig. 1 and Fig. 5 show text-based image retrieval results for CLIP and ContextCLIP. The results of our proposed framework are visually better in terms of consistency with the textual descriptions.

Contributions. The contributions are summarized as follows.

- We analyze contrastive learning for representation learning jointly over image and text modalities and identify non-alignment between the representation space of image and text modalities of CLIP (Fig. 6 and Fig. 7).
- We propose ContextCLIP, a framework for contextual alignment of image-text space by employing contextual loss between non-aligned feature space of text and image embeddings (Sec. 3).
- We show that our approach learns a more robust visual representation than CLIP and shows positive zero-shot transfer on standard datasets where CLIP had negative zero-shot transfer performance. To show better representations obtained, we also performed fine-tuning experiments. (Table 1 and 2).

2 RELATED WORK

Multimodal data is learned in a way that can represent and summarize the complementary information of all modalities and remove redundant information. It can be achieved either through joint representations or coordinated representations. Collective representations have been used in state-of-the-art vision-language pre-training models [48]. Further, to identify the direct relations between sub(elements) from two or more different modalities, two types of alignments have been used i.e., implicit and explicit. We have used both types of alignment, implicit (i.e., contrastive) as well as explicit (i.e., contextual), to find the relation between sub(elements) of two modalities.

Contrastive representation learning. It was originally proposed

(a) Image-Text Alignment through CLIP [37]
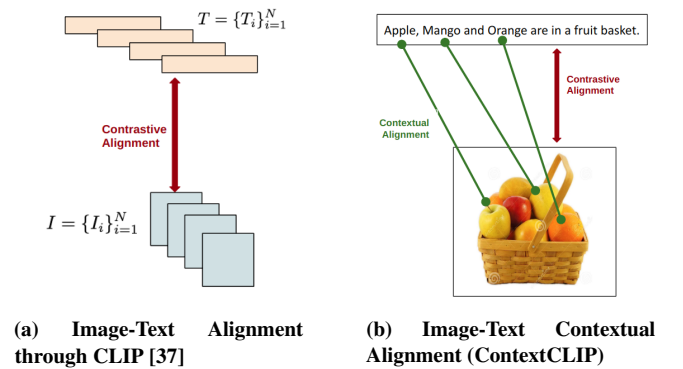
(b) Image-Text Contextual Alignment (ContextCLIP)

Figure 2: The figure (a) illustrate the contrastive alignment of the image-text pairs as performed by CLIP [37]. (b) shows the contextual alignment of the image-text pair as performed by ContextCLIP.

for self-supervised representation learning in the unimodal context where the embeddings of a sample are brought closer to an augmented version of the sample. In contrast, the embeddings are pushed away for other samples, and their augmentations [17] [5, 15, 16, 18, 46]. Some works like [51] impose additional constraints to remove redundancies and prevent dimensional collapse in the visual representations. Contrastive learning has also been used to learn robust representations of the multimodal data [38, 50].

The image and text multimodal data is encoded separately with their respective encoders, producing individual vision and language vectors embedded into a joint space using a contrastive loss. Contrastive approaches have the capability to learn a highly generic visual representation and have good strength to do classification on downstream tasks. Moreover, when pre-trained on a large and diverse dataset [21, 36] strong zero-shot vision-text retrieval and
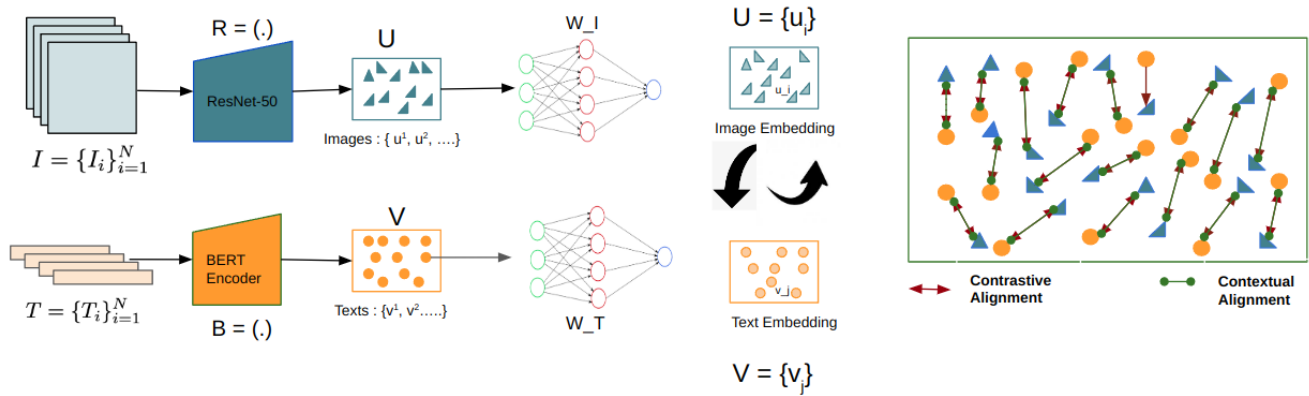
Figure 3: The figure shows the major components of ContextCLIP framework. It consists of image and text encoders with contextual and contrastive alignment objectives. The key idea is that contextual and contrastive alignment at embedding space helps to perform the image-Text alignment. The maroon arrow between the filled orange-colored circle (Text) and blue-colored triangle (Image) shows the contrastive alignment. The green arrow indicates the contextual alignment between Image-Text Pair.
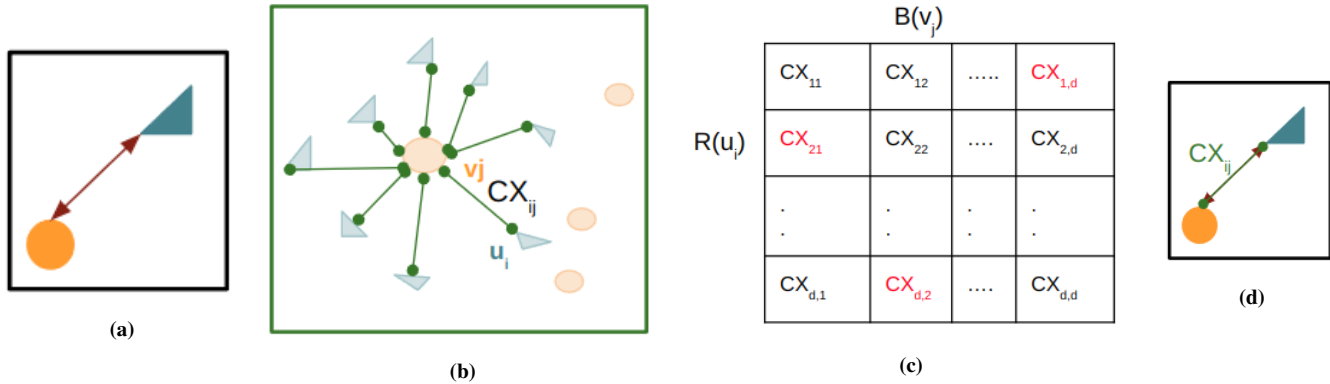

Figure 4: The figure shows contextual learning of Image-Text pairs. (a) Contrastive alignment of image (triangle) and text pair (circle) obtained through CLIP [37, 43, 49] (b) Many-to-One correspondence between text feature points (light orange colored circle v_j) and image feature points (light cyan colored triangle u_i). CX_{ij} is the contextual similarity between u_i and v_j. (c) Affinity matrix of contextual similarity between u_i and v_j. In each column, highlighted red color entry represents embedding u_i with the highest contextual similarity to the given v_j. (d) shows u_i and v_j are aligned contextually in addition to the contrastive alignment.

classification performance have been obtained. Unfortunately, as they are only trained to match visual data to text description, these models can only be adapted to close-ended tasks. Finally, it is challenging to adapt contrastive models using a handful of examples. In fact, Radford et al. (2021) indicated that using as few as two training examples per class actually decreased the CLIP zero-shot performance. But, Flamingo models have significantly improved with as few as four examples. Many works use additional losses to imbibe extra supervisory multimodal knowledge during the training process [12, 27, 42, 52]. In this work, we focus on having contextual loss[33] in addition to the contrastive loss to learn more robust and aligned image-text representations.

Contextual Learning. Contextual learning is useful for various applications such as image-to-image transformation tasks [33], style transfer [9, 31] and image restoration [29, 32]. Mastan et al. have shown contextual learning is also useful for image enhancement when training with limited training samples [7, 8, 28, 30]. The key challenge in image-to-image transformation tasks comes in the presence of non-aligned image data, where employing a pixel-based loss function will be less effective. Contextual learning on images first extracts contextual features using a pre-trained feature extractor and then compares the extracted representations using the contextual loss. The critical observation here is that similar features will be closer in the embedding space. Therefore, the contextual loss will be able to match the feature even in the presence of the non-alignment of the image data. We employ contextual loss on the pre-trained embeddings of the image and text pair captured from the feature extractor.

Image-Text alignment Models. In recent years rapid progress has been made in vision-language pre-training (VLP), which aims to jointly encode vision and language in a fusion model. Image-Text alignment models subsume the vision-language pre-training. CLIP [37] and ALIGN [21] demonstrate that dual encoder models pre-trained with contrastive objectives on image-text pairs can learn the strong image and text representations for cross-modal alignment task and zero-shot classification. Contrastive dual approaches [22] rely on a relatively similar procedure with the steps as follows i) Extract discriminative image features using a deep neural network ii) Extract text features using another deep neural network iii) use a loss function that measures as accurately as possible the distance between two embedding. These models align image-text pairs only at the data point level but not at the level of semantics of image and text. In this paper, we align the context of image and text by incorporating contextual loss at the joint image-text embedding space.

3 OUR APPROACH

We propose ContextCLIP framework for the alignment of contextual features extracted from image-text pair to enhance the alignment between image and text pair. The overview of our proposed framework is illustrated in Fig. 3. We start with feeding an image and its corresponding text description to image encoder and text encoder respectively to get their encoded representations. These joint image-text representations are then projected to low-dimensional space through the projection head. There, we represent each image and text having the same size subspace. We then measure the similarity between the image and text pairs as a similarity measure between these point sets. Contextual loss is computed at this point, which makes contrastively aligned image-text pairs to be aligned at the feature level.

3.1 Model Architecture

Our models use the same architecture as the original CLIP model presented in [37]. In addition to CLIP architecture, we have introduced a projection head for image (W_I) after image and text (W_T) as shown in Fig. 3 to bring different dimensional representational sizes of image and text to be of the same sizes.

Image Encoder. We used ResNet-50 architecture [19] as the image encoder (R) for learning image representations as shown in Fig. 3. We considered image representations of 2048 size. All images are resized to 224×224.

Text Encoder. It learns the feature representations from natural language description, which is input to the text encoder as shown in Fig. 3. For the text encoder (B), we used the pre-trained BERT model [13]. It is a deep language model that leverages Transformer architecture to learn the contextual relations between words in a textual description. During training, it also uses a word masking mechanism that masks 15% of the words with a token and enables BERT to learn the robust embeddings. We considered text representations of 768 size.

Projection Head. Once image and text representations of 2048 and 768 dimensions respectively are obtained, we projected them to the same subspace of size 256 through the projection head. The incorporation of contextual loss on top of contrastive loss requires the

two non-aligned sub-spaces to be on the same embedding subspace. Therefore, we have used the Projection Head for these representations to project them onto the space. We used image and text projection layers as W_I and W_T respectively, as shown in Fig. 3 with an output dimension d = 256. We followed a change introduced by [1] i.e., not to use the non-linear projection between the representation and the contrastive embedding space, instead use only a linear projection to map from each encoder's representation to the multi-modal embedding space.

3.2 Loss Function

This section describes the loss function used in ContextCLIP. The total loss is calculated as:

$$\mathcal{L} = \mathcal{L}_{CLIP} + \alpha \times \mathcal{L}_{CX} \tag{1}$$

Here, α is a constant with value of 0.5. $\mathcal{L}_{CLIP}$ is the contrastive loss computed on image and text representations of size 2048 and 768, respectively, to bring similar image-text pairs together and dissimilar image-text pairs farther apart. $\mathcal{L}_{CX}$ is the contextual loss calculated through contextual similarity defined in Sec. 3.2.2 and is applied at the individual features of contrastive aligned image-text pair.

3.2.1 **CLIP Loss** ($\mathcal{L}_{CLIP}$): Our work is most closely related to Contrtastive Language-Image pre-training (CLIP) [37]. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. An overview of our method is illustrated in Fig. 3. At a high-level view, our method first converts each of the images and text to its equivalent representations u and v by passing them through pre-trained image and text encoders. Image representations (v) are of dimension 2048, and the text representations (u) are of dimension 768. Image and text representation are connected through two contrastive losses at the data point level. The first loss function is an image-to-text contrastive loss [53] for the ith pair:

$$\ell_i^{(v \to u)} = \log \frac{\exp \langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau}{\sum_{k=1}^{N} \exp \langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau} \tag{2}$$

where, $\langle \mathbf{v}_i, \mathbf{u}_i \rangle$ represents the cosine similarity between image and text representations, i.e., $\langle \mathbf{v}, \mathbf{u} \rangle = \mathbf{v}^\top \mathbf{u} / ||\mathbf{v}|| ||\mathbf{u}||$; and $\tau \in \mathbb{R}^+$ represents a temperature parameter. This loss is InfoNCE [46] that tries to predict $\langle \mathbf{v}_i, \mathbf{u}_i \rangle$ as the true pairs and maximizes the mutual information between image and text representations. Hence, similar image-text pairs are put together, and the dissimilar image-text representations are put apart farther from each other. Here, contrastive loss is used between inputs of different modalities. Therefore, a similar text-to-image contrastive loss [53] is defined as follows:

$$\ell_i^{(u \to v)} = \log \frac{\exp \langle \mathbf{u}_i, \mathbf{v}_i \rangle / \tau}{\sum_{k=1}^{N} \exp \langle \mathbf{u}_i, \mathbf{v}_k \rangle / \tau} \tag{3}$$

Thus, final training loss is a weighted average of image-to-text and text-to-image loss over all positive image-text pairs in each minibatch:

$$\mathcal{L}_{CLIP} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \ell_i^{(v \to u)} + (1 - \lambda) \ell_i^{(u \to v)} \right) \tag{4}$$

where $\lambda \in [0,1]$ is a scalar weight. The maroon arrow in Fig. 4a shows the contrastive alignment between the image and text pair.

3.2.2 **Contextual loss** $\mathcal{L}_{CX}$.
The connected representations learned through CLIP [37] loss are projected to 256-dimensional space, where they are brought closer together at the feature level. Contextual loss is the loss function targeted at non-aligned data. i.e., it does not require the two domains to be spatially aligned. It is based on both context and semantics and compares the regions of image and text with similar meanings while considering the context of the entire image. Our idea is to consider a batch of images and texts as a collection of features and measure the contextual similarity between image and text features based on the similarity between their features.

Contextual Similarity: We define a measure of similarity between a pair of text and image. To accomplish this, we represent each image U encoded by ResNet [19] and each text V encoded by BERT [13] as a set of low-dimensional feature points in the same subspace via their projection heads W_I and W_T, as shown in Fig. 3. Feature points of image and text are represented as $U = \{u_i\}$ and $V = \{v_i\}$ as shown in Fig. 3. We assume $|U| = |V| = 256$ in a batch of image-text pairs. To calculate the similarity between the images and texts, we find for each v_j, the feature u_i that is most similar to it, and then sum the corresponding feature similarity values overall v_j. Formally, the contextual similarity CX_{ij} between U and V is defined as follows.

$$CX(U,V) = \frac{1}{N} \sum_j \max_j CX_{ij} \qquad (5)$$

where, CX_{ij} is the similarity between features u_i and v_j.

The contextual similarity CX_{ij} between points at the embedding space of image-text modalities helps for contextual cross-modal alignment by encouraging the distance between the image and text embedding to be close. Specifically, we consider feature u_i as contextually similar to feature v_j, if it is significantly closer to it than to all other features in U. If this is not the case, i.e., u_i is not more similar to any specific v_j, then its contextual similarity to all y_j should be low. This method is resistant to distance scales, i.e., if u_i is far from all v_j, then CX_{ij} will be low for all j.

On Image-Text latent space: When the image and text pairs are similar, a one-to-one mapping exists between all the feature points. When they are dissimilar, a many-to-one mapping exists between text and image. A many-to-one mapping exists between the image and text pair, as shown in Fig. 4b. This indicates the image-text pair are dissimilar in nature. We measure contextual similarity and aim to find a one-to-one mapping between image and text pairs to make them similar at the feature level to make them similar. The distance between image and text is a function of the near field. Contextual learning objectives use the nearest neighbor field in order to measure the similarity between images and texts. To avoid the need for geometric alignment between two modalities, it considers a particular feature of one modality and identifies the most similar features of the other modality. This yields the nearest neighbor field of matches between two modalities. The green arrow represents the nearest neighboring field. The loss function will use the green arrows and their associated weights. We start by computing a full affinity

matrix between the point sets representing the images and texts, as shown in Fig. 2b. Taking the maximum over each affinity yields the nearest neighbor field. To incorporate the context of each point, we use contextual affinity. We consider the context by normalizing the affinities in a softmax manner. The normalization is done in each row of the affinity matrix. Thus, we measure the similarity between two points representing an image and text using contextual affinity.

Matches between features are formed by considering all the features in the batch of images. This incorporates the cross-modal global image context into the similarity measure. The similarity between images and texts is then defined based on the similarity between the matched features. This allows the aligned image to spatially deform with respect to the text description. Let d_{ij} represent the cosine similarity of u_i and v_j. When $d_{ij} << d_{ik}$ and for all $k \neq j$, we consider features u_i and v_j to be similar. To begin, we normalize the distances, as done by [33]:

$$\tilde{d}_{ij} = \frac{d_{ij}}{\min_k d_{ik} + \epsilon} \qquad (6)$$

for a fixed $\epsilon = 1e-5$. We shift from distances to similarities by exponentiation:

$$w_{ij} = \exp\left(\frac{1 - \tilde{d}_{ij}}{h}\right) \qquad (7)$$

where $h > 0$ is a band-width parameter. Finally, the contextual similarity between features is defined to be a scale invariant version of the normalized similarities. In the extreme case, $CX(X,Y)$ is equivalent to counting how many features in V are the nearest neighbors of a feature in U, which is exactly the template matching measure proposed by [44].

$$CX_{ij} = w_{ij} / \sum_k w_{ik} \qquad (8)$$

Therefore, contextual loss between image-text space is defined as

$$\mathcal{L}_{CX}(u,v,d) = -\log\left(CX\left(R^d(u), B^d(v)\right)\right) \qquad (9)$$

The optimal solution to the contrastive loss formulation for CLIP would push the similarity between the normalized embeddings of the matched pairs towards one while forcing all other pairs of similarities to zero. Thus, aligning the cross-modal similarity matrix and minimizing the cross-modal alignement loss. However, this idealized scenario does not occur in practice, and we find that explicit alignment via contextual loss in ContextCLIP facilitates improved learning, as we show in our experiments. The pseudo-code for our approach is present in the supplementary material.

4 EXPERIMENTAL RESULTS

We performed zero-shot transfer, fine-tuning, and text-to-image retrieval experiments on the CLIP and ContextCLIP methods. In a zero-shot transfer experiment, we observe that ContextCLIP improves top-1 accuracy over CLIP by 1 on CIFAR-10, by 0.5 on CIFAR-100, and by 2 on the BirdsNap dataset. ContextCLIP also outperforms CLIP in fine-tuning experiments, with an average relative gain of top-1 accuracy of +8 on CIFAR-10, +3 on RESISC45, and +1.5 on BirdsNap datasets. We also observed improved structure of embeddings on ContextXLIP in the representation space when projected to 2 dimensional space with t-SNE [47] and discovered
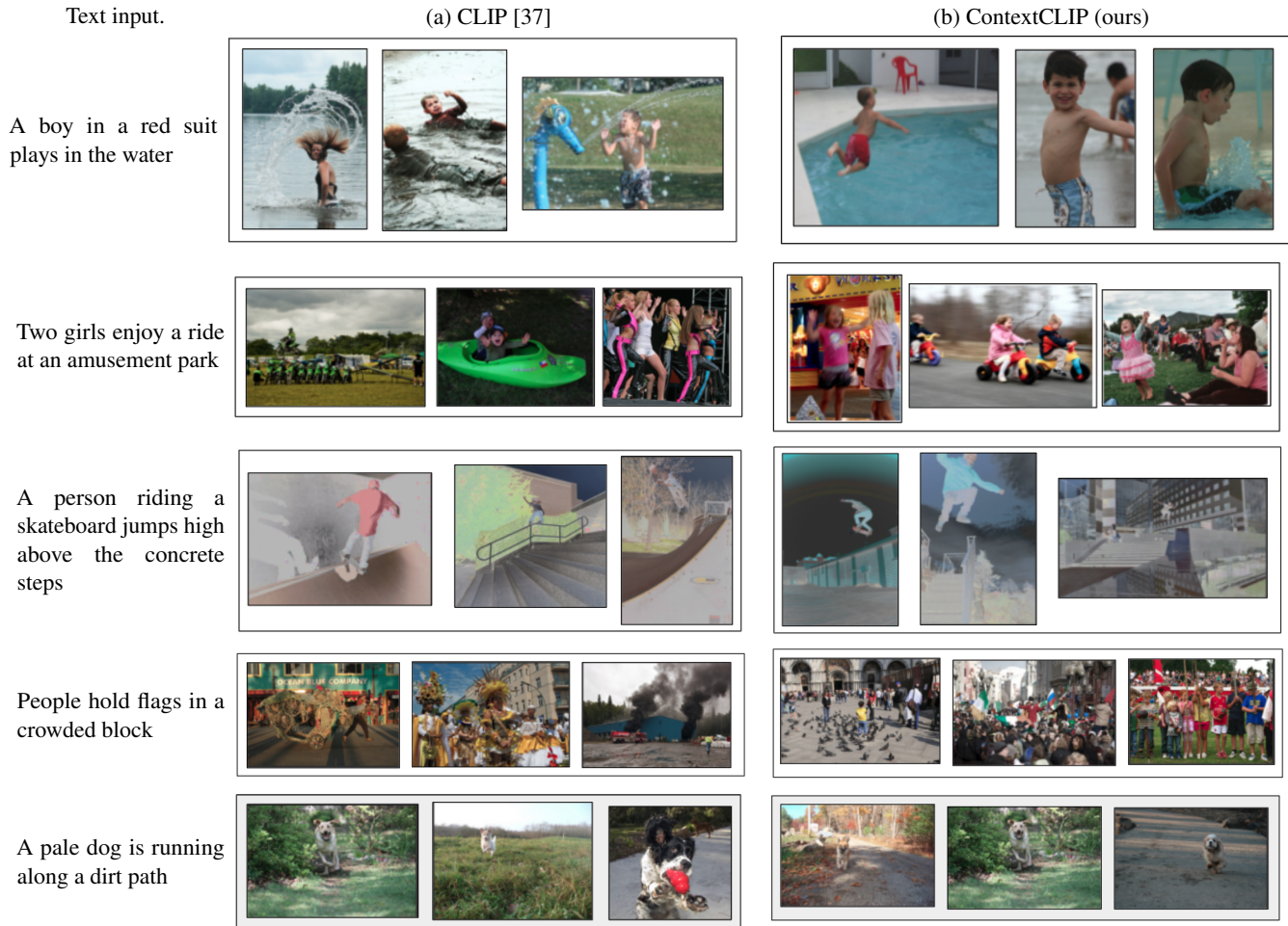
Text input.

(a) CLIP [37]

(b) ContextCLIP (ours)

A boy in a red suit plays in the water

Two girls enjoy a ride at an amusement park

A person riding a skateboard jumps high above the concrete steps

People hold flags in a crowded block

A pale dog is running along a dirt path

Figure 5: Text-to-image Retrieval examples. (a) Represents the retrieval from CLIP [37] model. (b) Represents the retrieval from ContextCLIP model.

that representations with contextual loss are more structured than representations only with contrastive loss.

4.1 Evaluation

A zero-shot image classification task (Sec. 4.2) has been performed for both CLIP [37] and ContextCLIP, i.e., classifying test images into categories not seen at training time. We have used top-1 and top-5 accuracy to report the accuracy of predictions of classes given an image and a text template. We follow the evaluation strategy as suggested by [37] for zero-shot classification using prompt engineering. We used the following datasets for evaluating zero-shot transfer and fine-tuning experiments : CIFAR-10 [25], CIFAR-100 [25], MNIST [11], RESISC45 [6], BirdsNap [2]. For each dataset, we use the names of classes to form a set of natural sentences such as 'a photo of the class name', 'a sketch of the class name' and more. Then, the similarity of the test image to each text template is computed (e.g., cosine distance), and the model predicts the category for which the image-caption similarity is highest. For evaluation of

representations obtained, we performed fine-tuning experiments as mentioned in Sec. 4.3

4.2 Zero-Shot Transfer

For this experiment, we have trained our model on the Conceptual Captions [41] dataset. The complete details of the dataset and experimental setup are mentioned in Sec. 6. There is a pre-defined text template corresponding to each dataset as adopted from [37]. These templates are passed through text encoder to get a set of text embeddings for that class. To obtain a single text embedding for that class, this set of text embeddings is ell_2-normalized, averaged, and further ell_2-normalized. For a given image, the image embedding is obtained by passing it through an image encoder as defined in Sec. 3. The class whose text embedding is closest to the test image is taken to be the predicted label. The top-1 and top-5 accuracies of zero-shot performance are presented in Table 1.

CLIP [37] showed positive zero-shot transfer on CIFAR-10 and CIFAR-100 but negative zero-shot transfer on BirdsNap, RESISC45, and MNIST. Unlike CLIP, our method ContextCLIP showed positive

zero-shot transfer performance on the BirdsNap (+1.91) and MNIST (+6.16) datasets in addition to CIFAR-10 (+1.03) and CIFAR-100 (+0.23) but negative zero-shot transfer performance on RESISC45 (-1.93). We reason that the positive zero-shot transfer of ContextCLIP on BirdNap and MNIST is because of the contextual loss in addition to the contrastive objective, which helps to align context (word) information from text description to the style information of the image. For RESISC45, ContextCLIP has shown negative zero transfer because of not being able to capture the alignment information from textual description and image. Among all of the evaluated datasets, MNIST has shown the most positive zero-shot transfer. We reason that this is because of the simple alignment relation between image and text space, as opposed to BirdsNap, CIFAR-10, and CIFAR-100, which have comparatively complex attributes and hence complex alignment relations.

4.3 Fine-tuning

The representation capabilities of our model ContextCLIP are evaluated through fine-tuning experiment. There are many ways to evaluate the quality of representations, e.g., One approach involves fitting a linear classifier on a representation extracted from the model and measuring its performance on various datasets. An alternative to this is measuring the performance of end-to-end fine-tuning the model. We adopted fine-tuning of the model because it increases the flexibility, and prior work has demonstrated that fine-tuning outperforms linear classification on most image classification datasets [24]. We conducted fine-tuning experiments of the pre-trained representations obtained from the image encoder and text encoder i) only with contrastive loss and ii) contextual loss in addition to contrastive loss. The results are shown in the Table 2. We observe that top-1 classification accuracy improved on all datasets with the addition of contextual loss.

We observe that the maximum improvement in this experiment is obtained on CIFAR-10 (+7.94). Followed by this is RESISC45 (+2.55); CIFAR-100 (+.22); MNIST (+1.69); and BirdsNap (+1.64). CIFAR-10 showed maximum improvement as compared to other datasets because it has only 10 class categories. MNIST also has the same number of classes, but because of the very simple dataset and the complex formulation of the model, it is not able to learn well. Improvement in the evaluated datasets is shown in below order, i.e., CIFAR-10 RESISC45, which has 45 classes, followed by CIFAR-100 (45 classes), and BirdNap (200 classes). We observed that with the increase in the number of classes in datasets, the quality of representations learned from the supervision of other modalities reduces.

4.4 Text-to-Image Retrieval

For qualitative performance evaluation, we performed text-to-image retrieval tasks, i.e., given a text description, we retrieved a set of images with the CLIP and ContextCLIP methods, respectively. As shown in Fig. 5, the figures present on the left hand side are the images retrieved with the CLIP and on the right hand side are the images retrieved from ContextCLIP. We found that with the ContextCLIP method, the retrieved images are more semantically matched with the text description as compared to the CLIP method. e.g., corresponding to the text "A boy in red suit plays in the water,"

images retrieved with ContextCLIP contain either complete text information or most parts of text in the image. In the CLIP output, it does not show the boy in a red suit. In one of the retrievals, it shows two kids playing in water. Such specific number and color information is missing in the CLIP method, but ContextCLIP is capable of capturing such minute details of the text. Similarly, considering the other text examples like "People hold flags in a crowded block. Here, CLIP has not retrieved any images where either people or a flag can be seen in any image. As opposed to this, ContextCLIP has retrieved somewhat relevant results here. In some of the images, we can see people holding flags, a crowd somewhere, etc.

5 ABLATION STUDIES

We projected high-dimension representations of image and text embeddings (256) obtained from their respective encoders (R and B) and followed by their projection heads (W_I and W_T) to reduced-dimensional space. Low-dimensional projection is of size 2 and is done with t-SNE [47] for CIFAR-10 and MNIST with and without contextual loss as shown in Fig. 6 and Fig. 7. For both the datasets, image embeddings are more structured as compared to text embeddings. This is because, through ContextCLIP, we only aim to modify and learn robust image representations through natural language supervision. Once the image and text representations are obtained from the projection head, text representations do not get modified. With the addition of contextual loss, only image embeddings are updated with respect to the content and style of a given text description.
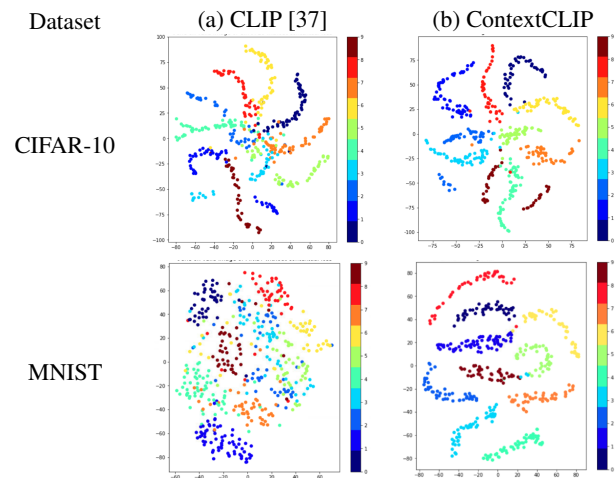

Figure 6: Image Embeddings. The top and bottom rows are the embeddings of CIFAR-10 and MNIST, respectively. a) The image embeddings without contextual loss are shown in the left column; b) The image embeddings with contextual loss are shown in the right column.

As shown in Fig. 6, image embeddings clearly visualises all of the classes for both the datasets with the different colours. Image embedings of all classes are more clustered with ContextCLIP than CLIP. As opposed to image embeddings, there is no structured information in text embeddings 7, neither in CLIP nor in ContextCLIP, but still comparatively better clustering information is available with

Table 1: The table shows quantitative performance for Zero-shot Transfer on standard datasets.

Dataset	Top-1 Accuracy		Top-5 Accuracy	
	CLIP [37]	ContextCLIP (ours)	CLIP [37]	ContextCLIP (ours)
CIFAR-10	9.37	**10.40**	49.77	**50.14**
CIFAR-100	0.92	**1.15**	4.34	**5.32**
RESISC45	3.41	1.48	12.19	11.05
BirdsNap	8.5	**10.41**	41.11	**46.86**
MNIST	12.26	**18.42**	50.90	**59.82**

Table 2: The table shows quantitative performance for Fine-Tuning on standard datasets.

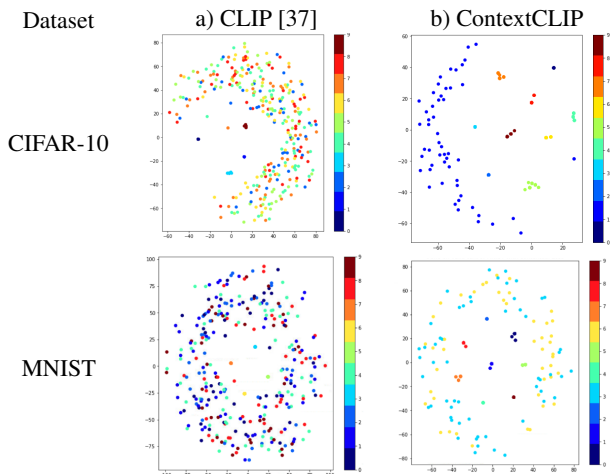| Dataset	Top-1 Accuracy		Top-5 Accuracy	
	CLIP [37]	ContextCLIP (ours)	CLIP [37]	ContextCLIP (ours)
CIFAR-10	2.46	**10.40**	60.78	**68.14**
CIFAR-100	1.00	**1.22**	5.37	**7.24**
RESISC45	82.83	**85.38**	98.30	**98.37**
BirdsNap	14.25	**15.89**	35.9	**43.85**
MNIST	92.78	**94.47**	99.12	**99.64**

Figure 7: Text Embeddings. The top and bottom rows are the embeddings of CIFAR-10 and MNIST, respectively. a) The left column shows text embeddings that have no contextual loss; b) the right column shows image embeddings that have contextual loss.

ContextCLIP than CLIP. We tested text embeddings though BERT [13] as well as SimCSE[15]. In the figure 7, we have shown the results with SimCSE. In SimCSE, text representations are learned in an unsupervised way.

6 EXPERIMENTAL SETUP

Training Details. We have used [34] code for producing all tables and figures of our work. For contrastive learning, we used a temperature value of $\tau = 1.0$ and a loss weight of $\lambda = 0.75$. We used the Adam optimizer [23] with a weight decay of 1e-3. A learning rate of 1e-4 and 1e-6 was used for the image encoder and text encoder respectively. We use a fixed batch size of 32. We train the model for 30 epochs. The network is trained jointly to minimise the distance between the modality-specific embeddings.

Datasets. We used Conceptual Captions 1M [41] (CC1M) image-caption pairs as the source of image and text pre-training data for zero-shot transfer experiments. Although this dataset is smaller than the custom dataset (400 million pairs) used in the original work on CLIP [37], it is suitable for our available data and compute. Many subsequent works on language-image pre-training have used CC1M for benchmark evaluations[3, 26, 35, 45].

Fine-tuning: For fine-tuning purposes, we trained variants of the ContextCLIP model end-to-end with the CIFAR-10 [25], CIFAR-100 [25], MNIST [11], RESISC45 [6], and BirdsNap [2] datasets. We considered 80–20–10% as the training, validation, and testing configurations. We considered the same datasets for evaluating the results of zero-shot transfer experiments.

7 CONCLUSION

We presented ContextCLIP, a framework for contextual alignment of image-text pairs on CLIP visual representations. We studied the misalignment of image and text representations on CLIP representations. The main benefit of our method comes from incorporating contextual alignment into the joint image-text embedding space in addition to contrastive alignment. Empirically, we showed that ContextCLIP

performs better than CLIP on zero-shot classification on datasets like MNIST, CIFAR-10, CIFAR-100, BirdsNap and RESISC45 where CLIP had shown negative zero-shot transfer performance. For future work, we aim to evaluate our proposed method on ImageNet and its variants. We also showed that the representations learned by ContextCLIP are consistent with the concept-level knowledge of text description, as evidenced by text-to-image retrieval tasks. Also, with the fine-tuning experiments, we observed more robustness in visual representations than CLIP. We believe this work can motivate further studies on identifying conditions and regularization strategies under which the learned representations are synergistic across the image-text modalities for downstream applications. One significant future direction for ContextCLIP is to extend the work from text-to-image retrieval tasks to text-to-image style transfer and manipulation tasks.

REFERENCES

[1] Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning Representations by Maximizing Mutual Information Across Views. (2019), 15509–15519. https://proceedings.neurips.cc/paper/2019/hash/ddf354219aac374f1d40b7e760ee5bb7-Abstract.html

[2] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. 2014. Birdsnap: Large-Scale Fine-Grained Visual Categorization of Birds. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2019–2026. https://doi.org/10.1109/CVPR.2014.259

[3] Nicholas Carlini and Andreas Terzis. 2022. Poisoning and Backdooring Contrastive Learning. (2022). https://openreview.net/forum?id=iC4UHbQ01Mp

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 4 (2018), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607. http://proceedings.mlr.press/v119/chen20j.html

[6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* 105, 10 (2017), 1865–1883.

[7] Indra Deep Mastan and Shanmuganathan Raman. 2019. Multi-Level Encoder-Decoder Architectures for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[8] Indra Deep Mastan and Shanmuganathan Raman. 2020. DeepCFL: Deep Contextual Features Learning from a Single Image. *arXiv e-prints* (2020), arXiv–2011.

[9] Indra Deep Mastan and Shanmuganathan Raman. 2020. DeepObjStyle: Deep Object-based Photo Style Transfer. *arXiv e-prints* (2020), arXiv–2012.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[11] Li Deng. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Process. Mag.* 29, 6 (2012), 141–142. https://doi.org/10.1109/MSP.2012.2211477

[12] Karan Desai and Justin Johnson. 2021. VirTex: Learning Visual Representations From Textual Annotations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 11162–11173. https://doi.org/10.1109/CVPR46437.2021.01101

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[14] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014 (JMLR Workshop and Conference Proceedings, Vol. 32)*. JMLR.org, 647–655. http://proceedings.mlr.press/v32/donahue14.html

[15] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. (2020). https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html

[17] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010 (JMLR Proceedings, Vol. 9)*, Yee Whye Teh and D. Mike Titterington (Eds.). JMLR.org, 297–304. http://proceedings.mlr.press/v9/gutmann10a.html

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[20] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. 2017. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 3296–3297. https://doi.org/10.1109/CVPR.2017.351

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 4904–4916. http://proceedings.mlr.press/v139/jia21b.html

[22] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Cascade Attention Network for Person Search: Both Image and Text-Image Similarity Selection. *CoRR* abs/1809.08440 (2018). arXiv:1809.08440 http://arxiv.org/abs/1809.08440

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. (2015). http://arxiv.org/abs/1412.6980

[24] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. 2019. Do Better ImageNet Models Transfer Better?. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2661–2671. https://doi.org/10.1109/CVPR.2019.00277

[25] Alex Krizhevsky et al. 2009. Learning multiple layers of features from tiny images. (2009).

[26] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. (2022). https://openreview.net/forum?id=zq1iJkNk3uN

[27] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2021. Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis. *CoRR* abs/2109.01797 (2021). arXiv:2109.01797 https://arxiv.org/abs/2109.01797

[28] Indra Deep Mastan and Shanmuganathan Raman. 2019. Multi-level encoder-decoder architectures for image restoration. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1728–1737.

[29] Indra Deep Mastan and Shanmuganathan Raman. 2020. Dcil: Deep contextual internal learning for image restoration and image retargeting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2366–2375.

[30] Indra Deep Mastan and Shanmuganathan Raman. 2021. DeepCFL: Deep contextual features learning from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2897–2906.

[31] Indra Deep Mastan and Shanmuganathan Raman. 2021. Deepobjstyle: deep object-based photo style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 711–719.

[32] Indra Deep Mastan, Shanmuganathan Raman, and Prajwal Singh. 2022. DILIE: deep internal learning for image enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 24–33.

[33] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-aligned Data. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV (Lecture Notes in Computer Science, Vol. 11218)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 800–815. https://doi.org/10.1007/978-3-030-01264-9_47

[34] moein shariatnia. 2021. OpenAI-CLIP. https://github.com/moein-shariatnia/OpenAI-CLIP. [Online; accessed Mar, 2022].

[35] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. 2021. SLIP: Self-supervision meets Language-Image Pre-training. *CoRR* abs/2112.12750 (2021). arXiv:2112.12750 https://arxiv.org/abs/2112.12750

[36] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. 2021. Combined Scaling for Zero-shot Transfer Learning. *CoRR* abs/2111.10050 (2021). arXiv:2111.10050 https://arxiv.org/abs/2111.10050

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. http://proceedings.mlr.press/v139/ramesh21a.html

[39] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 512–519. https://doi.org/10.1109/CVPRW.2014.131

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 2556–2565. https://doi.org/10.18653/v1/P18-1238

[42] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. FLAVA: A Foundational Language And Vision Alignment Model. *CoRR* abs/2112.04482. arXiv:2112.04482 https://arxiv.org/abs/2112.04482

[43] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. LightningDOT: Pre-training Visual-Semantic Embeddings for Real-Time Image-Text Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 982–997. https://doi.org/10.18653/v1/2021.naacl-main.77

[44] Itamar Talmi, Roey Mechrez, and Lihi Zelnik-Manor. 2017. Template Matching with Deformable Diversity Similarity. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 1311–1319. https://doi.org/10.1109/CVPR.2017.144

[45] Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. 2021. A Fistful of Words: Learning Transferable Visual Models from Bag-of-Words Supervision. arXiv:2112.13884 https://arxiv.org/abs/2112.13884

[46] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018). arXiv:1807.03748 http://arxiv.org/abs/1807.03748

[47] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[48] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *CoRR* abs/2205.01917 (2022). https://doi.org/10.48550/arXiv.2205.01917 arXiv:2205.01917

[49] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen

Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021. Florence: A New Foundation Model for Computer Vision. *CoRR* abs/2111.11432 (2021). arXiv:2111.11432 https://arxiv.org/abs/2111.11432

[50] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal Contrastive Training for Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 6995–7004. https://doi.org/10.1109/CVPR46437.2021.00692

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12310–12320. http://proceedings.mlr.press/v139/zbontar21a.html

[52] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 5579–5588. https://doi.org/10.1109/CVPR46437.2021.00553

[53] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *CoRR* abs/2010.00747 (2020). arXiv:2010.00747 https://arxiv.org/abs/2010.00747

ContextCLIP: Contextual Alignment of Image-Text pairs on CLIP visual representations
(Supplementary Material)

```
# Pseudo-code for ContextCLIP
1. encoder_image - ResNet Model
2. encoder_text - BERT Model
3. I[n,h,w,c] - minibatch of images
4. T[n,l]        - minibatch of texts
5. W_I[d_i,d_e] - learned projection of image to embed
6. W_T[d_t,d_e] - learned projection of text to embed
7. t                      - learned temperature parameter
# extract feature representations of each modality
8. I_f = encoder_image(I) - [n,d_i]
9. T_f = encoder_text(T) - [n,d_t]

# joint multimodal embedding [n, d_e]
10. I_e = l_2normalize(np.dot(I_f, W_I), axis = 1)
11. T_e = l_2normalize(np.dot(T_f, W_T), axis = 1)

# scaled pairwise cosine similarities [n, n]
12. logits = np.dot(I_e, T_e.T) * np.exp(t)

# contrastive loss function
13. labels = np.arange(n)
14. loss_i = crossEntroLoss(logits, labels, axis=0)
15. loss_t = crossEntroLoss(logits, labels, axis=1)
16. loss_contrastive = (loss_i + loss_t)/2

# contextual loss function
17. loss_contextual = contextualLoss(I_e, T_e)

18. loss = loss_contrastive + lambda × loss_contextual
```

Algorithm 1. This describes sequence of steps taken to add contextual loss on CLIP joint image-text embedding space.

This details of Algorithm 1 for the proposed ContextCLIP is described as follows.

- Line 1 and 2 are the ResNet image encoder and BERT text encoder.
- Line 3 and 4 are the mini-batch of size n of images (I) and texts (T).
- Line 5 and 6 are the projection heads for image (W_I) and text (W_T) which projects the different size image (d_i) and text (d_t) representations to the same size (d_e).
- Line 7 is the temperature value used for calculating the contrastive loss between image and text pairs.
- Line 8 and 9 are the image (I_f) and text (T_f) representations obtained from the image-encoder and text-encoder respectively.
- Line 10 and 11 are the normalised image (I_e) and text (T_e) embeddings of their representations obtained by dot product of image and text representations with their projection heads.
- Line 12 calculates the logits of image and text embeddings i.e. it calculates the image and text similarity with a specific temparture value.
- Line 13 defines the true labels of all images.
- Line 14 and 15 calculates the cross entropy loss $loss_i$ and $loss_t$ between image and text embeddings based on true and predicted image-text pairs.
- Line 16 calculates the contrastive loss between image and text embeddings.
- Line 17 calculates the contextual loss at the semantic level of image and text embeddings.
- Line 18 calculates the sum of contrastive and contextual loss. This completes the total loss calculation for a mini-batch of texts and images pair with the weightage of $\lambda = [0 - 1]$.

The same sequence of steps is performed for another batch of images and so on. This code was iterated for 25 epochs on training datasets: Conceptual Captions for zero-shot transfer experiments; CIFAR-10, CIFAR-100, MNIST, BirdsNap, RESISC45 for fine-tuning experiments; Flickr-8K for text-to-image retrieval experiments.