Self-Distillation for Few-Shot Image Captioning (Supplementary Materials)

Xianyu Chen, Ming Jiang, Qi Zhao
University of Minnesota, Twin Cities
{chen6582, mjiang}@umn.edu, qzhao@cs.umn.edu

In this supplementary material, we provide additional information and results of our proposed few-shot image captioner on the Microsoft COCO dataset [7] and Flickr30K dataset [8]. In Section 1, we present the algorithm of our few-shot image captioning. Next, in Section 2, we show a comprehensive ablation study for the hyper-parameters and the effects on different base models. Further, in Section 3, we present the quantitative results for Flickr30K [8] with state-of-the-art approach and our different baselines. Finally, in Section 4, we report and discuss qualitative results of various baselines and state-of-the-art approaches.

1. Algorithm

We summarize our proposed few-shot image captioning approach in Algorithm 1. Please refer to Section 3 in our main paper for more details.

2. Ablation Study

To investigate the effectiveness of the various hyper-parameters, we conduct an ablation study by comparing a set of baseline models in Table 1.

Model Selection with λ_x and λ_y. As shown in The second and third panels of Table 1, we optimize the unsupervised loss term for unpaired images and unpaired captions in Equation (1), respectively. In each panel, we fix the other optimal parameters and change the value of the specific ones. As mentioned above, we select the CIDEr score as the criterion for model selection. The corresponding results demonstrate that our approach achieves the best performance at $\lambda_x = 0.1$ and $\lambda_y = 1$, respectively. Hence, we select $\lambda_x = 0.1$ and $\lambda_y = 1$ in all experiments trained with the total loss (1).

Model Selection with Smoothness α. In the fourth panel of Table 1, we evaluate the effects of the hyper-parameter α in Equation (3). When $\alpha = 0$, the Mean Teacher becomes a naïve model without a temporal ensemble. Among these settings, our approach achieves the best performance at $\alpha = 0.99$. This validates the necessity of the employment of the Mean Teacher for providing more robust pseudo captions.

Model Selection with σ for Pseudo Feature Generation. In the fifth panel of Table 1, we fine-tune the standard deviation of the initialized pseudo feature. The corresponding results show that we can set the initialized noise standard derivation σ as 0.1 to achieve the best performance. Similarly, we select $\sigma = 0.1$ in all the other experiments.

Effects on different base models. Since most of the few-shot image captioner [4] and unsupervised image captioners [2, 3, 5] use the NIC as a base model. For a fair comparison, we adopt the this base model into our design framework in a large portion of our experiments. However, to demonstrate the generalization of our proposed method in different base models, we use other state-of-the-art base models, the Att2in [9] and Up-Down [1] in Table 2. This table shows a reasonable improvement based on different base models. The performances of different baselines are also consistent with different base models.

3. Quantitative Results

Captioning with Flickr30K. In Table 3, we demonstrate the performance of our three baselines as well as the state-of-the-art approach Pivoting [3] in Flickr30K [8] dataset with only 1% image-caption pairs, while the remaining data is used as the unpaired data. The performance is verified by the Flickr30K test set. The results show our method can significantly improve the performance in this dataset. To sum up, The improvement among different datasets and data settings demonstrates the generalizability of our self-distillation for the few-shot image captioning task.

4. Qualitative Results

Figure 1 presents qualitative examples of captions generated with three different baselines: Ours (P), Ours (P+UI) and Ours (P+UI+UC). As shown in the examples, trained with all the paired and unpaired data using the self-distillation method, our few-shot image captioner generates better captions. The advantages of our methods can be observed as follows: (1) It describes important objects better (*e.g.,* 'cows' in Figure 1a, 'cat' in Figure 1b, 'woman' and 'tennis court' in Figure 1c, 'clock' and 'building' in

Algorithm 1: Ensemble-Based Self-Distillation for Few-Shot Image Captioning

Data: Image-caption pairs $\mathcal{D}_{x,y}$, unpaired image set $\mathcal{D}_x$ and unpaired caption set $\mathcal{D}_y$;
Input: The number of base models M, smoothing coefficient α for Equation (3) to update the parameters of the Mean Teacher, hyper-parameters λ_x and λ_y for Equation (1) to balance different loss terms, standard derivation σ for the initialization of pseudo latent feature and the probability $\boldsymbol{p}$ to sample integer n from the categorical distribution.
Initialization: Randomly initialize weights $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ and set $\boldsymbol{\Theta}_1 = \boldsymbol{\theta}_1, \ldots, \boldsymbol{\Theta}_M = \boldsymbol{\theta}_M$;
for *epoch* in [*1,num_epochs*] **do**
  for *iter* in [*1,max_iters*] **do**
    1: Generate mini-batch $\mathcal{B}_{x,y} \subset \mathcal{D}_{x,y}$, $\mathcal{B}_x \subset \mathcal{D}_x$ and $\mathcal{B}_y \subset \mathcal{D}_y$;
    2: Randomly select n from $\mathrm{Cat}(M, \boldsymbol{p})$, then the n-th base model would be taught from the pseudo labels;
    3: Generate the pseudo captions for $\mathcal{B}_x$ by the ensemble $\{\tilde{\boldsymbol{y}}^1, \ldots, \tilde{\boldsymbol{y}}^K\} = F(\boldsymbol{x}|\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M),\ \boldsymbol{x} \in \mathcal{B}_x$;
    4: Calculate the softmax normalization $\boldsymbol{\gamma} = \mathrm{softmax}(\boldsymbol{s})$ for the generated pseudo captions, where $\boldsymbol{s}$ is obtained from Equation (4);
    5: Generate the pseudo latent features for $\mathcal{B}_y$ by Equation (6) with Gradient Descent;
    6: Joint update parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ by back-propagation of the Equation (1);
    7: Update temporally average model weights $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M$ by Equation (3).
  end
end

Hyper-parameters				**COCO validation**								
λ_x	λ_y	α	σ	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE	WMD
0.1	1	0.99	0.1	64.4	45.8	31.9	22.2	19.9	46.8	**61.3**	12.7	14.7
0	1	0.99	0.1	64.5	45.9	31.9	22.0	20.0	47.0	61.0	12.6	14.6
0.01	1	0.99	0.1	64.6	46.1	32.0	22.2	19.9	46.9	61.1	12.6	14.6
1	1	0.99	0.1	64.5	45.8	31.9	22.0	20.0	46.7	61.2	12.6	14.6
0.1	0	0.99	0.1	63.7	44.9	30.9	21.4	19.4	46.3	57.5	12.1	13.9
0.1	0.5	0.99	0.1	64.5	45.7	31.7	21.9	20.0	46.9	61.2	12.8	14.7
0.1	2	0.99	0.1	64.2	45.7	31.7	21.9	19.8	46.6	60.6	12.5	14.3
0.1	1	0	0.1	62.9	44.1	30.4	20.8	19.4	45.8	57.7	11.8	14.1
0.1	1	0.9	0.1	64.5	45.7	31.8	22.0	20.0	46.9	60.6	12.6	14.5
0.1	1	0.999	0.1	63.7	44.8	30.9	21.3	19.8	46.2	59.4	12.5	14.3
0.1	1	0.99	0.01	64.4	45.7	31.7	22.0	19.9	46.7	61.2	12.6	14.6
0.1	1	0.99	1	64.5	45.7	31.7	21.8	19.9	46.6	60.8	12.6	14.3

Table 1. Ablation study on the selection of hyper-parameters on the COCO validation set. The first panel shows the evaluation results with the optimal hyper-parameters. The four panels below show the effects of λ_x, λ_y, α, σ, respectively.

Figure 1d, 'bench' in Figure 1e, 'skateboard' Figure 1g); (2) It describes multiple similar objects more specifically and accurately (*e.g.,* 'a man and a woman' in Figure 1e, 'a group of people' in Figure 1f); (3) Self-distillation allows to describe precise actions between the subject and the object (*e.g.,* 'grazing' in Figure 1a, 'eating food' in Figure 1f). These characteristics suggest that the self-distillation method can simultaneously make use of the unpaired images and captions to model the image feature representations and language structures, which gradually improves the quality of the generated captions over the training process.

Furthermore, we also present qualitative examples to compare the results of our self-distillation few-shot image captioner with state of the art. Figure 2 compares our results with captions generated with Pseudo Label [6] and Deep Mutual Learning [11]. With the self-distillation method, our image captioner can leverage a large number of unpaired images and captions to improve the performance of a few-shot image captioning. Similar to the observations illustrated in the main paper, the self-distillation plays two important roles: (1) Self-distillation provides an accurate action description (*e.g.,* 'flying kites' in Figure 1a, 'riding a wave' in Figure 1b,

Method	Base model	COCO test								
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE	WMD
Mean Teacher (P)	NIC [10]	62.0	43.1	29.4	20.1	18.7	45.1	53.8	11.4	13.4
Ours (P)		62.9	44.1	30.4	20.9	19.2	45.7	56.2	11.9	13.7
Ours (P+UI+UC)		64.5	45.9	32.1	22.5	20.0	46.7	**62.4**	12.7	14.7
Mean Teacher (P)	Att2in2 [9]	64.3	45.4	31.6	21.9	18.9	46.2	54.7	11.5	13.3
Ours (P)		64.5	46.0	32.5	22.7	19.4	46.7	58.8	11.8	14.2
Ours (P+UI+UC)		66.9	48.6	34.5	24.3	20.8	48.2	**66.3**	13.2	15.4
Mean Teacher (P)	Up-Down [1]	65.2	46.9	33.0	23.1	20.2	47.2	63.7	13.0	14.9
Ours (P)		66.2	48.2	34.2	24.2	20.9	48.1	67.6	13.5	15.8
Ours (P+UI+UC)		67.9	49.8	35.4	25.0	21.7	49.3	**73.0**	14.5	16.6

Table 2. Quantitative comparison with various base models on the COCO test set. We select three commonly used base models (*e.g.,*NIC [10], Att2in2 [9] and Up-Down [1]) which achieve state-of-the-art performance in image captioning task. To have a fair comparison with most exist few-shot/unsupervised image captioners, we do not apply policy gradient update [9] to these baselines.

Method	Flickr30K test								
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE	WMD
Pivoting [3]	49.7	27.8	14.8	7.9	13.6	-	16.2	-	-
Mean Teacher (P)	50.6	30.2	17.4	10.3	13.1	34.7	13.7	6.8	8.2
Ours (P)	50.6	30.2	17.9	11.1	13.4	35.7	14.1	7.1	8.2
Ours (P+UI+UC)	54.3	33.4	20.0	12.0	14.0	36.6	**16.5**	7.7	8.8

Table 3. Quantitative comparison on the Flickr30K test set among our three main baselines and the state-of-the-art unsupervised [3] image captioner.

'playing soccer' in Figure 1c); (2) Self-distillation helps to describe important objects correctly (*e.g.,* 'vase' in Figure 1d, 'dog' in Figure 1e, 'frisbee' in Figure 1f, 'red stop sign' in Figure 1g). These examples demonstrate that self-distillation is effective for generating accurate and robust pseudo captions and pseudo features, which leads to better performance. Specifically, compared with Pseudo Label [6] generating the hard captions and the Mutual Deep Learning [11] mutually teaching each base model with the other base models, our self-distillation method utilizes the soft captions generated from an ensemble, providing a more stable training process. Furthermore, with Gradient Descent, the generated pseudo latent features also help to bridge the performance gap between model outputs and human annotations.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. *European Conference on Computer Vision (ECCV)*, 2018.

[4] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[5] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[6] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *International Conference on Machine Learning (ICML)*, 2013.

[7] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)*, 2014.

[8] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

(a)

Ours (P): A herd of sheep in a field.
Ours (P+UI): A herd of cows standing in a field.
Ours (P+UI+UC): A group of cows grazing in a field.
Ground-truth: A field with a bunch of cows grazing.

(b)

Ours (P): A small room with a bed and closet.
Ours (P+UI): A bedroom is sitting on a bed in the room.
Ours (P+UI+UC): A cat sitting on a bed in a room.
Ground-truth: A cat sleeps on a bed in a small room.

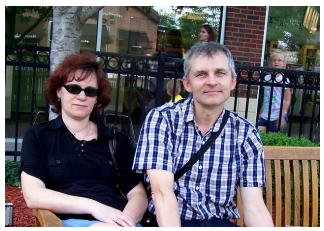
(c)

Ours (P): A young boy is holding a tennis racket.
Ours (P+UI): A young boy holding a tennis racket on a tennis court.
Ours (P+UI+UC): A woman on a tennis court holding a tennis racket.
Ground-truth: A woman on a tennis court with a racquet.

(d)

Ours (P): A train that is on the side of a bridge.
Ours (P+UI): A train is standing in front of a building.
Ours (P+UI+UC): A large building with a clock on it.
Ground-truth: A very big building with a mounted clock.

(e)

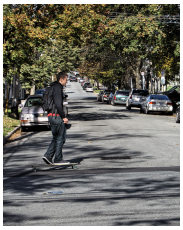


Ours (P): A man is sitting at a table with a cell phone.
Ours (P+UI): A man sitting on a bench with a cell phone.
Ours (P+UI+UC): A man and a woman sitting on a bench.
Ground-truth: A man and woman sitting on a wooden bench together.

(f)

Ours (P): A man is sitting at a table with a knife.
Ours (P+UI): A man is sitting at a table with some food.
Ours (P+UI+UC): A group of people at a table eating food.
Ground-truth: A group of people sitting around a table eating food.

(g)



Ours (P): A man riding a motorcycle down the road.
Ours (P+UI): A man riding a motorcycle down a street.
Ours (P+UI+UC): A man riding a skateboard down a street.
Ground-truth: A man riding a skateboard across a street.

Figure 1. Qualitative examples of various baselines on COCO validation set. With the use of unpaired images and unpaired captions, our approach is able to generate more accurate and fluent captions. We use different colors to demonstrate the advantages of our proposed method (*e.g.,*red color indicates that it can describe important objects better, green color shows that it describes multiple similar objects more specifically and accurately, and blue color demonstrates that it can describe precise actions between the subject and the object).

(a)

Pseudo Label: A group of people on a beach with a kite.
Deep Mutual Learning: A group of people standing on a beach.
Ours (P+UI+UC): A crowd of people flying kites on the beach.
Ground-truth: A group of people flying kites at the beach.

(b)

Pseudo Label: A man on a surfboard in the ocean.
Deep Mutual Learning: A man on a surfboard in the water.
Ours (P+UI+UC): A man on a surfboard riding a wave.
Ground-truth: A man who is on a surfboard riding a wave.

(c)

Pseudo Label: A group of people playing a game of frisbee.
Deep Mutual Learning: A group of people playing a game of frisbee.
Ours (P+UI+UC): A group of people playing soccer on a field.
Ground-truth: A group of kids that are playing soccer on a field.

(d)

Pseudo Label: A table with a bunch of flowers on it.
Deep Mutual Learning: A table with two flowers on it.
Ours (P+UI+UC): A vase filled with flowers on a table.
Ground-truth: A small green vase with some pretty yellow flowers.

(e)

Pseudo Label: A man in a field with a frisbee.
Deep Mutual Learning: A man in a field with a frisbee.
Ours (P+UI+UC): A dog running on a field with a frisbee.
Ground-truth: A dog chasing a red frisbee across a field.

(f)

Pseudo Label: A man in a field with a hat.
Deep Mutual Learning: A man is standing in front of a giraffe.
Ours (P+UI+UC): A man is holding a frisbee in a field.
Ground-truth: A close up of a person holding a frisbee in a field.

(g)

Pseudo Label: A street sign on the side of a road.
Deep Mutual Learning: A street sign in front of a building.
Ours (P+UI+UC): A red stop sign in front of a building.
Ground-truth: A red stop sign in front of a building.

Figure 2. Qualitative examples of various state-of-the-art approaches on COCO validation set. Our proposed image captioner can generate more natural and accurate captions, which also sheds light on the effectiveness of self-distillation. We use two different colors to demonstrate the advantages of self-distillation (*e.g.*,red color indicates that it can provide an accurate action description, and green color shows that it can describe important objects correctly).