Weakly-supervised Spatially Grounded Concept Learner for Few-Shot Learning

*Gaurav Bhatt[1], *Deepayan Das[2], Leonid Sigal[1], Vineeth N Balasubramanian[2]
[1]The University of British Columbia, Vancouver
[2]Indian Institute of Technology, Hyderabad

Aim

- We propose a **visually grounded concept learner** (VG-CoL) that enforces semantic structure over spatial representations, overcoming limitations of existing methods[1,2] that either **lack semantics** or **strong supervision**
- Introduce a regularization technique to ensure learned **concepts** are **semantic**, **disentangled**, and aligned with weights of image-level concept/attribute classifiers

Concepts

Fig: Images showing different species of animals and birds and the various traits that help us identify them.

- Humans quickly identify species using key features like stripes or crown colors, eliminating the need for countless examples.
- In this work we seek to localize semantic concepts in images, aiming to enhance deep networks' efficiency in classification with minimal examples.

Problem Formulation

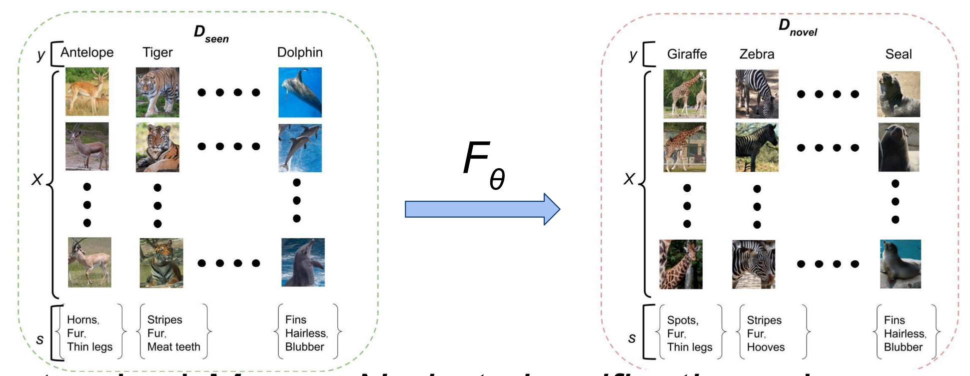
Tasks:

1. Train a model on a dataset with abundant annotations from *seen* classes D_{seen}
2. Adapt it to samples from a disjoint set of novel classes D_{novel} with limited labels

Overview:

1. Learn feature extractor F_θ on D_{seen} by minimizing cross entropy loss over seen classes

*Fig: Feature extractor F_θ Is trained on D_{seen} and adapted to D_{novel}. We use attributes **s** to ensure that model learns to identify and localize concepts for better **transferability** and **interpretability***



2. For Inference, we use the standard *M-way, N-shot classification*, where we minimize the distance between query and cluster center of the support set

$$\hat{y} = \arg\max_m d(\mathcal{F}_\theta(\mathbf{x}^q), c_m); \ \mathbf{c}_m = \frac{1}{N} \sum_{(\mathbf{x},y,\mathbf{s}) \in \mathcal{S}, y=m} \mathcal{F}_\theta(\mathbf{x}).$$

VG-CoL Network

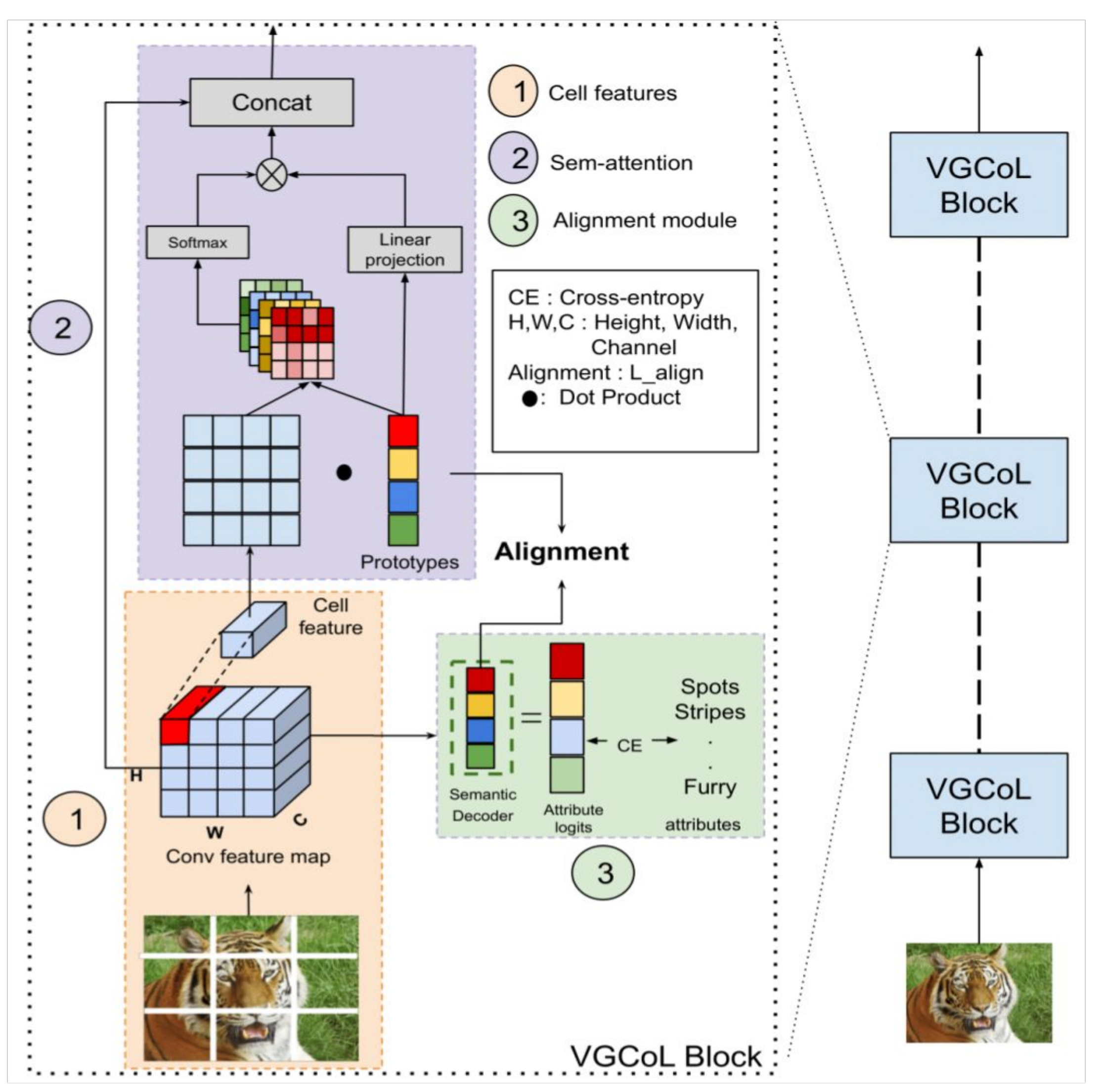


Fig: Given an input x to a convolutional layer, we derive features f where $f \in \mathbb{R}^{H \times W \times C}$. Each column of f, denoted as f_{ij} and belonging to $\mathbb{R}^C$, represents cell features that capture local information at specific spatial locations. Our objective is to ensure f_{ij} encodes visual concepts like color and texture, facilitating swift adaptation to novel categories with few examples.

- Once all the similarity maps are generated, we compute the attention score over each similarity matrix

- Given input **x** to a conv layer, we extract features **f** where **f**∈$\mathbf{R}^{HxWxC}$
- We consider columns of **f** as cell features, $\mathbf{f}_{ij} \in \mathbf{R}^C$
- Our goal is to make $\mathbf{f}_{ij}$ encode **visual concepts**, such as **color** and **texture**, enabling **quick generalization to novel categories** with **minimal examples.**

Semantic Co-Attention

- We introduce semantic/concept prototypes $\mathbf{P}_s$ consisting of $\mathbf{p}_k$ for each concept ***k*** with each prototype's dimension equal to the image feature channels (**C**).
- For each prototype $\mathbf{p}_k$, a similarity map $\mathbf{M}_k$ is computed using dot product between $\mathbf{f}_{ij}$ and $\mathbf{P}_k$

Learning Objective

- A^k enhances feature matching for concept *k*, but does not reveal the spatial interactions between the concepts.
- We address this by introducing a linear layer, computing weighted sum of prototypes:

$$\bar{\mathbf{P}}_s = \mathtt{Linear}([\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_K]), \text{ where } \bar{\mathbf{P}}_s \in \mathbb{R}^{H \times W \times K}$$

- Next we compute the product of the attention score matrix **A** and $\bar{\mathrm{P}}_s$ then concatenate this with the original features to produce the VGCoL block output:

$$VGCoL_{out} = \mathbf{f} \oplus \left(\mathbf{A} \odot \sigma\left(\bar{\mathbf{P}}_s\right)\right)$$

- Our Network **has multiple blocks of VGCoL** stacked on top of each other
- We perform **classification** on the output of **final VGCoL** block
- During testing, the output of the **final VGCoL** block is averaged to obtain prototype representation of each class $\mathbf{c}_k$

- To induce semantics into concept prototypes we introduce **semantic decoder**, a simple neural network that **outputs logits** equivalent to **attributes** present in a dataset
- The semantic decoder computes softmax over concepts:

$$p(\mathbf{s}_k = 1|\mathbf{x}) = \frac{\exp\left(\mathbf{W}^s_{[k,:]} \cdot \mathtt{AvgPool}(\mathbf{f})\right)}{\sum_k \exp\left(\mathbf{W}^s_{[k,:]} \cdot \mathtt{AvgPool}(\mathbf{f})\right)}$$

- W^s are trainable parameters of the semantic decoder. $W^s \in$ KxC. Here K^{th} row is associated with K^{th} concept

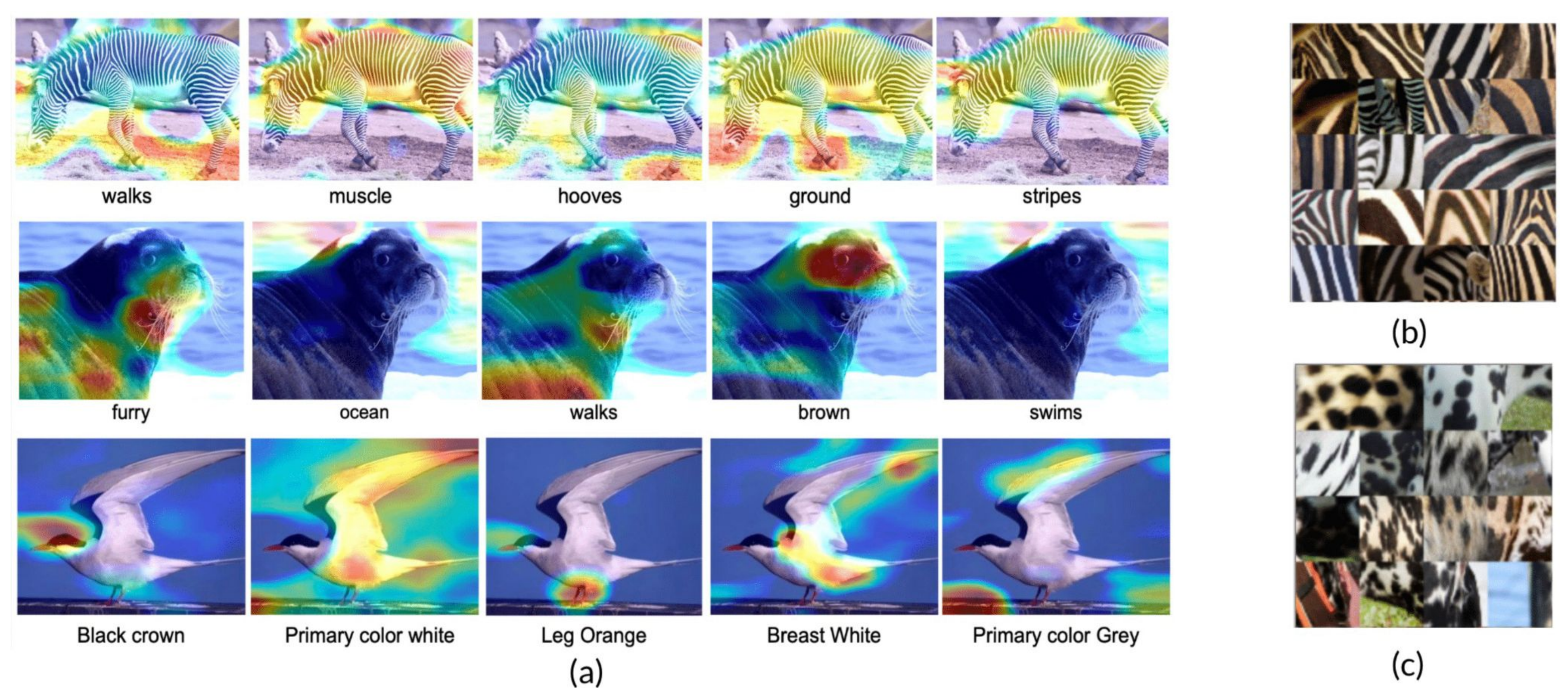
Optimization

In addition to training **VGCoL** network using classification loss ($\mathbf{L}_{cls}$) on the D_{seen}

1. We use cross entropy loss, termed $\mathbf{L}_{sem}$ to ensure each row of W^s of semantic decoder corresponds to semantic concepts such as "stripes" or "spot".
2. We align the semantic decoder weights with prototypes using $\mathbf{L}_1$ loss, termed as $\boldsymbol{L}_{align}$ in our work.
3. Given the frequent co-occurrence and correlation of visual attributes, we employ an orthogonality constraint to prevent concept entanglement, formulated as:

$$\mathcal{L}_{ortho} = |\mathbf{W^s} \cdot (\mathbf{W^s})^T - \mathbf{I}|$$

4. The network is jointly trained to optimize all losses: $Loss = \mathcal{L}_{cls} + \alpha \mathcal{L}_{sem} + \beta \mathcal{L}_{align} + \lambda \mathcal{L}_{ortho}$



*(a)Visualizing the similarity matrix **M**. Three samples and 5 concepts are illustrated. Red corresponds to strong grounding of the concept. **(b)** and **(c)** shows extracted patches around the concepts **stripes** and **spots**, respectively.*

Experiments and Results

- **Diverse data set:**
  - We use Caltech UCSD Birds (CUB), Scene Classification with Attributes (SUN) and Animals with Attributes 2 (AWA2).
- **Methods:**
  - We compare the performance of our model with state-of-the-art few-shot methods
  - We also present a strategy where we freeze the entire network except the semantic decoder and optimize by minimizing the combined semantic and alignment losses, aiming to instill prior semantic knowledge from the support set of novel classes.
- **Results:**

Method	CUB 1-shot	CUB 5-shot	SUN 1-shot	SUN 5-shot	AWA2 1-shot	AWA2 5-shot
ProtoNets [33]	43.4	67.8	37.1	63.1	41.9±0.8	54.86±0.7
MatchingNets [38]	48.5	69.2	41.0	60.4	-	-
RelationNets [35]	39.5	67.1	35.1	63.7	-	-
COMET [1]	67.9±0.9	85.3±0.5	-	-	-	-
CompoNets [37]	53.6	74.6	45.9	67.1	-	-
ConstellationNet - Conv-4	67.8±0.9	85.7±0.6	49.7±0.8	68.2±0.7	44.4±0.7	60.0±0.6
ConstellationNet - ResNet-12	70.1±0.8	86.3±0.5	50.3±0.8	70.1±0.7	47.3±0.7	63.3±0.6
Ours - Conv-4	66.7±0.5	83.1±0.6	52.5±0.8	69.1±0.7	45.7±0.7	61.5±0.6
Ours - ResNet-12	70.5±0.3	87.3±0.5	54.6±0.7	71.2±0.6	47.5±0.6	65.9±0.6
Ours - Conv-4 finetune	66.8±0.9	83.2±0.6	54.4±0.8	71.5±0.7	46.6±0.3	62.1±0.7
Ours - ResNet-12 finetune	**73.8±0.8**	**90.0±0.3**	**57.9±0.7**	**75.6±0.7**	**50.1±0.9**	**70.0±0.9**

References

1. Weijian Xu and Yifan xu and Huaijin Wang and Zhuowen Tu. *Attentional Constellation Nets for Few-Shot Learning.* In ICLR 2021
2. Tokmakov, Pavel and Wang, Yu-Xiong and Hebert, Martial. *Learning compositional representations for few-shot recognition.* In ICCV 2019

Conclusion

- In this work we introduced a weakly supervised, visually grounded concept learner enhancing few-shot performance, yielding interpretable VGCoL predictions for novel classes, and demonstrating potential in zero-shot object segmentation.