Self-Distillation for Few-Shot Image Captioning

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

Abstract

The development of large-scale image-captioning datasets is expensive, while the abundance of unpaired images and text corpus can potentially help reduce the efforts of manual annotation. In this paper, we study the few-shot image captioning problem that only requires a small amount of annotated image-caption pairs. We propose an ensemble-based self-distillation method that allows image captioning models to be trained with unpaired images and captions. The ensemble consists of multiple base models trained with different data samples in each iteration. For learning from unpaired images, we generate multiple pseudo captions with the ensemble and allocate different weights according to their confidence levels. For learning from unpaired captions, we propose a simple yet effective pseudo feature generation method based on Gradient Descent. The pseudo captions and pseudo features from the ensemble are used to train the base models in future iterations. The proposed method is general over different image captioning models and datasets. Our experiments demonstrate significant performance improvements and meaningful captions generated with only 1% of paired training data. Source code is available at `https://github.com/chenxy99/SD-FSIC`.

1. Introduction

The advances in Deep Neural Networks (DNNs) have demonstrated promising performances in vision and natural language processing tasks. Driven by such advances, research in image captioning, a cross-modal task that requires both visual and language modeling, has been developing rapidly in recent years. Most image captioning methods learn a deep neural network model in a supervised learning manner based on manually labeled image-caption pairs [5, 54, 57]. Despite their success, the training of these supervised models requires a large corpus of captions paired with images, which is extraordinarily labor-intensive. With over 123k images annotated with 5 captions each, the most popular image captioning dataset Microsoft COCO [39] is still considered relatively small compared with general datasets such as ImageNet [47] and OpenImages [32]. Therefore, the high cost of manual annotations has limited the generalizability of image captioning models.

To alleviate the expensive cost of annotating image-caption pairs, recent studies propose semi-supervised learning [8, 28, 40] and unsupervised image captioning [13, 19, 33] approaches, allowing image captioners to learn from unpaired images and captions. These methods utilize externally trained object detectors [45, 59, 61] and external sentence corpus [13]. Semi-supervised image captioners also leverage external modeling [28, 40] and language data [8] to establish semantic alignments between visual and language data and hence boost the performance of image captioners. Despite their success, their performances are highly dependent on the availability of external data and models, especially when only few image-caption pairs are annotated.

To address the few-shot image captioning problem, we for the first time propose an ensemble-based self-distillation method that does not depend on any external knowledge. Specifically, we train multiple base models using annotated image-caption pairs together with unpaired images and captions, which forms an ensemble that performs better than the individual models. Pseudo captions and image features are generated with the ensemble, and added to the training of the base models. This method is considered a self-distillation approach, in which the ensemble serves as a teacher and the base models serve as the students. To improve the accuracy and robustness of the ensemble, different weights of loss are assigned to the pseudo captions depending on their confidence levels. Further, we introduce a simple yet effective method to generate pseudo features from unpaired captions, and use these features to train the base models. With the proposed method, we can leverage the large number of unpaired images and captions to improve the performance of few-shot image captioning.

In sum, the contributions of this work include: 1) a novel approach to **few-shot image captioning** based on temporal ensemble and multi-model ensemble, 2) a self-distillation method with Confidence Reweighting (CR) for learning from unpaired images, and 3) a pseudo feature generation method based on Gradient Descent for learning from unpaired cap-

tions.

2. Related Work

2.1. Image Captioning with Unpaired Data

Different methods have been proposed to train image captioners with partially annotated image-caption pairs [7, 8, 18, 19, 21, 40] or with only unpaired data [13, 33]. Chen *et al.* [8] first exploits external language data to improve the performance of image captioners. The cross-domain image captioner [7] uses adversarial training to transfer a supervised image captioner to a target domain without paired training data. Non-autoregressive image captioning [22] formulates a multi-agent reinforcement learning system to cooperatively maximize a sentence-level reward in a semi-supervised setting. Liu *et al.* [40] proposes a self-retrieval approach to make use of unpaired images. Gu *et al.* [18] captures the characteristics of an image captioner from a pivot language and aligns it with the target language. Kim *et al.* [28], Yang *et al.* [13], Laina *et al.* [33] and Gu *et al.* [19] use Generative Adversarial Networks (GANs) to generate pseudo images from captions or to project images and captions into a common latent space [20, 65]. Guo *et al.* [21] implements a concepts-to-sentence memory translator correlating the relational reasoning between visual concepts and generated captions. Different from these works, our ensemble-based self-distillation approach focuses on maximizing the use of existing data for few-shot image captioning. It does not require additional models to generate pseudo features or searching for the unpaired captions over the training set, which is more efficient than previous methods.

2.2. Novel Object Captioning

A related problem to few-shot image captioning is novel object captioning. It aims to describe images of objects absent from training data. Novel object captioning methods highly depend on externally trained image taggers or object detectors [37, 60, 64] to generalize pretrained image captioners for describing near-domain or out-of-domain images [2]. Deep Compositional Captioner (DCC) [24] and Novel Object Captioner (NOC) [53] are the first methods to address this problem with the incorporation of external knowledge. Following a template-based framework, Lu *et al.* [41], Wu *et al.* [56], and Feng *et al.* [12] further propose Neural Baby Talk (NBT), Decoupled Novel Object Captioner (DNOC) and Cascaded Revision Network (CRN), respectively. Mogadala *et al.* [42], Yao *et al.* [60] and Li *et al.* [37] propose a copying mechanism with knowledge guided attention, LSTM with Copying (LSTM-C) and LSTM with Pointing (LSTM-P), respectively. Finally, inference-time strategies are proposed for generating sentences with specific novel objects, namely image Captions with Guiding Objects (CGO) [64] and Constrained Beam Search (CBS) [4]. These novel object captioning methods assume the absence of training data for novel objects, but for few-shot image captioning, such knowledge is available in the unpaired training data and can be learned with semi-supervised learning methods. Therefore, instead of focusing on the use of external knowledge and specialized architectures for describing novel objects, our ensemble-based self-distillation approach is more general and more feasible to address the few-shot image captioning problem.

2.3. Ensemble-Based Semi-Supervised Learning

Our work is also related to a number of ensemble-based semi-supervised learning approaches [34, 51, 62]. Ensemble [14] is commonly used in semi-supervised learning to generate pseudo labels for unlabeled data, which has been applied in various tasks, such as object detection [9, 63], person re-identification [16], machine translation [55] and natural language inference [58]. For example, in image classification, multi-model ensemble [62] uses different base models to generate pseudo class labels and integrates their results, whereas temporal ensemble [34, 51] integrates models at different training iterations to generate pseudo labels. Despite the success of these methods, their application in the image captioning task has not been explored, and how to utilize unpaired data from both vision and language domains remains an open question. In this work, we bridge this gap by generating pseudo captions and pseudo image features based on an integration of multi-model ensemble and temporal ensemble.

3. Approach

The goal of few-shot image captioning is to develop an image captioner $\boldsymbol{y} = F(\boldsymbol{x}|\boldsymbol{\theta})$ that generates a caption $\boldsymbol{y} = \{y_1, \ldots, y_t\}$ to describe an input image $\boldsymbol{x}$. Its parameters $\boldsymbol{\theta}$ can be jointly optimized on three datasets: a scarcely annotated set of image-caption pairs $\mathcal{D}_{x,y} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{N_{x,y}}$, a set of unpaired images $\mathcal{D}_x = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N_x}$, and a set unpaired captions $\mathcal{D}_y = \{\boldsymbol{y}^{(i)}\}_{i=1}^{N_y}$. The training of an image captioner can be achieved by minimizing the loss function

$$\mathcal{L} = \mathcal{L}_{x,y} + \lambda_x \mathcal{L}_x + \lambda_y \mathcal{L}_y, \tag{1}$$

where $\mathcal{L}_{x,y}$, $\mathcal{L}_x$, and $\mathcal{L}_y$ are computed on the three datasets $\mathcal{D}_{x,y}$, $\mathcal{D}_x$, $\mathcal{D}_y$, respectively, and λ_x and λ_y balance the weights of the corresponding loss terms.

In the rest of this section, we introduce our ensemble method (see Section 3.1) to generate pseudo captions of the unpaired images (see Section 3.2) and pseudo features of the unpaired captions (see Section 3.3). A complete summary of our algorithm is described in the Supplementary Materials.

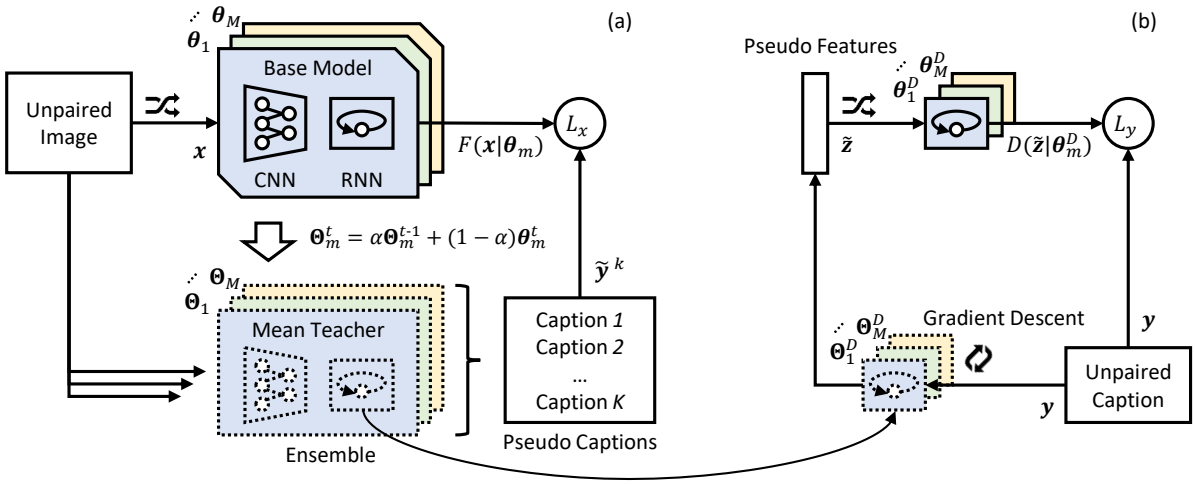

Figure 1. Overview of the ensemble-based self-distillation method. We train M encoder-decoder networks (*i.e.,* base models) and build an ensemble to generate pseudo captions and pseudo features. (a) We feed the unpaired images to the base models and obtain K output captions from the ensemble. These captions are used in later iterations of the training process as pseudo captions, with their normalized confidences as the weights of loss. (b) Given an unpaired caption, we use Gradient Descent to find the optimal latent features for the ensemble to generate the caption. The M Mean Teacher models are sequentially selected during the optimization.

3.1. Ensemble Method

Assume that an image captioner is designed following the encoder-decoder framework [54]. It is composed of an encoder $\boldsymbol{z} = E(\boldsymbol{x}|\boldsymbol{\theta}^E)$ that projects the input image $\boldsymbol{x}$ into a latent feature vector $\boldsymbol{z}$, and a decoder $\boldsymbol{y} = D(\boldsymbol{z}|\boldsymbol{\theta}^D)$ that generates the output caption $\boldsymbol{y}$ from the vector $\boldsymbol{z}$. In practice, the encoder is commonly implemented using a Convolutional Neural Network (CNN), and the decoder is commonly implemented using a Recurrent Neural Network (RNN). To generate accurate yet diverse pseudo captions and pseudo features, we train M image captioners as base models and develop an ensemble out of the base models. Training on the paired data is achieved by minimizing the supervised loss

$$\mathcal{L}_{x,y} = \sum_{m=1}^{M} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_{x,y}} \ell_{\text{CE}}(\boldsymbol{y}, F(\boldsymbol{x}|\boldsymbol{\theta}_m)), \qquad (2)$$

where $\boldsymbol{\theta}_m$ is the parameters of the m-th base model, and $\ell_{\text{CE}}(\cdot,\cdot)$ measures the sequential cross entropy loss.

We develop an ensemble by computing a Mean Teacher [51] for each of the M base models, and the predictions of the M Mean Teachers are averaged into the final output. Mean Teacher is a temporal ensemble method that averages model weights instead of their outputs to prevent overfitting. Specifically, given the parameters $\boldsymbol{\theta}_m^t$ of the m-th base model at the t-th training iteration, the parameters for the m-th Mean Teacher are computed as

$$\boldsymbol{\Theta}_m^t = \alpha\boldsymbol{\Theta}_m^{t-1} + (1-\alpha)\boldsymbol{\theta}_m^t,\ m = 1,\ldots,M, \qquad (3)$$

where α is a smoothing coefficient, and the parameters at $t = 0$ are initialized as $\boldsymbol{\Theta}_m^0 = \boldsymbol{\theta}_m^0$.

Thus, the temporal and multi-model ensemble not only generates more accurate and robust captions than the base models, but also enables the generation of pseudo captions and pseudo features that can be used to train the base models. As shown in Figure 1, for the diversity of base models and the robustness of the ensemble, we randomly assign the generated pseudo captions and features to different base models. Specifically, at each iteration, we randomly split the batched training data into M blocks and train each base model with only one block of the data. Compared with other resampling strategies such as Monte Carlo cross-validation [10] or Bootstrap [50], the non-repeated M-fold splitting leads to relatively noisier samples, which improves the diversity of different base models and prevents them from generating similar captions. By iteratively including pseudo captions and features into the training set, this method is referred to as self-distillation.

3.2. Self-Distillation with Unpaired Images

Figure 1a shows the process of self-distillation with unpaired images. With beam search [30], the ensemble generates K pseudo captions $\{\tilde{\boldsymbol{y}}^1,\ldots,\tilde{\boldsymbol{y}}^K\}$ that describe the input image with different confidence levels (*i.e.,* sum of log-likelihood). Using less confident pseudo captions as training labels may result in error accumulation and gradually degrade the performance of the models. To address this issue, we propose the Confidence Reweighting (CR) method to assign different weights to each training sample according to the confidence of the pseudo captions. Given the k-th pseudo caption $\tilde{\boldsymbol{y}}^k = \{\tilde{y}_1^k,\ldots,\tilde{y}_{L_k}^k\}$, where L_k represents

its length, its confidence is computed as

$$s_k = \sum_{j=1}^{L_k} \log p(\tilde{y}_j^k | \tilde{y}_{1:j-1}^k, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_M), \quad k = 1, \ldots, K. \tag{4}$$

Given the confidence levels of all pseudo captions $\boldsymbol{s} = \{s_1, \ldots, s_k\}$, the weights $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_k\}$ can be obtained with a softmax normalization $\boldsymbol{\gamma} = \text{softmax}(\boldsymbol{s})$.

With the normalized weights, the unsupervised loss term for the unpaired images is defined as

$$\begin{aligned} \mathcal{L}_x = & M \sum_{m=1}^{M} \sum_{\boldsymbol{x} \in \mathcal{D}_x} \sum_{k=1}^{K} \gamma_k \mathbb{1}(m = n) \ell_{\text{MSE}}(\tilde{\boldsymbol{y}}^k, F(\boldsymbol{x} | \boldsymbol{\theta}_m)), \\ & n \sim \text{Cat}(M, \boldsymbol{p}), \end{aligned} \tag{5}$$

where $\ell_{\text{MSE}}(\cdot, \cdot)$ is the distillation loss [25, 49, 51], $\mathbb{1}(\cdot)$ is the indicator function, and $\text{Cat}(M, \boldsymbol{p})$ is the categorical distribution with probabilistic parameter $\boldsymbol{p}$.

The proposed CR method allows the base models to avoid accumulating the errors while learning from less confident pseudo captions. Unlike the adversarial semi-supervised learning [28] and the fluency-guided cross-lingual approach [35], our method directly uses the log-likelihood from the existing decoder to balance the weights of the generated captions. Therefore, without additional parameters, it avoids overfitting due to the few paired data.

3.3. Self-Distillation with Unpaired Captions

To include the unpaired captions in the training of base models, we propose to generate pseudo image features by applying Gradient Descent to the trained ensemble (see Figure 1b). Previous studies typically use GANs [17, 65] to generate pseudo image features [13, 28, 33] or model parameters [48], which is less effective under the few-shot condition. Differently, our method can effectively generate valid pseudo features without introducing additional parameters, avoiding overfitting the few training examples.

Suppose we have learned M base-model decoders with their parameters $\{\boldsymbol{\theta}_1^D, \ldots, \boldsymbol{\theta}_m^D\}$, as well as their Mean Teacher parameters $\{\boldsymbol{\Theta}_1^D, \ldots, \boldsymbol{\Theta}_m^D\}$. Given an unpaired caption $\boldsymbol{y}$, we aim to find its corresponding latent features

$$\tilde{\boldsymbol{z}} = arg\,min_{\boldsymbol{z}} \sum_{m=1}^{M} \ell_{\text{CE}}(\boldsymbol{y}, D(\boldsymbol{z} | \boldsymbol{\Theta}_m^D)). \tag{6}$$

Starting from a Gaussian noise $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, we sequentially select one of the Mean Teachers to calculate the sequential cross entropy and update the features $\boldsymbol{z}$ using Gradient Descent:

$$\boldsymbol{z} := \boldsymbol{z} - \eta \frac{\partial \ell_{\text{CE}}(\boldsymbol{y}, D(\boldsymbol{z} | \boldsymbol{\Theta}_m^D))}{\partial \boldsymbol{z}}, \tag{7}$$

where η is the learning rate of this inner optimization problem. This strategy can reduce the computational complexity and improve the robustness of self-distillation. Its convergence can be guaranteed by the online learning with non-convex losses [15]. The optimal features $\tilde{\boldsymbol{z}}$ can be used as a pseudo feature vector to train the base models, so that they can generate more fluent and accurate captions. Thus, the unsupervised loss term for unpaired captions is denoted as

$$\begin{aligned} \mathcal{L}_y = & M \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{D}_y} \mathbb{1}_n(m = n) \ell_{\text{CE}}(\boldsymbol{y}, D(\tilde{\boldsymbol{z}} | \boldsymbol{\theta}_m^D)), \\ & n \sim \text{Cat}(M, \boldsymbol{p}). \end{aligned} \tag{8}$$

4. Experiments

In this section, we report our experiments and results to demonstrate the effectiveness of the proposed approach. First, we introduce datasets, evaluation, and implementation details. Next, we conduct quantitative comparisons with the state of the art and various baselines. Finally, we present the qualitative results, and analyze the model complexity.

4.1. Datasets and Evaluation

Our experiments are mainly conducted on the Karpathy splits [26] of the Microsoft COCO dataset [39], with 113k training images, 5k validation images, and 5k test images. Following [28], we randomly sample 1% of the image-caption pairs for training, and use the rest images and captions as unpaired training data.

In addition, following the practices of [13] and [28], we introduce Shutterstock [1] as an external sentence corpus. We use $2,322,628$ distinct image descriptions from this website as the unpaired captions and randomly sample a small portion of image-captions pairs from the COCO dataset. The remaining training images from the COCO dataset are used as the unpaired training data.

We use the common evaluation metrics for image captioning: BLEU [43], METEOR [6], ROUGE [38], CIDEr [52], SPICE [3] and WMD [27]. Since CIDEr [52] is well-accepted to measure the information and smoothness of the sentences, the hyper-parameters of our models are tuned on the validation set for the best CIDEr, and the final results are evaluated on the test set.

4.2. Implementation Details

We use the Neural Image Caption (NIC) [54] with a ResNet-101 [23] backbone as our base model, as well as the Att2in2 [46] and the Up-Down [5] models. We train the models using a minibatch size of 50 and the Adam [29] optimizer with learning rate 2.5×10^{-3}. We initialize the hyperparameters λ_x and λ_y as 0, and linearly increase them to $\lambda_x = 0.1$ and $\lambda_y = 1$ with the number of epochs. For inner optimization of the latent feature vector $\tilde{\boldsymbol{z}}$, we run the

Method	Base Model	COCO test								
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE	WMD
Adversarial Learning [28]	NIC [54]	63.0	-	-	18.7	20.7	-	55.2	-	-
Pseudo Label [36]	NIC [54]	63.3	44.6	20.9	21.3	19.4	46.0	57.2	12.0	14.0
Deep Mutual Learning [62]	NIC [54]	63.7	44.9	31.1	21.6	19.5	46.2	58.3	12.3	14.1
Pivoting [18]	NIC [54]	46.2	24.0	11.2	5.4	13.2	-	17.7	-	-
GAN [13]	NIC [54]	58.9	40.3	27.0	18.6	17.9	43.1	54.9	11.1	-
SME [33]	NIC [54]	-	-	-	19.3	20.2	45.0	61.8	12.9	-
SGA [19]	SGAE [59]	67.1	47.8	32.3	21.5	20.9	47.2	69.5	15.0	-
Ours	NIC [54]	64.5	45.9	32.1	22.5	20.0	46.7	62.4	12.7	14.7
Ours	Att2in2 [46]	66.9	48.6	34.5	24.3	20.8	48.2	66.3	13.2	15.4
Ours	Up-Down [5]	67.9	49.8	35.4	25.0	21.7	49.3	**73.0**	14.5	16.6

Table 1. Quantitative comparisons on the COCO test set between our method and state-of-the-art semi-supervised [28, 36, 62] and unsupervised [13, 18, 19, 33] image captioners.

Adagrad [11] optimizer for $N = 100$ iterations with learning rate $\eta = 1$. The standard derivation of the initialized pseudo feature is $\sigma = 0.1$. We set the smoothing coefficient $\alpha = 0.99$ for the Mean Teacher method and the weight decay as 0.0005. The total epoch is set as 100, where the first two epochs are used to pretrain the models on the unpaired captions only. We also use $M = 3$ base models to form our ensemble model and set beam search size as $K = 5$. The elements of $\boldsymbol{p}$ for the categorical distribution are set as $1/M$. We implement our experiments in PyTorch [44].

4.3. Quantitative Results

Quantitative results on the COCO dataset. We compare our method with state-of-the-art approaches on COCO test set. First, our method is compared with three few-shot image captioning methods (see the first panel of Table 1): Adversarial Learning [28] uses a GAN model to match the distribution of latent feature from images and captions. Pseudo Label [36] is a conventional semi-supervised classification model. Deep Mutual Learning [62] is an ensemble of students learning collaboratively and teaching each other throughout the training process. For a fair comparison with our method, we have similarly applied ensembles to the Pseudo Label [36] and Deep Mutual Learning [62] methods with $M = 3$ and $\alpha = 0.99$, and adapted them for the few-shot image captioning task. In addition, we also compare our method with four unsupervised methods (see the second panel of Table 1): Pivoting [18] uses a joint learning framework with an image-to-pivot captioning model and a pivot-to-target neural machine translation model. GAN [13] generates an adversarial caption, reconstructs an alignment of visual and sentence embedding space and uses gradient policy to optimize the awards. SME [33] aligns images and sentences in a shared latent representation structured through visual concepts. SGA [19] presents a scene graph-based approach for unpaired image captioning. For our proposed method, we use three image captioners as the base model, *e.g.,*NIC [54], Att2in2 [46] and Up-Down [5] (see the last panel of Table 1) and compare their performances with the state of the arts.

As shown in Table 1, most of the compared methods are based on the NIC [54] model, while only SGA uses the state-of-the-art Auto-Encoding Scene Graphs (SGAE) [59] and policy gradient [46] to improve its performance. In comparison with SME [33], the best NIC-based approach, our method performs 1.0% better in CIDEr (from 61.8 to 62.4) using the same base model. Different from SME [33] that depends on external object detectors, our method simply depends on the base models, without any externally trained models. Therefore, our approach can be easily applied to different state-of-the-art image captioners to handle the few-shot situation. For example, our method based on the Up-Down base model performs significantly better than the SGA method [19], while the Up-Down model [5] is inferior to the SGAE [59] captioner used in SGA [19]. Further comparisons across different base models are reported in the Supplementary Materials.

Quantitative results on the Shutterstock dataset. Further, following the practices of [13] and [28], we use the image descriptions from Shutterstock as unpaired captions and test the model performances on the COCO test set. In Table 2, we compare our method with three state-of-the-art semi-supervised [28] and unsupervised [13, 21] image captioners, including Adversarial Learning [28], GAN [13], and R^2M [21]. Our method significantly outperforms these methods, even with only 0.5% of paired data (*i.e.,*566 image-caption pairs). The performance of our method increases with the number of paired data, while being consistently better than the Adversarial Learning [28] method. Interestingly, the discrepancy between the unpaired Shutterstock captions and COCO images affects our method and Adversarial Learning [28] differently. On one hand, with 1% paired data, replacing the unpaired COCO captions with the Shutterstock captions only causes a minor degradation

Method	Paired	COCO test								
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE	WMD
GAN [13]	0%	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1	-
R^2M [21]		44.0	25.4	12.7	6.4	13.0	31.3	29.0	9.1	-
Adversarial Learning [28]	0.5%	-	-	-	5.4	12.0	34.6	10.5	4.2	-
Ours		61.9	42.4	28.5	18.9	17.3	44.6	**46.5**	9.8	11.5
Adversarial Learning [28]	0.8%	-	-	-	12.2	15.1	41.6	29.0	7.6	-
Ours		63.9	44.9	31.0	21.1	18.8	46.1	**54.7**	11.6	13.2
Adversarial Learning [28]	1%	-	-	-	15.2	16.9	43.3	39.7	9.4	-
Ours		64.1	45.2	31.3	21.5	19.3	46.4	**58.4**	12.1	14.1

Table 2. Quantitative comparisons on the COCO test set between our method and state-of-the-art semi-supervised [28] and unsupervised [13, 21] image captioners trained with Shutterstock captions. Both our method and Adversarial Learning [28] are trained with $0.5 - 1\%$ of COCO image-caption pairs in addition to the unpaired COCO images and Shutterstock captions. The unsupervised GAN [13] and R^2M [21] methods use an additional set of 36 million images from the OpenImage [32] dataset.

Method	Data			COCO test								
	P	UI	UC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	ROUGE_L	CIDEr	SPICE	WMD
Mean Teacher (P)	✓			62.0	43.1	29.4	20.1	18.7	45.1	53.8	11.4	13.4
Mean Teacher (P+UI)	✓	✓		63.0	43.9	30.1	20.6	18.9	45.6	55.8	11.6	13.7
Mean Teacher (P+UC)	✓		✓	62.5	43.5	29.8	20.5	19.8	45.2	56.1	12.0	13.7
Mean Teacher (P+UI+UC)	✓	✓	✓	62.8	44.2	30.6	21.3	19.5	45.5	59.3	12.2	14.3
Ours (P)	✓			62.9	44.1	30.4	20.9	19.2	45.7	56.2	11.9	13.7
Ours (P+UI) w/o CR	✓	✓		64.2	45.2	31.1	21.3	19.4	46.2	58.0	12.1	13.9
Ours (P+UI)	✓	✓		63.8	45.0	31.2	21.6	19.6	46.3	58.7	12.3	14.2
Ours (P+UC)	✓		✓	64.4	45.6	31.9	22.2	19.9	46.7	60.4	12.5	14.4
Ours (P+UI+UC) w/o CR	✓	✓	✓	64.3	45.8	32.1	22.4	19.9	46.5	60.7	12.5	14.5
Ours (P+UI+UC)	✓	✓	✓	64.5	45.9	32.1	22.5	20.0	46.7	**62.4**	12.7	14.7
Ours+ (Visual Genome)	✓	✓✓	✓	65.2	46.9	33.0	23.3	20.4	47.6	64.9	13.1	15.1
Ours+ (Unlabeled COCO)	✓	✓✓	✓	65.8	47.5	33.5	23.6	20.6	47.9	**65.3**	13.3	15.4

Table 3. Quantitative comparison with various baselines on the COCO test set. The baselines are trained with different data settings, including paired data (P), unpaired images (UI), and unpaired captions (UC). The Ours+ model is trained with additional unpaired images from the Visual Genome [31] dataset or the COCO [39] dataset.

(6.4% in CIDEr, from 62.4 to 58.4), while that of Adversarial Learning [28] is 28.1% (from 55.2 to 39.7). On the other hand, when the percentage of paired data decreases from 1% to 0.5%, our method has a less significant performance drop (20.4% in CIDEr, from 58.4 to 46.5) than Adversarial Learning [28] (73.6% in CIDEr, from 39.7 to 10.5). Since Adversarial Learning [28] adopts a pseudo-label assignment strategy to utilize the unpaired data, it cannot avoid the error propagation due to the domain discrepancy between COCO images and Shutterstock captions, especially with few paired data. Differently, our method does not rely on a matching mechanism, but uses ensemble models and gradient descent to prevent models from the severe error propagation. In practice, it allows real-world applications to collect unpaired images and captions from different domains, which can effectively reduce the labor-intensive efforts of data collection.

Effects of multi-model ensemble. Since our method is based on an ensemble of multiple Mean Teachers, when $M = 1$, our ensemble model degrades to a single Mean Teacher model. As shown in Table 3, Mean Teacher (P+UI+UC) with fewer parameters has already outperformed the state-of-the-art methods. Hence, we compare our method with Mean Teacher as a strong baseline, to demonstrate the effects of multi-model ensemble. As shown in Table 3, with paired data only (P), the multi-model ensemble leads to a significant improvement of 4.5% in CIDEr (from 53.8 to 56.2); with both paired data and unpaired images (P+UI), the improvement is 5.2% (from 55.8 to 58.7); with both paired data and unpaired captions (P+UC), the improvement is 7.7% (from 56.1 to 60.4); with all paired and unpaired data (P+UI+UC), the improvement is 5.2% (from 59.3 to 62.4). These improvements suggest the effectiveness of the integration of temporal ensemble and multi-model ensemble.

Further, we observe that the performance of our method (P+UI+UC) increases when $M \leq 5$ and then start to decrease when $M \geq 5$ (see Figure 2a). We also observe

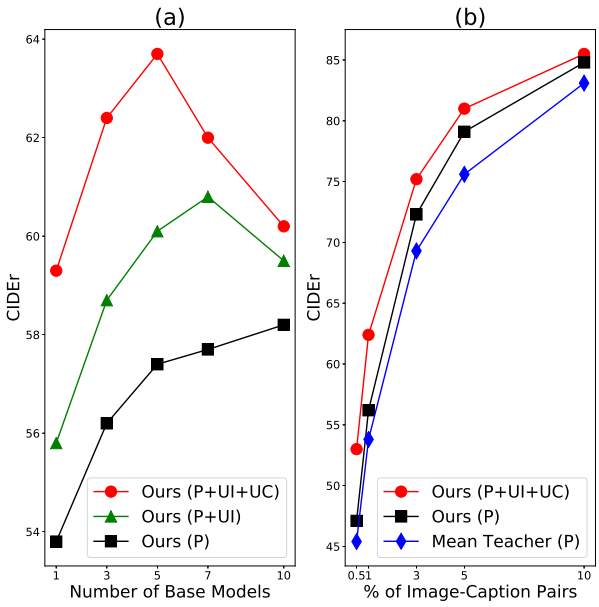

Figure 2. Results on the COCO test set w.r.t (a) different number of base models from the ensemble given 1% paired data and (b) different ratios of paired training data with a fixed number of base models $M = 3$.

similar trends in our (P+UI) baseline. It suggests that error propagation may occur due to the distillation with the pseudo captions/features instead of ground-truth supervision. Training with only paired data does not suffer from error propagation because it is completely supervised. Interestingly, even if we use 10 base models with paired data only (P), its performance cannot surpass 3 base models with unpaired images (P+UI) or 1 base models with all the unpaird data (P+UI+UC), which demonstrate the effectiveness of our approach in leveraging unpaired data.

Finally, as shown in Figure 2b, with a fixed number of the models $M = 3$, the performance improvements between different baselines are consistent across various ratios of paired data from 0.5% to 10%, suggesting the effectiveness of our method and the inclusion of unpaired data.

Effects of self-distillation with unpaired images and captions. By including the unpaired images, our self-distillation method increases its CIDEr score from 56.2 to 58.7 (*i.e.,* Ours (P+UI) in Table 3). Similarly, by only including the unpaired captions, our approach also achieves a significant improvement in CIDEr (from 56.2 to 60.4). Moreover, with the pseudo features generated from unpaired captions, our method achieves a remarkable 6.3% improvement in CIDEr (from 58.7 to 62.4), suggesting that the knowledge distilled from the pseudo features is effective. To further verify the performance gain from the unpaired images, we additionally include 112k Visual Genome [31] training images or 123k COCO unlabeled images in the experiments. These images provide abundant knowledge to improve the generalization of the model. As shown Table 3, Visual Genome and COCO unlabeled images further increase the model's CIDEr scores to 64.9 and 65.3, respectively. The improvements confirm the effectiveness of pseudo caption generation and suggests potential benefits from larger image datasets.

Note that the main technical novelty of our self-distillation method is the Confidence Reweighting (CR) applied to the pseudo captions. As shown in Table 3, the performance degrades to CIDEr=58.0 without using this method (*i.e.,*by averaging the cross entropy). The difference is also significant when unpaired captions are added to the training data (2.8% in CIDEr, from 60.7 to 62.4), which suggests the effectiveness of CR in protecting model convergence from the propagation of captioning errors. Moreover, we observe that the confidence levels of generated pseudo captions are strongly correlated with their CIDEr scores. For the top-5 most confident captions, their average CIDEr scores on the COCO validation set are 61.3, 59.5, 59.3, 57.8, and 57.6, respectively. The most confident caption has the highest CIDEr score, while the least confident caption has the lowest score. This correlation justifies the use of confidence level as a reweighting strategy in the distillation process.

Analysis of SPICE F-scores. To further understand the contributions of each technical component, we report a breakdown of SPICE F-scores over various subcategories on the COCO test set. As illustrated in the first panel and the second panel in Table 4, after adding unpaired images and unpaired captions, the generated captions are more comprehensiveness in relationships between objects, counting number, sizes and objects. For example, the higher score on objects suggests that the generated captions can describe the object in the image more precisely. For more detailed ablation studies, please refer to the Supplementary Materials.

4.4. Qualitative Analysis

In addition to the quantitative results, we further demonstrate the effectiveness of our method with qualitative examples on the COCO validation set. Figure 3 compares the results of our method (P+UI+UC) with the two state-of-the-art approaches: Pseudo Label [36] and Deep Mutual Learning [62], as well as two baseline methods: Ours (P) and Ours (P+UI). The Adversarial Learning [28] method is not compared because of the inaccessibility of source code. As shown in Figure 3, the Pseudo Label and Deep Mutual Learning methods, as well as the baselines, can describe part of the scenes correctly (*e.g.,* 'broccoli' and 'beach'), but fail to describe the image contents completely (*e.g.,* missing the 'meat' and 'surfboard'). By training with images and unpaired captions, our approach generates more accurate and fluent captions to describe the input images, and objects in the images are correctly recognized (*e.g.,* 'meat' and 'surfboard'). It also verifies that the F-score of objects of our method is higher than that of other baselines. The qualita-

Method	Data			COCO test						
	P	UI	UC	SPICE	Relation	Cardinality	Attribute	Size	Color	Object
Mean Teacher (P)	✓			11.4	2.3	0.8	2.8	0.9	2.5	23.1
Mean Teacher (P+UI)	✓	✓		11.6	2.6	1.0	2.8	1.5	2.3	23.5
Mean Teacher (P+UC)	✓		✓	11.9	2.3	**2.2**	**3.6**	1.4	**3.8**	23.9
Mean Teacher (P+UI+UC)	✓	✓	✓	**12.2**	**2.9**	1.2	**3.6**	**2.1**	1.5	**24.3**
Ours (P)	✓			11.9	2.6	0.5	2.9	1.3	3.2	24.1
Ours (P+UI) w/o CR	✓	✓		12.1	2.6	0.6	3.1	1.4	3.4	24.3
Ours (P+UI)	✓	✓		12.3	2.7	0.6	3.2	1.5	3.5	24.6
Ours (P+UC)	✓		✓	12.6	2.6	**2.2**	**3.6**	1.3	3.5	**25.3**
Ours (P+UI+UC) w/o CR	✓	✓	✓	12.5	2.7	1.0	3.4	**1.6**	3.5	25.1
Ours (P+UI+UC)	✓	✓	✓	**12.7**	**2.8**	1.4	**3.6**	**1.6**	**3.6**	**25.3**
Ours+ (Visual Genome)	✓	✓✓	✓	13.1	3.1	1.0	3.1	1.5	1.3	26.2
Ours+ (Unlabeled COCO)	✓	✓✓	✓	**13.3**	**3.2**	**1.4**	**3.2**	**2.0**	**1.9**	**26.6**

Table 4. Breakdown of SPICE F-scores over various subcategories on the COCO test set.

```
Pseudo Label:           A close up of a plate of broccoli and vegetables.
Deep Mutual Learning:   A plate of food with broccoli and broccoli.
Ours (P):               A plate of food with broccoli and broccoli.
Ours (P+UI):            A plate of food with broccoli and vegetables.
Ours (P+UI+UC):         A plate of food with meat and broccoli.
Ground-truth:           A plate with meat chops, broccoli and pasta on it
```

```
Pseudo Label:           A man flying a kite on the beach.
Deep Mutual Learning:   A man holding a kite on a beach.
Ours (P):               A man is flying a kite on a beach.
Ours (P+UI):            A man holding a skateboard on a beach.
Ours (P+UI+UC):         A person holding a surfboard on a beach.
Ground-truth:           A guy on a beach holding a surf board.
```

Figure 3. Qualitative examples of our model and various baselines on COCO validation set.

tive results suggest the effectiveness of using unpaired data in few-shot image captioning. For more qualitative results, please refer to the Supplementary Materials.

4.5. Model Complexity

Despite its significantly improved performance, compared with non-ensemble approaches, the proposed method is relatively more complex in model size and computational cost. Its complexity is mostly decided by the base model architecture, the number of base models M, and the number of iterations N for the inner optimization of $\boldsymbol{z}$. On the one hand, the model size increase linearly with M, and the training time increases with $\mathcal{O}(M+N)$. In our experiments, we set $N = 100$ and $M = 3$, which takes a total of ~ 3 hours for training on a single NVIDIA 2080 Ti GPU. To accelerate the training for faster deployment, a smaller N can be adopted. With a reduced $N = 20$, our method can achieve a 61.2 CIDEr score with $M = 3$ by converging in only ~ 1 hour. On the other hand, to reduce the inference time, one can selectively choose any number of the base models to form a new ensemble. For example, a base model can achieve 59.6 ± 0.09 CIDEr scores, and an ensemble of two base models can achieve 61.4 ± 0.08 CIDEr scores, significantly better than the state-of-the-art approaches.

5. Conclusion

In this paper, we have introduced an ensemble-based self-distillation method for image captioning with few paired data and a large number of unpaired images and captions. It is an effective method to generate accurate and robust pseudo captions and pseudo features, and use them to train the base models. Our approach significantly outperforms the state-of-the-art approaches, demonstrating its effectiveness of utilizing the unpaired images and captions. Future efforts will be focused on the exploration of the connection of the unpaired images and unpaired captions to make the best use of these unpaired datasets in few-shot image captioning and other related vision tasks.

References

[1] Shutterstock. https://www.shutterstock.com.

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. *European Conference on Computer Vision (ECCV)*, 2016.

[4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[5] 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.

[6] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. *Annual Conference of the Association for Computational Linguistics Workshop (ACLW)*, 2005.

[7] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[8] Wenhu Chen, Aurelien Lucchi, and Thomas Hofmann. A semi-supervised framework for image captioning. *CoRR, abs/1611.05321v3*, 2016.

[9] Xuanyi Dong, Liang Zheng, Fan Ma, Yi Yang, and Deyu Meng. Few-example object detection with model communication. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[10] Werner Dubitzky, Martin Granzow, and Daniel Berrar. *Fundamentals of Data Mining in Genomics and Proteomics*. Springer Science & Business Media, 2007.

[11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research (JMLR)*, 2011.

[12] Qianyu Feng, Yu Wu, Hehe Fan, Chenggang Yan, and Yi Yang. Cascaded revision network for novel object captioning. *CoRR, abs/1908.02726*, 2019.

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

[14] Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. *International Conference on Machine Learning (ICML)*, 2018.

[15] Xiang Gao, Xiaobo Li, and Shuzhong Zhang. Online learning with non-convex losses and non-stationary regret. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[16] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *International Conference on Learning Representations (ICLR)*, 2020.

[17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

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

[19] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[21] Dan Guo, Yang Wang, Peipei Song, and Meng Wang. Recurrent relational memory network for unsupervised image captioning. *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2020.

[22] Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2020.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *CoRR, abs/1503.02531*, 2015.

[26] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[27] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. *Annual Conference of the Association for Computational Linguistics (ACL)*, 2017.

[28] 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.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[30] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 2017.

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR, abs/1811.00982*, 2018.

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

[34] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *International Conference on Learning Representations (ICLR)*, 2017.

[35] Weiyu Lan, Xirong Li, and Jianfeng Dong. Fluency-guided cross-lingual image captioning. *ACM Multimedia (ACM-MM)*, 2017.

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

[37] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. *Annual Conference of the Association for Computational Linguistics Workshop (ACLW)*, 2004.

[39] 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.

[40] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. *European Conference on Computer Vision (ECCV)*, 2018.

[41] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] Aditya Mogadala, Umanga Bista, Lexing Xie, and Achim Rettinger. Describing natural images containing novel objects with knowledge guided assitance. *CoRR, abs/1710.06303*, 2017.

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. *Annual Conference of the Association for Computational Linguistics (ACL)*, 2002.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, Gregory Chanan James Bradbury, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[46] 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.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International journal of computer vision (IJCV)*, 2015.

[48] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, and Raia Hadsell Simon Osindero. Meta-learning with latent embedding optimization. *International Conference on Learning Representations (ICLR)*, 2019.

[49] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[50] Michael Steinbach, Pang-Ning Tan, and Vipin Kumar. *Introduction to Data Mining*. Pearson, 2nd edition, 2018.

[51] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[52] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[53] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[54] 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.

[55] Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Transductive ensemble learning for neural machine translation. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

[56] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. *ACM Multimedia (ACM-MM)*, 2018.

[57] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning (ICML)*, 2015.

[58] Yige Xu, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang. Improving bert fine-tuning via self-ensemble and self-distillation.

Association for the Advancement of Artificial Intelligence (AAAI), 2020.

[59] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[60] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[61] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[63] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[64] Yue Zheng, Yali Li, and Shengjin Wang. Intention oriented image captions with guiding objects. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[65] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE International Conference on Computer Vision (ICCV)*, 2017.