Paper_ID: 3874

Few-shot Image Classification with Multi-facet Prototypes

Kun Yan[1], Zied Bouraoui[2], Ping Wang[1], Shoaib Jameel[3], Steven Schockaert[4]

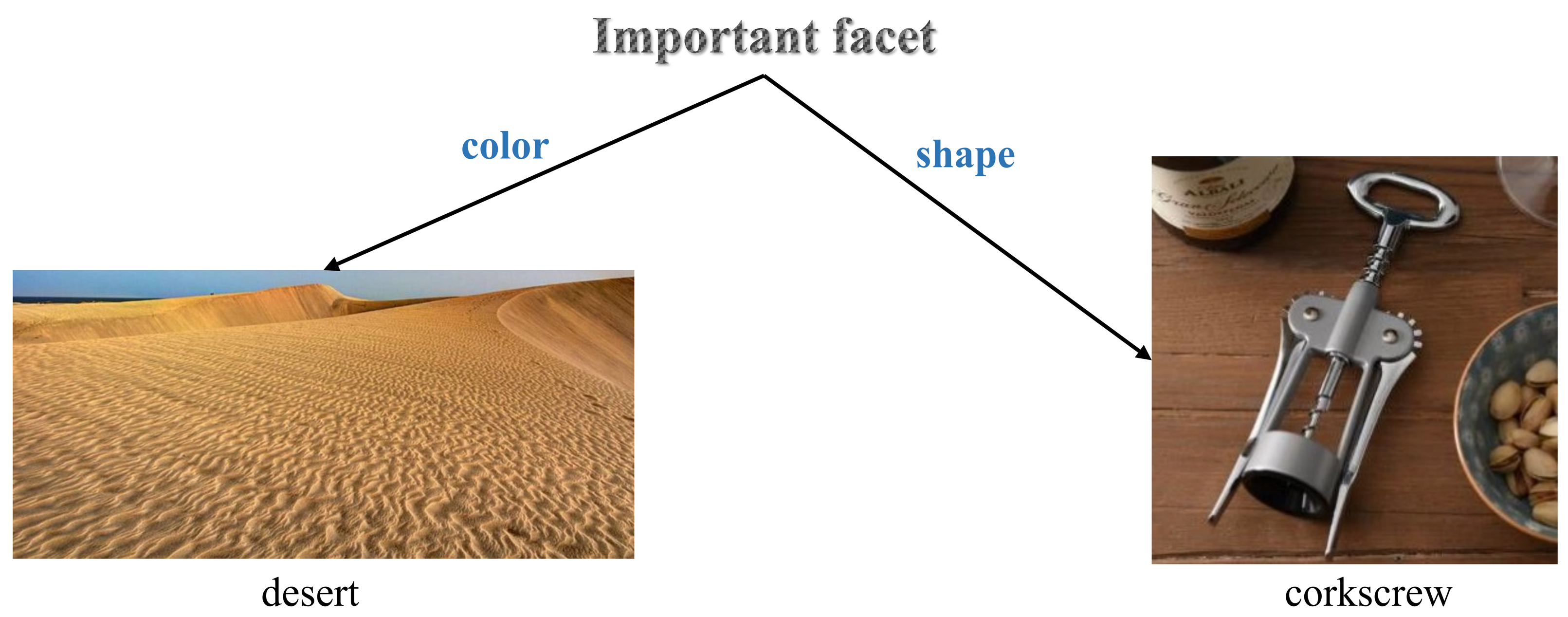
[1]Peking University, [2]CRIL – University of Artois & CNRS, [3]University of Essex, [4]Cardiff University

Introduction

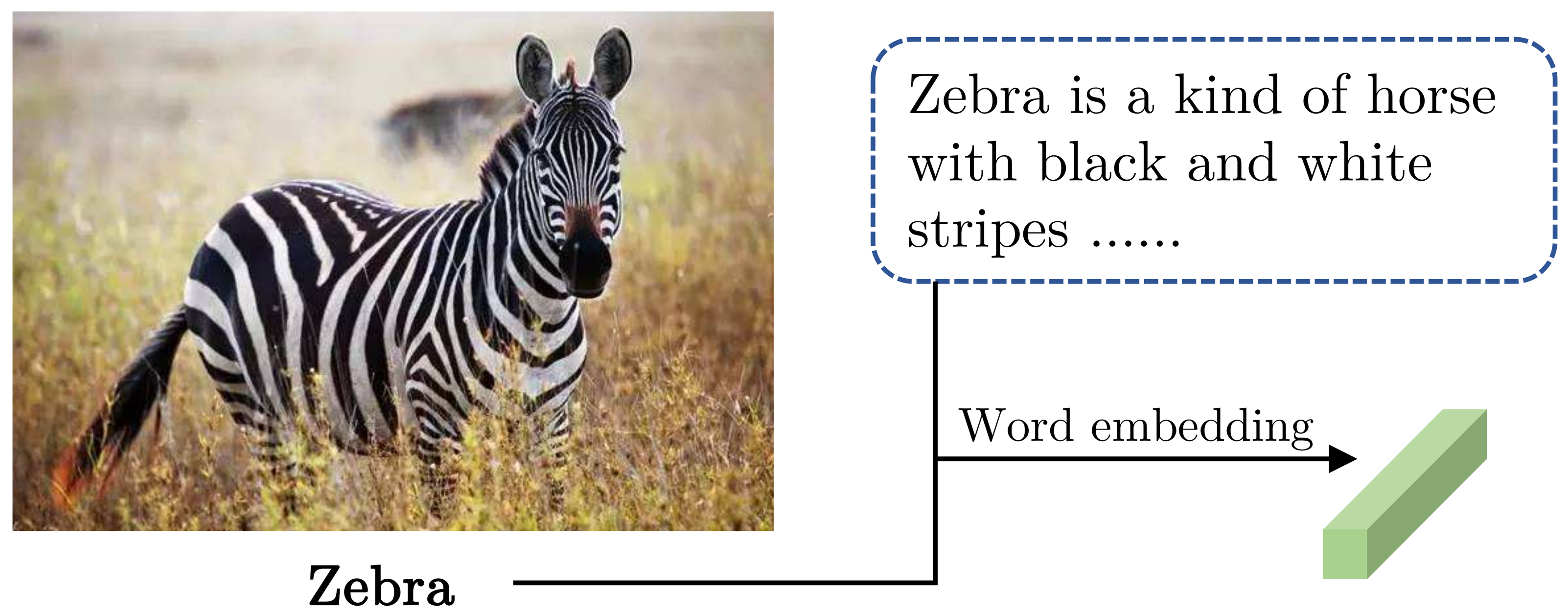
- Few-shot learning aims to recognize unseen images of new classes with only a few training examples.
- A central challenge is that the available training examples are normally insufficient to determine which visual features are most characteristic of the considered categories.

Motivation

- The importance of each facet differs from category to category.



- It is possible to predict facet importance from a pre-trained embedding of the category names.



Class Name Embeddings

- For each class c, we sample 1000 sentences from the May 2016 English Wikipedia dump;
- we replace the name of the class by [MASK], and take the sentence as the input to BERT. The class name embedding can be obtained from the output of the BERT.

Facet Identification

Our aim is to group the coordinates of the visual feature vectors $f_\theta(x)$, such that coordinates from the same group intuitively refer to similar aspects.

- Given a visual feature vector $f_\theta(x)$, we define $X_1, \dots, X_F$ as the set of coordinate indices of $f_\theta(x)$ of F different facets.
- We define a_c^i as the importance of the i^{th} coordinate for the class c, the formula is as below:

$$a_c^i = \sum_{(x_j,c)} RELU\left(\frac{\sum_{d\in C_p\{c\}} ||f_\theta(x_j[i] - \mathrm{v}_d[i])||_2^2}{(N-1)||f_\theta(x_j)[i] - \mathrm{v}_c[i]||_2^2} - 1\right)$$

$$a^i = \frac{1}{N}\sum_{c\in C_p} a_c^i$$

- We construct $m\times n$ matrix A by repeating the above computation for m episodes, where each time n classes are sampled.
- We computed the Kendall τ statistic between the i^{th} and j^{th} column of A. Let us write $e_{ij} \in [-1,1]$ for the resulting value.
- We use average-link agglomerative hierarchical clustering to partition the set $\{1, \dots, n\}$ into the facets $X_1, \dots, X_F$, where the values $e(i,j)$ are used to measure similarity.

Similarity Computation

- Given a word embedding n^c for class c, we introduce a facet-importance generation network g_e, which maps n^c onto an F-dimensional vector:

$$\mathrm{b}_c = g_e(\mathrm{n}^c)$$

- We obtain the final facet importance weights by applying a softmax layer:

$$(\eta_c^1, \dots, \eta_c^1) = \mathrm{SOFTMAX}\big(g_e(\mathrm{n}^c)\big)$$

$$\eta^i = \frac{1}{N}\sum_{c\in C_p} \eta_c^i$$

- The distance between a query image q and the prototype of class c as a weighted sum of facet-specific distances, as follow:

$$fdist(q,c) = \sum_{i=1}^{F} \eta^i ||f_\theta^i(q) - \mathrm{v}_c^i||_2^2$$

- Rather than using $fdist(q,c)$ directly, we combine $fdist(q,c)$ with the standard Euclidean distance, as used in ProtoNet, as follows:

$$dist(q,c) = |f_\theta(q) - \mathrm{v}_c| + \lambda \cdot fdist(q,c)$$

Experiments

- Ablation study of different word embeddings

Method	Backbone	Word Embeddings	5-way 5-shot
ProtoNet	ResNet-10	None	73.24 ± 0.63
Ours(ProtoNet)	ResNet-10	GloVe	74.10 ± 0.61
Ours(ProteNet)	ResNet-10	BERT	75.24 ± 0.76

- The mean accuracies (%) with a 95% confidence interval on the miniImageNet dataset

Method	Backbone	5-way 1-shot	5-way 5-shot
MAML [2]	Conv-64	48.70 ± 1.75	63.15 ± 0.91
Reptile [18]	Conv-64	47.07 ± 0.26	62.74 ± 0.37
LEO [19]	WRN-28	61.76 ± 0.08	77.59 ± 0.12
MTL [20]	ResNet-12	61.20 ± 1.80	75.50 ± 0.80
MetaOptNet-SVM [21]	ResNet-12	62.64 ± 0.61	78.63 ± 0.46
Matching Net [7]	Conv-64	43.56 ± 0.84	55.31 ± 0.73
ProtoNet [5]	Conv-64	49.42 ± 0.78	68.20 ± 0.66
RelationNet [4]	Conv-64	50.44 ± 0.82	65.32 ± 0.70
ProtoNet [5]	ResNet-12	56.52 ± 0.45	74.28 ± 0.20
TADAM [22]	ResNet-12	58.50 ± 0.30	76.70 ± 0.38
AM3(ProtoNet, BERT)	ResNet-12	62.11 ± 0.39	74.72 ± 0.64
AM3(ProtoNet, GloVe)	ResNet-12	62.43 ± 0.80	74.87 ± 0.65
AM3(ProtoNet++) [10]	ResNet-12	65.21 ± 0.49	75.20 ± 0.36
TRAML(ProtoNet) [12]	ResNet-12	60.31 ± 0.48	77.94 ± 0.57
DSN-MR [23]	ResNet-12	64.60 ± 0.48	79.51 ± 0.50
DeepEMD [24]	ResNet-12	65.91 ± 0.82	82.41 ± 0.56
FEAT [6]	ResNet-12	66.78	82.05
Ours(ProtoNet)	ResNet-12	63.21 ± 0.37	77.84 ± 0.64
Ours(FEAT)	ResNet-12	**67.24 ± 0.58**	**82.51 ± 0.66**

- The mean accuracies (%) with a 95% confidence interval on the CUB dataset

Method	Backbone	5-way 1-shot	5-way 5-shot
MAML	Conv-64	55.92 ± 0.95	72.09 ± 0.76
Matching Net	Conv-64	61.16 ± 0.89	72.86 ± 0.70
ProtoNet	Conv-64	51.31 ± 0.91	70.77 ± 0.69
RelationNet	Conv-64	62.45 ± 0.98	76.11 ± 0.69
Baseline++	Conv-64	60.53 ± 0.83	79.34 ± 0.61
SAML [25]	Conv-64	69.35 ± 0.22	81.37 ± 0.15
DN4 [26]	Conv-64	53.15 ± 0.84	81.90 ± 0.60
Ours(ProtoNet)	Conv-64	**69.52 ± 0.76**	**82.34 ± 0.66**