Multimodal Prototypical Networks for Few-shot Learning

Frederik Pahde[1, †], Mihai Puscas[2, ‡], Tassilo Klein[3], Moin Nabi[3]
[1]Amazon.com, Inc., [2]Huawei Research, Ireland, [3]SAP AI Research, Berlin, Germany
frederikpahde@gmail.com, mihai.puscas@huawei.com, {tassilo.klein, m.nabi}@sap.com

1. Extended Quantitative Results

In this section we provide additional results for the quantitative evaluations of our method in our main paper.

1.1. Results for multimodal 50-way classification

We extend Tab. 1 from the main paper which shows the accuracies for 50-way classification in comparison with state-of-the-art multimodal few-shot learning methods and our baselines. While in the paper we only reported the top-5 accuracy, we provide top-1 and top-3 accuracy in addition in Tab. 1. We can observe the same trends as for the top-5 accuracy.

Dataset	Method	Metric	1-shot	2-shot	5-shot	10-shot	20-shot
CUB	Pahde et al. [1]	Top-1	24.90	25.17	34.66	44.00	53.70
		Top-3	37.59	39.75	49.86	59.62	67.99
		Top-5	57.67	59.83	73.01	78.10	84.24
	Image Only Baseline	Top-1	28.91±0.18	37.52±0.15	47.33±0.12	52.31±0.12	**55.62±0.14**
		Top-3	51.13±0.22	62.23±0.16	72.53±0.11	76.71±0.09	78.87±0.10
		Top-5	62.65±0.22	73.52±0.15	82.44±0.09	85.64±0.08	87.27±0.08
	ZSL Baseline	Top-1	22.39±0.18	27.15±0.16	32.24±0.15	34.65±0.13	36.42±0.17
		Top-3	45.37±0.21	52.42±0.18	59.09±0.14	61.73±0.13	63.27±0.15
		Top-5	58.28±0.22	65.62±0.19	71.79±0.14	74.15±0.11	75.32±0.13
	Our (Multimodal)	Top-1	**34.16±0.17**	**41.43±0.14**	**48.84±0.10**	**53.01±0.11**	55.58±0.14
		Top-3	**58.56±0.19**	**67.44±0.13**	**74.65±0.09**	**77.60±0.09**	**79.30±0.11**
		Top-5	**70.39±0.19**	**78.62±0.12**	**84.32±0.06**	**86.23±0.08**	**87.47±0.09**
Oxford-102	Pahde et al. [1]	Top-1	43.77	61.42	72.49	-	-
		Top-3	57.96	77.68	82.18	-	-
		Top-5	78.37	91.18	92.21	-	-
	Image Only Baseline	Top-1	49.39±0.34	60.02±0.27	70.24±0.18	74.49±0.16	**76.98±0.16**
		Top-3	74.12±0.32	82.76±0.21	89.11±0.12	91.34±0.11	92.46±0.10
		Top-5	83.88±0.27	90.27±0.16	94.25±0.09	95.61±0.08	96.21±0.08
	ZSL Baseline	Top-1	34.99±0.32	39.63±0.29	42.92±0.24	44.21±0.22	44.91±0.21
		Top-3	60.90±0.34	65.91±0.30	69.89±0.23	71.70±0.21	72.76±0.20
		Top-5	73.14±0.31	77.63±0.27	81.10±0.19	82.88±0.18	84.00±0.17
	Our (Multimodal)	Top-1	**53.70±0.30**	**62.46±0.24**	**71.08±0.19**	**74.73±0.17**	76.77±0.15
		Top-3	**77.71±0.25**	**84.35±0.18**	**89.51±0.11**	**91.53±0.11**	**92.48±0.10**
		Top-5	**86.34±0.21**	**91.30±0.15**	**94.58±0.09**	**95.77±0.08**	**96.34±0.07**

Table 1: 50-way classification top-1, top-3 and top-5 accuracies in comparison to other multimodal few-shot learning approaches and our baselines for CUB-200 and Oxford-102 datasets with $n \in \{1, 2, 5, 10, 20\}$. The best results are in bold.

[†]Work completed while at SAP AI Research, prior to joining Amazon.com, Inc.
[‡]Work completed while at SAP AI Research, prior to joining Huawei Research, Ireland

1.2. Results for additional k-way n-shot classification tasks

To extend the results in Tab. 2 in the main paper, we evaluate our method in additional k-way classification scenarios with $k \in \{5, 10, 20\}$. For this experiment we compare our full method to the image-only baseline. Further extending Tab. 2 from the main paper, we report top-3 and top-5 accuracy in addition to the top-1 accuracy. The results for different n-shot scenarios with $n \in \{1, 2, 5, 10, 20\}$ are shown in Tab. 2. It can be observed that in the most scenarios our method improves the classification accuracies and decreases the variance. However, the more visual data is available (higher n), the less impact have the generated features.

Task	Method	Metric	1-shot	2-shot	5-shot	10-shot	20-shot
5-way	Image Only Baseline	Top-1	68.85±0.86	79.99±0.66	83.93±0.57	86.95±0.49	**87.78±0.48**
		Top-3	94.71±0.34	97.80±0.16	98.39±0.12	98.91±0.09	98.95±0.09
		Top-5	-	-	-	-	-
	Our (Multimodal)	Top-1	**75.01 ± 0.81**	**80.90±0.64**	**85.30 ± 0.54**	**86.96±0.48**	87.67±0.51
		Top-3	**96.83±0.25**	**98.06±0.14**	**98.65±0.11**	**98.94±0.08**	**99.02±0.09**
		Top-5	-	-	-	-	-
10-way	Image Only Baseline	Top-1	59.34±0.76	68.25±0.65	75.72±0.61	78.12±0.60	**79.68±0.85**
		Top-3	86.19±0.51	91.67±0.33	94.66±0.25	**95.65±0.24**	**95.99±0.31**
		Top-5	94.50±0.29	97.20±0.15	98.24±0.11	98.55±0.11	**98.63±0.14**
	Our (Multimodal)	Top-1	**62.25±0.73**	**69.71±0.63**	**76.02**±0.62	78.12±0.64	79.29±0.82
		Top-3	**88.41±0.45**	**92.78±0.28**	**94.88±0.23**	95.55±0.25	95.89±0.31
		Top-5	**95.80±0.25**	**97.57±0.14**	**98.32±0.11**	**98.57±0.11**	98.62±0.12
20-way	Image Only Baseline	Top-1	46.55±0.40	55.94±0.36	64.64±0.39	**68.21±0.43**	**69.97±0.47**
		Top-3	73.60±0.38	81.94±0.28	87.93±0.23	89.66±0.22	**90.55±0.26**
		Top-5	84.37±0.30	90.56±0.19	94.06±0.14	94.96±0.14	**95.49±0.15**
	Our (Multimodal)	Top-1	**48.23±0.40**	**57.02±0.36**	**64.94±0.38**	68.08±0.43	69.62±0.44
		Top-3	**75.31±0.36**	**82.90±0.27**	**88.31±0.22**	**89.75±0.21**	90.51±0.24
		Top-5	**85.65±0.29**	**91.23±0.18**	**94.27±0.14**	**95.02±0.13**	95.42±0.14

Table 2: Top-1, top-3 and top-5 accuracies for different k-way classification tasks on the CUB-200 dataset of our approach compared to our image-only baseline. We report the average accuracy of 600 randomly sampled few-shot episodes including 95% confidence intervals. The best results are in bold.

2. Visualization of Embedding Space

In this section we provide additional visualizations for the embedding spaces in different few-shot learning scenarios. Therefore, we show visualizations for the embedding space in k-way classification problems with $k \in \{5, 10, 20, 50\}$. We use t-SNE for dimensionality reduction. The visualizations are shown in Figures 1, 2, 3 and 4 in which the different colors indicate the class membership. Although the dimensionality is reduced to two for being able to visualize the embedding space we can still see the clusters for different classes. It can be observed that in many cases the multimodal prototypes (triangles) are moved towards the center of the cluster for the particular class compared to the image-only prototypes (crosses).

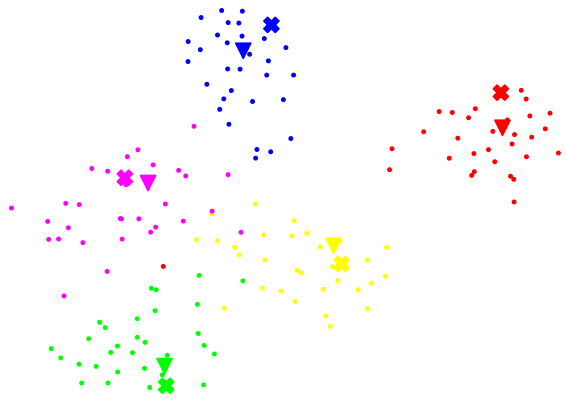

Figure 1: t-SNE graph for 5-way classification showing the image-only prototypes (crosses), updated multimodal prototypes (triangles) and unseen test samples (dots)

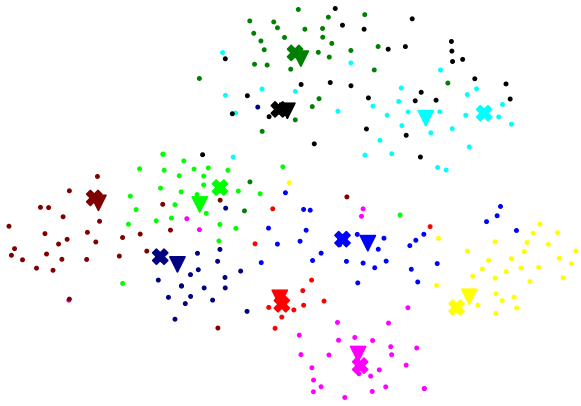

Figure 2: t-SNE graph for 10-way classification showing the image-only prototypes (crosses), updated multimodal prototypes (triangles) and unseen test samples (dots)

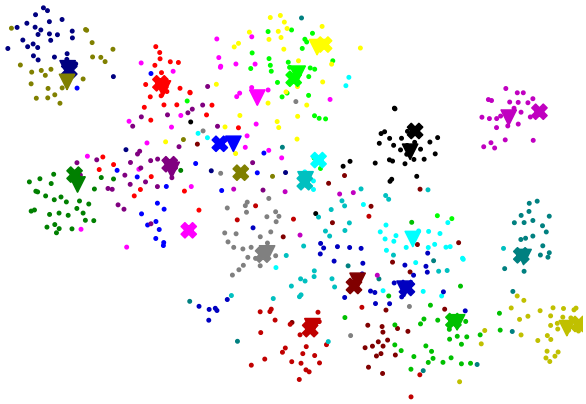

Figure 3: t-SNE graph for 20-way classification showing the image-only prototypes (crosses), updated multimodal prototypes (triangles) and unseen test samples (dots)

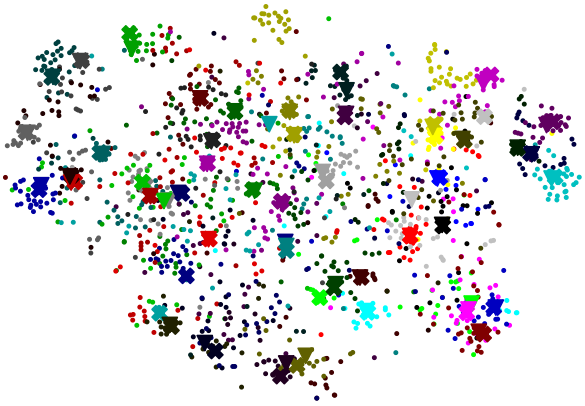

Figure 4: t-SNE graph for 50-way classification showing the image-only prototypes (crosses), updated multimodal prototypes (triangles) and unseen test samples (dots)

3. Retrieval Results

As an additional experiment we use the generated feature vectors to retrieve the closest unseen test images in the 5-way 1-shot learning scenario. The feature vectors are generated conditioned on the textual descriptions of the training images. We calculate the distance between feature vectors using the cosine distance measure. The retrieval results for some randomly selected feature vectors are shown in Fig. 5. This shows the effectiveness of the text-conditional feature generator. It can be seen that most of the retrieved images are from the correct class. Thus, the generated feature vector is close to unseen test images of the same class, facilitating classification with a nearest neighbor approach.

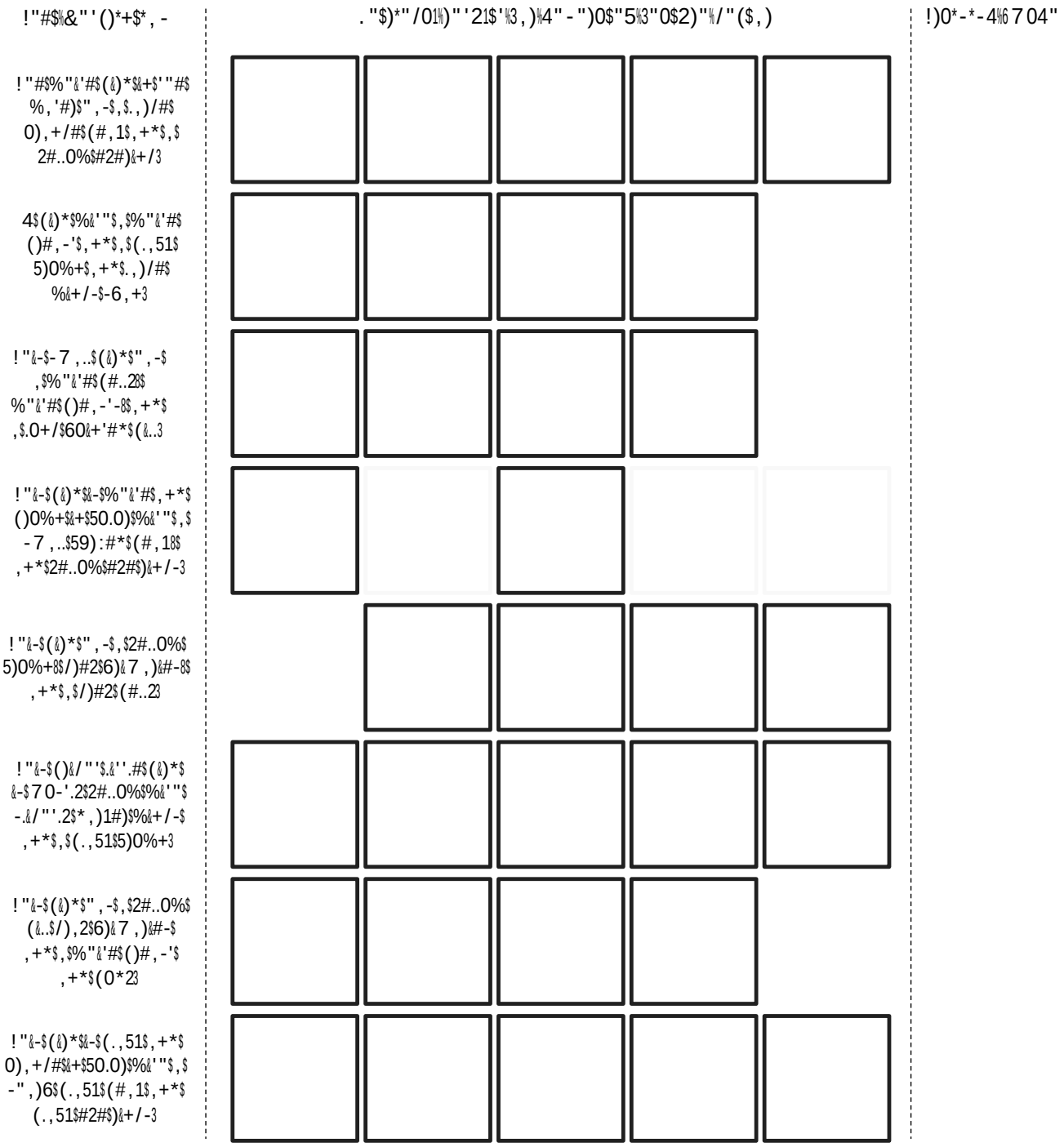

Figure 5: Retrieval Results for generated feature vectors in a 5-way classification task: The left column shows the textual description (one random caption out of the ten available descriptions per training image), in the middle are the top-5 retrieved unseen test images and in the right column is the training image for the particular class. The color of the surrounding box indicates whether the retrieved test image is from the correct class (green) or a wrong class (red).

References

[1] F. Pahde, O. Ostapenko, P. Jähnichen, T. Klein, and M. Nabi. Self-paced adversarial training for multimodal few-shot learning. *WACV*, 2019. 1