Image Captioning with Very Scarce Supervised Data: Adversarial Semi-Supervised Learning Approach

Dong-Jin Kim[1]   **Jinsoo Choi**[1]   **Tae-Hyun Oh**[2]   **In So Kweon**[1]

[1]KAIST, South Korea.   [2]MIT CSAIL, Cambridge, MA.

[1]{djnjusa,jinsc37,iskweon77}@kaist.ac.kr   [2]taehyun@csail.mit.edu

Abstract

Constructing an organized dataset comprised of a large number of images and several captions for each image is a laborious task, which requires vast human effort. On the other hand, collecting a large number of images and sentences separately may be immensely easier. In this paper, we develop a novel data-efficient *semi-supervised* framework for training an image captioning model. We leverage massive *unpaired* image and caption data by learning to associate them. To this end, our proposed semi-supervised learning method assigns pseudo-labels to unpaired samples via Generative Adversarial Networks to learn the joint distribution of image and caption. To evaluate, we construct scarcely-paired COCO dataset, a modified version of MS COCO caption dataset. The empirical results show the effectiveness of our method compared to several strong baselines, especially when the amount of the paired samples are scarce.

1 Introduction

Image captioning is a task of automatically generating a natural language description of a given image. It is a highly useful task, in that 1) it extracts the essence from an image into a self-descriptive form of representation, and 2) the output format is a natural language, which exhibits free-form and manageable characteristics useful to applications such as language based image or region retrieval [Johnson *et al.*, 2016; Karpathy and Fei-Fei, 2015; Kim *et al.*, 2019], video summarization [Choi *et al.*, 2018], navigation [Wang *et al.*, 2018a], vehicle control [Kim *et al.*, 2018b]. Image captioning exhibits free-form characteristics, since it is not confined to a few number of pre-defined classes. This enables descriptive analysis of a given image.

Recent research on image captioning has made impressive progress [Anderson *et al.*, 2018; Vinyals *et al.*, 2015]. Despite this progress, the
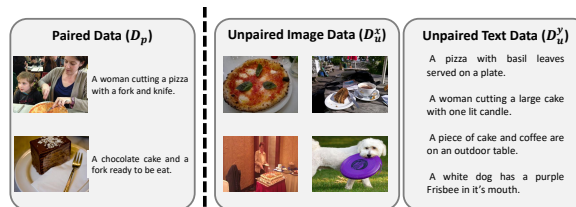

Figure 1: The proposed data setup utilizes both "paired" and "unpaired" image-caption data. We denote paired data as $\mathcal{D}_p$, and unpaired image and caption datasets as $\mathcal{D}_u^x$ and $\mathcal{D}_u^y$ respectively.

majority of works are trained via supervised learning that requires a large corpus of caption labeled images such as the MS COCO caption dataset [Lin *et al.*, 2014]. Specifically, the dataset is constructed with 1.2M images that were asked the annotators to provide five grammatically plausible sentences for each image. Constructing such human-labeled datasets are an immensely laborious task and time-consuming. This is a challenge of image captioning, because the task heavily requires large data, *i.e.*, data hungry task.

In this work, we present a novel way of leveraging *unpaired* image and caption data to train data hungry image captioning neural networks. We consider a scenario in which we have a small scale paired image-caption data of a specific domain. We are motivated by the fact that images can be easily obtained from the web and captions can be easily augmented and synthesized by replacing or adding different words for given sentences as done in [Zhang *et al.*, 2015]. Moreover, given a sufficient amount of descriptive captions, it is easy to crawl *corresponding but noisy* images through Google or Flickr image databases [Thomee *et al.*, 2016] to build an image corpus. In this way, we can easily construct a scalable *unpaired* dataset of images and captions, which requires no (or at least minimal) human effort.

Due to the unpaired nature of images (input) and captions (output supervision), the conven-

tional supervision loss can no longer be used directly. We propose to algorithmically assign supervision labels, termed as *pseudo-labels*, to make unpaired data paired. The pseudo-label is used as a *learned* supervision label. To develop the mechanism of pseudo-labeling, we use generative adversarial network (GAN) [Goodfellow *et al.*, 2014], for searching pseudo-labels for unpaired data. That is, in order to find the appropriate pseudo-labels for unpaired samples, we utilize an adversarial training method for training a discriminator model. Thereby, the discriminator learns to distinguish between real and fake image-caption pairs and to retrieve pseudo labels as well as to enrich the captioner training.

Our main contributions are summarized as follows. (1) We propose a novel framework for training an image captioner with the unpaired image-caption data and a small amount of paired data. (2) In order to facilitate training with unpaired data, we devise a new semi-supervised learning approach by the novel usage of the GAN discriminator. (3) We construct our *scarcely-paired COCO* dataset, which is the modified version of the MS COCO dataset without pairing information. On this dataset, we show the effectiveness of our method in various challenging setups, compared to strong competing methods.

2 Related Work

The goal of our work is to deal with unpaired image-caption data for image captioning. Therefore, we mainly focus on image captioning and unpaired data handling literature.

Data Issues in Image Captioning. Since the introduction of large-scale datasets such as MS COCO data [Lin *et al.*, 2014], image captioning has been extensively studied in vision and language society [Anderson *et al.*, 2018; Rennie *et al.*, 2017; Vinyals *et al.*, 2015; Xu *et al.*, 2015] by virtue of the advancement of deep convolutional neural networks [Krizhevsky *et al.*, 2012]. As neural network architectures become more advanced, they require much larger dataset scale for generalization [Shalev-Shwartz and Ben-David, 2014]. Despite the extensive study on better network architectures, the data issues for image captioning, such as noisy data, partially missing data, and unpaired data have been relatively barely studied.

Unpaired image-caption data has been just recently discussed. Gu *et al.* [2018] introduce

a third modal information, Chinese captions, for language pivoting [Utiyama and Isahara, 2007; Wu and Wang, 2007]. Feng *et al.* [2019] propose an unpaired captioning framework which trains a model without image or sentence labels via learning a visual concept detector with external data, OpenImage dataset [Krasin *et al.*, 2017]. Chen *et al.* [2017] approach image captioning as a domain adaptation by utilizing large paired MS COCO data as the source domain and adapting on a separate unpaired image or caption dataset as the target domain. Liu *et al.* [2018] use self-retrieval for captioning to facilitate training a model with partially labeled data, where the self-retrieval module tries to retrieve corresponding images based on the generated captions. As a separate line of work, there are novel object captioning methods [Anne Hendricks *et al.*, 2016; Venugopalan *et al.*, 2017] that additionally exploit unpaired image and caption data corresponding to a novel word.

Most of aforementioned works [Gu *et al.*, 2018; Anne Hendricks *et al.*, 2016; Venugopalan *et al.*, 2017; Feng *et al.*, 2019] exploit large auxiliary supervised data which is often beyond image-caption data. To the best of our knowledge, we are the first to study how to handle unpaired image and caption data for image captioning without any auxiliary information but by leveraging scarce paired image-caption data only. Although [Chen *et al.*, 2017] does not use auxiliary information as well, it requires large amounts of paired source data, of which data regime is different from ours. [Liu *et al.*, 2018] is also this case, where they use the full paired MS COCO caption dataset with an additional large unlabeled image set. Our method lies on a very scarce paired source data regime, of which scale is only 1% of the COCO dataset.

Multi-modality in Unpaired Data Handling. By virtue of the advancement on generative modeling techniques, *e.g.*, GAN [Goodfellow *et al.*, 2014], multi-modal translation recently emerged as a popular field. Among many possible modalities, image-to-image translation between two different (and unpaired) domains has been mostly explored. To tackle this issue, the cycle-consistency constraint between unpaired data is exploited in CycleGAN [Zhu *et al.*, 2017] and DiscoGAN [Kim *et al.*, 2017], and it is further improved in UNIT [Liu *et al.*, 2017].

In this work, we regard image captioning as a multi-modal translation. Our work has similar

motivation to the unpaired image-to-image translation [Zhu *et al.*, 2017], and unsupervised machine translation [Artetxe *et al.*, 2018; Lample *et al.*, 2018a,b] works or machine translation with monolingual data [Zhang *et al.*, 2018]. However, we show that the cycle-consistency does not work on our problem setup due to a significant modality gap. Instead, our results suggest that the traditional label propagation based semi-supervised framework [Zhou *et al.*, 2004] is more effective for our task.

Semi-supervised Learning. Our method is motivated by the generative model based semi-supervised learning [Chongxuan *et al.*, 2017; Gan *et al.*, 2017; Kingma *et al.*, 2014], which mostly deals with classification labels. In contrast, we leverage caption data and devise a novel model to train the captioning model.

3 Proposed Framework

In this section, we first brief the standard image caption learning, and describe how we can leverage the unpaired dataset. Then, we introduce an adversarial learning method for obtaining a GAN model that is used for assigning pseudo-labels and encouraging to match the distribution of latent features of images and captions. Lastly, we describe a connection to CycleGAN.

3.1 Model

Let us denote a dataset with N_p image-caption pairs as $\mathcal{D}_p = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_p}$. A typical image captioning task is defined as follows: given an image $\mathbf{x}_i$, we want a model to generate a caption y_i that best describe the image. Traditionally, a captioning model is trained on a large paired dataset $(\mathbf{x}, y) \in \mathcal{D}_p$, *e.g.*, MS COCO dataset, by minimizing the negative log likelihood against the ground truth caption as follows:

$$\min_{\theta} \sum_{(\mathbf{x},y)\in\mathcal{D}_p} L_{\mathsf{CE}}(y, \mathbf{x}; \theta), \tag{1}$$

where $L_{\mathsf{CE}} = -\log q_\theta(y|\mathbf{x})$ denotes the cross entropy loss, θ the set of learnable parameters, and $q(\cdot)$ the probability output of the model. Motivated by encoder-decoder based neural machine translation literature [Cho *et al.*, 2014], traditional captioning frameworks are typically implemented as a encoder-decoder architecture [Vinyals *et al.*, 2015], *i.e.*, CNN-RNN. The CNN encoder $F(\mathbf{x}; \theta_{enc})$ produces a latent feature vector $\mathbf{z}^x$

from a given input image $\mathbf{x}$, and the RNN decoder $H(\mathbf{z}^x; \theta_{dec})$ is followed to generate a caption y in a natural language form from $\mathbf{z}^x$, as depicted in Fig. 2. The term $q(\cdot)$ in Eq. (1) is typically implemented by the sum of the cross entropy of each word token. For simplicity, we omit θ from here if deducible, *e.g.*, $H(\mathbf{z}^x)$.

Learning with Unpaired Data. Our problem deals with unpaired data samples, where the image and caption sets $\mathcal{D}_u^x = \{\mathbf{x}_i\}_{i=0}^{N_x}$ and $\mathcal{D}_u^y = \{y_i\}_{i=0}^{N_y}$ are not paired. Given the unpaired datasets, due to the missing supervision, the loss in Eq. (1) cannot be directly computed. Motivated by the nearest neighbor aware semi-supervised framework [Shi *et al.*, 2018], we artificially generate *pseudo-labels* for respective unpaired datasets.

Specifically, we retrieve a best match caption $\tilde{y}_i$ in $\mathcal{D}_u^y$ given a query image $\mathbf{x}_i$, and assign it as a pseudo-label, and vice versa ($\tilde{\mathbf{x}}_i$ for y_i). We abuse the pseudo-label as a function for simplicity, *e.g.*, $\tilde{y}_i = \tilde{y}(\mathbf{x}_i)$. To retrieve a semantically meaningful match, we need a measure to assess matches. We use a discriminator network to determine real or fake pairs by GAN, which we will describe later. With the retrieved pseudo-labels, now we can compute Eq. (1) by modifying it as:

$$\min_{\theta} \lambda_x \sum_{\mathbf{x}\in\mathcal{D}_u^x} L_{\mathsf{CE}}(\tilde{y}(\mathbf{x}), \mathbf{x}; \theta) + \lambda_y \sum_{y\in\mathcal{D}_u^y} L_{\mathsf{CE}}(y, \tilde{\mathbf{x}}(y); \theta), \tag{2}$$

where $\lambda_{\{\cdot\}}$ denote the balance parameters.

Discriminator Learning by Unpaired Feature Matching. We train the criterion to find a semantically meaningful match, so that pseudo-labels for each modality are effectively retrieved. We learn a discriminator for that by using a small amount of paired supervised dataset.

We adopt a caption encoder, $G(y; \theta_{cap})$, which embeds the caption y into a feature $\mathbf{z}^y$. This is implemented with a single layer LSTM, and we take the output of the last time step as the caption representation $\mathbf{z}^y$. Likewise, given an image x, we obtain $\mathbf{z}^x$ by the image encoder $F(\mathbf{x}; \theta_{enc})$. Now, we have a comparable feature space of $\mathbf{z}^x$ and $\mathbf{z}^y$. We utilize the discriminator to distinguish whether the pair $(\mathbf{z}^x, \mathbf{z}^y)$ from true paired data $(\mathbf{x}, y) \in \mathcal{D}_p$, *i.e.*, the pair belongs to the real distribution $p(\mathbf{x}, y)$ or not. We could use random pairs of $(\mathbf{x}, y)$ independently sampled from respective unpaired datasets, but we found that it is detrimental due to uninformative pairs. Instead, we conditionally
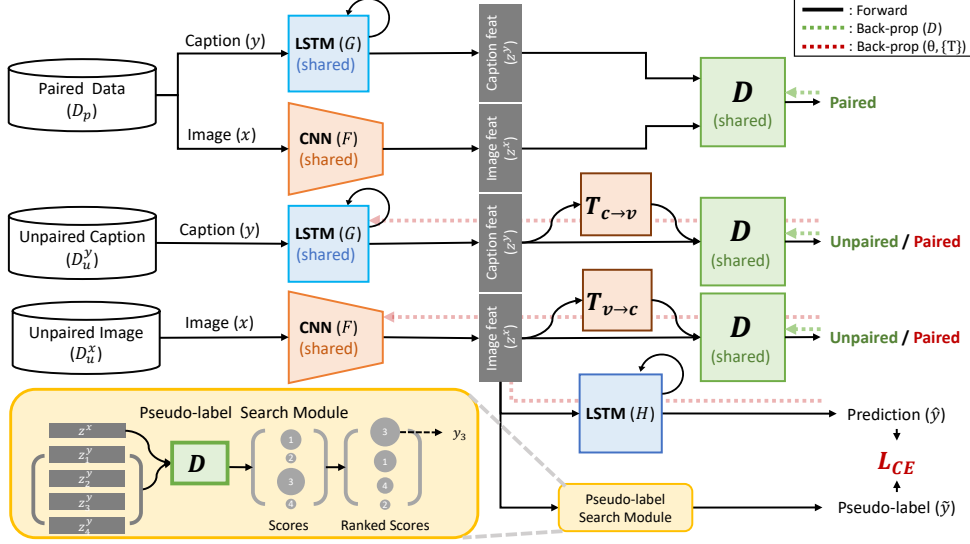
Figure 2: Description of the proposed method. Dotted arrows denote the path of the gradients via back-propagation. Given any image and caption pair, CNN and RNN (LSTM) encoders encode input image and caption into the respective feature spaces. A discriminator (D) is trained to discriminate whether the given feature pairs are real or fake, while the encoders are trained to fool the discriminator. The learned discriminator is also used to assign the most likely pseudo-labels to unpaired samples through the pseudo-label search module.

synthesize $\mathbf{z}^x$ or $\mathbf{z}^y$, to form a synthesized pair that appears to be as realistic as possible. We use the feature transformer networks $\tilde{\mathbf{z}}^y = T_{v \to c}(\mathbf{z}^x)$ and $\tilde{\mathbf{z}}^x = T_{c \to v}(\mathbf{z}^y)$, where $v \to c$ denotes the mapping from visual data to caption data and vice versa, and $\tilde{\mathbf{z}}$ denotes the conditionally synthesized feature. $\{T\}$ are implemented with multi-layer-perceptron with four FC layers with ReLU nonlinearity.

The discriminator $D(\cdot, \cdot)$ learns to distinguish features, real or not. At the same time, the other associated networks $F, G, T_{\{\cdot\}}$ are learned to fool the discriminator by matching the distribution of paired and unpaired data. Motivated by [Chongxuan et al., 2017], the formulation of this adversarial training is as follows:

$$\min_{\theta, \{T\}} \max_D \mathop{\mathbb{E}}_{\substack{(\mathbf{z}^x, \mathbf{z}^y) \\ \sim (F,G) \circ \mathcal{D}_p}} [\log(D(\mathbf{z}^x, \mathbf{z}^y)) + L_{reg}(\mathbf{z}^x, \mathbf{z}^y, \{T\})]$$

$$+ \frac{1}{2} \Bigg(\mathop{\mathbb{E}}_{\mathbf{x} \sim p(\mathbf{x})} [\log(1 - D(F(\mathbf{x}), T_{v \to c}(F(\mathbf{x}))))]$$

$$+ \mathop{\mathbb{E}}_{y \sim p(y)} [\log(1 - D(T_{c \to v}(G(y)), G(y)))] \Bigg),$$
(3)

where $L_{reg}(\mathbf{z}^x, \mathbf{z}^y, \{T\}) =$

$$\lambda_{reg}(\| \underset{v \to c}{T} (\mathbf{z}^x) - \mathbf{z}^y \|_F^2 + \| \mathbf{z}^x - \underset{c \to v}{T} (\mathbf{z}^y) \|_F^2),$$

and we use the distribution notation $p(\cdot)$ to flexibly refer to any type of the dataset regardless of $\mathcal{D}_p$ and $\mathcal{D}_u$. Note that the first log term is not used for updating any learnable parameters related to $\theta, \{T\}$, but only used for updating D. Through alternating training of the discriminator

(D) and generator ($\theta, \{T\}$) similar to [Kim et al., 2018a], the latent feature distribution of paired and unpaired data should be close to each other, i.e., $p(\mathbf{z}^x, \mathbf{z}^y) = p(\mathbf{z}^x, T_{v \to c}(\mathbf{z}^x)) = p(T_{c \to v}(\mathbf{z}^y), \mathbf{z}^y)$ (refer to the proof in [Chongxuan et al., 2017]). In addition, as the generator is trained, the decision boundary of the discriminator tightens. If the unpaired datasets are sufficiently large such that semantically meaningful matches exist between the different modality datasets, and if the discriminator D is plausibly learned, we can use D to retrieve a proper pseudo-label. Our architecture is illustrated in Fig. 2.

Pseudo-label Assignment. Given an image $\mathbf{x} \in \mathcal{D}_u^x$, we retrieve a caption in the unpaired dataset, i.e., $\tilde{y} \in \mathcal{D}_u^y$, that has the highest score obtained by the discriminator, i.e., the most likely caption to be paired with the given image as

$$\tilde{y}_i = \tilde{y}(\mathbf{x}_i) = \underset{y \in \mathcal{D}_u^y}{\arg\max} \ D\left(F(\mathbf{x}_i), G(y)\right), \quad (4)$$

vice versa for unpaired captions:

$$\tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}(y_i) = \underset{\mathbf{x} \in \mathcal{D}_u^x}{\arg\max} \ D\left(F(\mathbf{x}), G(y_i)\right). \quad (5)$$

By this retrieval process over all the unpaired dataset, we have image-caption pairs $\{(x_i, y_i)\}$ from the paired data and the pairs with pseudo-labels $\{(x_j, \tilde{y}_j)\}$ and $\{(\tilde{x}_k, y_k)\}$ from the unpaired

data. However, these pseudo-labels are not noise-free, thus treating them equally with the paired ones is detrimental. Motivated by learning with noisy labels [Lee *et al.*, 2018; Wang *et al.*, 2018b], we define a confidence score for each of the assigned pseudo-labels. We use the output score from the discriminator, as the confidence score, *i.e.*, $\alpha_i^x = \hat{D}(\mathbf{x}_i, \tilde{y}_i)$ and $\alpha_i^y = \hat{D}(\tilde{\mathbf{x}}_i, y_i)$, where we denote $\hat{D}(\mathbf{x}, y) = D(F(\mathbf{x}), G(y))$, and $\alpha \in [0, 1]$ due to the sigmoid function of the final layer of the discriminator. Therefore, we utilize the confidence scores to assign weights to the unpaired samples. We compute the weighted loss as follows:

$$\min_\theta \sum_{(\mathbf{x},y)\in\mathcal{D}_p} L_{\mathsf{CE}}(y, \mathbf{x}; \theta) + \lambda_x \sum_{\mathbf{x}\in\mathcal{D}_u^x} \alpha_i^x L_{\mathsf{CE}}(\tilde{y}(\mathbf{x}), \mathbf{x}; \theta)$$

$$+ \lambda_y \sum_{y\in\mathcal{D}_u^y} \alpha_i^y L_{\mathsf{CE}}(y, \tilde{\mathbf{x}}(y); \theta). \quad (6)$$

We jointly train the model on both paired and unpaired data. To ease the training further, we add an additional triplet loss function:

$$\sum_{\substack{(x_p,y_p)\in\mathcal{D}_p, \\ x_u\in\mathcal{D}_u^x, y_u\in\mathcal{D}_u^y}} \log \frac{p_\theta(y_p|x_p)}{p_\theta(y_p|x_u)} + \log \frac{p_\theta(y_p|x_p)}{p_\theta(y_u|x_p)}, \quad (7)$$

by regarding random unpaired samples as negative. This slightly improves the performance.

3.2 Connection with CycleGAN

As a representative fully unpaired method, CycleGAN [Zhu *et al.*, 2017] would be the strong baseline at this point, which has been popularly used for unpaired distribution matching. Since it is designed for image-to-image translation, we describe the modifications to it to fit our task, so as to understand relative performance of our method over the CycleGAN baseline. When applying the cycle-consistency on matching between images and captions, since the input modalities are totally different, we modify it to a translation problem over feature spaces as follows:

$$\min_{\theta,\{T\}} \max_{D_{\{x,y\}}} L_{cycle}(\{T\})$$
$$+ \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})} [\log(D_x(F(\mathbf{x}))) + \log(1 - D_x(T_{v\to c}(F(\mathbf{x}))))]$$
$$+ \mathop{\mathbb{E}}_{y\sim p(y)} [\log(D_y(G(y))) + \log(1 - D_y(T_{c\to v}(G(y))))],$$
$$(8)$$

where $\{T\}$ and $D_{\{x,y\}}$ denotes the feature translators and the discriminators for image domain and caption domain, and

$$L_{cycle}(\{T\}) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\|T_{c\to v}(T_{v\to c}(F(\mathbf{x}))) - F(\mathbf{x})\|_2]$$
$$+ \mathbb{E}_{y\sim p(y)}[\|T_{v\to c}(T_{c\to v}(G(y))) - G(y)\|_2].$$
$$(9)$$

The discriminator D is learned to distinguish whether the latent features is from the image or the caption distribution. This is different with our method, in that we distinguish the correct *semantic match of pairs*. We experimented with Cycle-GAN purely on unpaired datasets, but we were not able to train it; hence, we add the supervised loss (Eq. (1)) with the paired data, which is a fair setting with ours.

4 Experiments

In this section, we describe the experimental setups, competing methods and provide performance evaluations of unpaired captioning with both quantitative and qualitative results.

4.1 Implementation Details

We implement our neural networks by PyTorch library [Paszke *et al.*, 2017]. We use ResNet101 [He *et al.*, 2016] model as a backbone CNN encoder for image input, and initialize with weights pretrained on ImageNet [Russakovsky *et al.*, 2015]. Each region is represented as 2,048-dimensional output from the last residual block of this network. For NIC model [Vinyals *et al.*, 2015] and Att2in model [Rennie *et al.*, 2017] in Table 3, we use visual features with size $2048 \times 7 \times 7$. For Up-Down [Anderson *et al.*, 2018] and the model in Table 3, we use Faster R-CNN [Ren *et al.*, 2015] pretrained on Visual Genome [Krishna *et al.*, 2017] object dataset, thereby we extract visual features for 10 to 100 regions.

We set the channel size of the hidden layers of LSTMs to be 1024, 512 for the attention layer, and 1024 for the word embeddings. For inference stage, we empirically choose the beam size to be 3 when generating a description, which shows the best performance.

We use a minibatch size of 100, the Adam [Ba and Kingma, 2015] optimizer for training (learning rate $lr = 5e^{-4}$, $b_1 = 0.9$, $b_2 = 0.999$). For hyperparameters, we empirically choose λ_x and λ_y to be equal to 0.1. The total loss function for training our model is as follows:

$$\mathcal{L} = \mathcal{L}_{cap} + 0.1\mathcal{L}_{GAN} + 0.1\mathcal{L}_{triplet}, \quad (10)$$

where $\mathcal{L}_{cap}$ denotes the captioning loss defined in Eq. (6), $\mathcal{L}_{GAN}$ the loss for adversarial training defined in Eq. (3), and $\mathcal{L}_{triplet}$ the triplet loss defined in Eq. (7).

	(A)	(B)	(C)	BLEU1	BLEU4	METEOR	CIDEr
Fully paired (100%)				72.5	29.4	25.0	95.7
Paired only (1%)				58.1	13.4	15.9	36.0
CycleGAN (Zhu *et al.*)				58.7	14.1	16.2	37.7
Ours ver1	✓			60.1	15.7	16.5	45.3
Ours ver2	✓	✓		61.3	17.1	19.1	51.7
Ours (final)	✓	✓	✓	**63.0**	**18.7**	**20.7**	**55.2**
[Gu *et al.*, 2018]				46.2	5.4	13.2	17.7
[Feng *et al.*, 2019]				58.9	18.6	17.9	54.9

Table 1: Captioning performance comparison on MS COCO testset. The "Paired only" baseline is trained only with 1% of paired data from our *scarcely-paired COCO* dataset, while CycleGAN and Ours {ver1, ver2, final}, are trained with our *scarcely-paired COCO* dataset (1% of paired data and unpaired image and caption datasets). We indicate the ablation study by: (A) the usage of the proposed GAN that distinguishes real or fake image-caption pairs, (B) pseudo-labeling, and (C) noise handling by sample re-weighting. We also compare with [Gu *et al.*, 2018] and [Feng *et al.*, 2019], which are trained with unpaired datasets.

4.2 Experimental Setups

We utilize MS COCO caption [Lin *et al.*, 2014] as our target dataset, which contains 123k images with 5 caption labels per image. To validate our model, we follow *Karpathy* splits [Karpathy and Fei-Fei, 2015], which have been broadly used in image captioning literature. The Karpathy splits contain 113k training, 5k validation, and 5k test images in total. In our experiments, in order to simulate the scenario that both paired and unpaired data exist, we use two different data setups: 1) the proposed scarcely-paired COCO, and 2) partially labeled COCO [Liu *et al.*, 2018] setup.

For the *scarcely-paired* COCO setup, we remove the pairing information of the MS COCO dataset, while leaving a small fraction of pairs unaltered. We randomly select only one percent of the total data, *i.e.*, 1,133 training images for the paired data, and remove the pairing information of the rest to obtain unpaired data. We call the small fraction of samples given as pairs as *paired* data ($\mathcal{D}_p$) and call the samples without pairing information as *unpaired* data ($\mathcal{D}_u$). We study the effects of other ratios of paired data used for training (in Figs. 4 and 3). This dataset allows evaluating the proposed framework to measure whether such small paired data can be leveraged to learn plausible pseudo-label assignment, and what performance can be achieved compared to the fully supervised case.

For *partially labeled* COCO setup, we follow [Liu *et al.*, 2018] and use the whole MS COCO data (paired) and add the *"Unlabeled-COCO"*
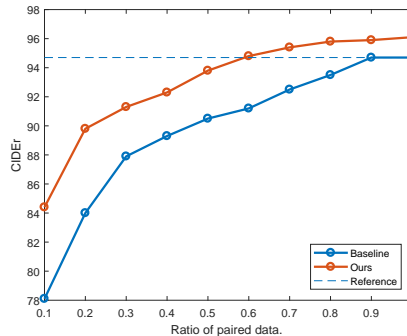

Figure 3: Performance w.r.t. amount of paired data for training. Baseline denotes our *Paired Only* baseline, Ours is our final model, and Reference is *Paired Only* trained with full paired data.

split of the officially MS-COCO [Lin *et al.*, 2014] for unpaired images, which involves 123k images without any caption label. To compute the cross entropy loss, we use the pseudo-label assignment for the Unlabeled-COCO images.

For evaluation, we use the following metrics conventionally used in image captioning: BLEU [Papineni *et al.*, 2002], ROUGUE-L [Lin, 2004], SPICE [Anderson *et al.*, 2016], METEOR [Denkowski and Lavie, 2014], and CIDEr [Vedantam *et al.*, 2015]. All the evaluation is done on MS COCO testset.

In order to avoid high time complexity of the pseudo-labeling process, we do not search the pseudo-labels from the whole unpaired data. Pseudo-label retrieval is done on a subset of one-hundredth of the unpaired data (yielding 1000 samples). Thus, the complexity of each minibatch becomes $O(B \times N \times 0.01)$, where B denotes the minibatch size, N denotes the size of the unpaired dataset. Since we also apply label-noise handling, a plausible pseudo-label assignment is sufficient in helping the learning process. We can improve performance by using a larger subset or by using a fast approximate nearest neighbor search.

4.3 Evaluation on Scarcely-paired COCO

We follow the same setting with [Vinyals *et al.*, 2015], if not mentioned. Table 1 shows the results on the scarcely-paired COCO dataset. We compare with several baselines: *Paired Only*; we train our model only on the small fraction (1%) of the paired data, *CycleGAN*; we train our model with the cycle-consistency loss [Zhu *et al.*, 2017] (as in Eq. (8)). Additionally, we train variants of our model denoted as *Ours* (ver1, ver2, and final). The model denoted as *Our ver1*, is trained

Figure 4: Sampled qualitative results of our model. We compare with the baseline models trained only on $N\%$ of paired samples out of the full MS COCO. Despite the use of only 1% paired data, our model generates plausible captions similar to that of the baseline models trained with more data (10% and above).

via our GAN model (Eq. (3)) that distinguishes real or fake image-caption pairs. For *Our ver2*, we add training with pseudo-labels for unpaired data with Eq. (6) while setting the confidence scores $\alpha^x = \alpha^y = 1$ for all training samples. For our final model denoted as *Our (final)*, we apply the noise handling technique by the confidence scores α^x and α^y computed by Eq. (6) and re-weighting the loss for each sample. We present the performance of the fully supervised (*Fully paired*) model using 100% of the COCO training data for reference.

As shown in Table 1, in a scarce data regime, utilizing the unpaired data improves the captioning performance in terms of all metrics by noticeable margins. Also, our models show favorable performance compared to the CycleGAN model in all metrics. Our final model with pseudo-labels and noise handling achieves the best performance in all metrics across all the competitor, hereafter referred to this model as our model.

We also compare the recent unpaired image captioning methods [Gu *et al.*, 2018; Feng *et al.*, 2019] in Table 1. Both of the methods are evaluated on MS COCO testset. In the case of Gu *et al.*, AIC-ICC image-to-Chinese dataset [Wu *et al.*, 2017] is used as unpaired images $\mathcal{D}_u^x$ and captions from MS COCO are used as unpaired captions $\mathcal{D}_u^y$. Note that this is not a fair comparison to

our method but they have more advantages, in that Gu *et al.* use large amounts of additional labeled data (10M Chinese-English parallel sentences of AIC-MT dataset [Wu *et al.*, 2017]) and Feng *et al.* use 36M samples of additional OpenImages dataset, whereas our model only uses a small amount of paired samples (1k) and 122k unpaired data. Despite far lower reliance on paired data, our model shows favorable performance against recent unpaired captioning works.

We study our final model against our *Paired Only* baseline according to varying amounts of paired training data in Fig. 3, so that we can see how much information can be gained from the unpaired data. From 100% to 10%, as the amount of paired samples decreases, the fluency and the accuracy of the descriptions get worse. In particular, we observed that most of the captions generated from the *Paired Only* baseline trained with 10% of paired data (11,329 pairs) show erroneous grammatical structure. In contrast, by leveraging unpaired data, our method can generate more fluent and accurate captions, compared to *Paired Only* trained on the same amount of paired data. It is worthwhile to note that our model trained with 60% of paired data (67,972 pairs) achieves similar performance to the *Paired Only* baseline trained with fully paired data (113,287 pairs), which sig-

	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	SPICE	METEOR	CIDEr
Self-Retrieval [Liu *et al.*, 2018] (w/o unlabeled)	79.8	62.3	47.1	34.9	56.6	20.5	27.5	114.6
Self-Retrieval [Liu *et al.*, 2018] (with unlabeled)	80.1	63.1	48.0	35.8	57.0	21.0	27.4	117.1
Baseline (w/o unlabeled)	77.7	61.6	46.9	36.2	56.8	20.0	26.7	114.2
Ours (w/o unlabeled)	80.8	65.3	49.9	37.6	58.7	22.7	28.4	122.6
Ours (with unlabeled)	**81.2**	**66.0**	**50.9**	**39.1**	**59.4**	**23.8**	**29.4**	**125.5**

Table 2: Comparison with the semi-supervised image captioning method, "Self-Retrieval" [Liu *et al.*, 2018]. Our method shows improved performance even without Unlabeled-COCO data (denoted as *w/o unlabeled*) as well as with Unlabeled-COCO (*with unlabeled*), although our model is not originally proposed for such scenario.

a large bear standing on a grass.　　a bus is parked on a street.　　a motorcycle parked on a street.

Figure 5: Examples of the pseudo-labels (captions) assigned to the unpaired images. Our model is able to sufficiently assign image-caption pairs through the proposed adversarial training.

Two bears **growling** in the water.　　A broccoli **grows** on a green plant.　　Cows being **herded** by a fence in a field.

Figure 6: Example captions containing words that do not exist in the paired dataset $\mathcal{D}_p$. The novel words that are not in $\mathcal{D}_p$ but in $\mathcal{D}_u^y$ are highlighted in bold.

nifies that our method is able to save *near half* of the human labeling effort of constructing a dataset.

We also show qualitative samples of our results in Fig. 4. *Paired Only* baseline trained with 1% paired data produces erroneous captions, and the baseline with 10% paired data starts to produce plausible captions. It is based on *ten times* more paired samples, compared to our model that uses only 1% of them. We highlight that, in the two examples on the top row of Fig. 4, our model generates more accurate captions than the *Paired Only* baseline trained on the 100% paired data ("baseball" to "Frisbee" on the top-left, and "man" to "woman" on the top-right). This suggests that unpaired data with our method effectively boosts the performance especially when paired data is scarce.

In order to demonstrate the meaningfulness of our pseudo-label assignment, we show the pseudo-labels assigned to the unpaired samples. We show the pseudo-labels (captions) assigned to unlabeled images from Unlabeled-COCO images in Fig. 5. As shown in the figure, despite not knowing the real pairs of these images, the pseudo labels are sufficiently assigned by the model. Note that even though there is no ground truth caption for unlabeled images in the searching pool, the model is able to find the most likely (semantically correlated) image-caption pair for the given images.

Fig. 6 highlights interesting result samples of our caption generation, where the results contain words that do not exist in the paired data of the scarcely-paired COCO dataset. The semantic meaning of the words such as "growling," "grow-

ing," and "herded," which involve in the unpaired caption data but not in the paired data, may have been learned properly via pseudo-label assignment during training. These examples suggest that our method is capable to infer the semantic meaning of unpaired words to some extent, which would have been unable to be learned with only little paired data. This would be an evidence that our framework is capable to align abstract semantic spaces between two modalities, *i.e.*, visual and caption.

4.4 Evaluation on Partially Labeled COCO

For a more realistic setup, we compare with the recent semi-supervised image captioning method, called *Self-Retrieval* [Liu *et al.*, 2018], on their problem regime, *i.e.*, "partially labeled" COCO setup, where the full amount of the paired MS COCO (113k) plus 123k uncaptioned images of the Unlabeled-COCO set are used (no additional unpaired caption is used). While their regime is not our direct scope, but we show the effectiveness of our method in the regime. Then, we extend our framework by replacing our backbone architecture with recent advanced image caption architectures. In this setup, a separate *unpaired* caption data $\mathcal{D}_u^y$ does not exist, thus we use captions from paired COCO data $\mathcal{D}_p$ as pseudo-labels.

Table 2 shows the comparison with Self-Retrieval. For a fair comparison with them, we replace the cross entropy loss from our loss with the policy gradient method [Rennie *et al.*, 2017] to directly optimize our model with CIDEr score as in [Liu *et al.*, 2018]. As our baseline

	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	SPICE	METEOR	CIDEr
NIC [Vinyals *et al.*, 2015]	72.5	55.1	40.4	29.4	52.7	18.2	25.0	95.7
NIC **+Ours**	**74.2**	**56.9**	**42.0**	**31.0**	**54.1**	**19.4**	**28.8**	**97.8**
Att2in [Rennie *et al.*, 2017]	74.5	58.1	44.0	33.1	54.6	19.4	26.2	104.3
Att2in **+Ours**	**75.5**	**58.8**	**45.0**	**34.3**	**55.4**	**20.0**	**26.8**	**105.3**
Up-Down [Anderson *et al.*, 2018]	76.7	60.7	47.1	36.4	56.6	20.6	27.6	113.6
Up-Down **+Ours**	**77.3**	**64.1**	**47.6**	**37.0**	**57.1**	**21.0**	**27.8**	**114.6**

Table 3: Evaluation of our method with different backbone architectures. All models are reproduced and trained with the cross entropy loss. By adding Unlabeled-COCO images, our training method was applied in a semi-supervised way, which shows consistent improvement in all the metrics.

model (denoted as *Baseline*), we train a model only with policy gradient method without the proposed GAN model. When only using the 100% paired MS COCO dataset (denoted as *w/o unlabeled*), our model already shows improved performance over Self-Retrieval. Moreover, when adding Unlabeled-COCO images (denoted as *with unlabeled*), our model performs favorably against Self-Retrieval in all the metrics. The results suggest that our method is also advantageous in the semi-supervised setup.

To further validate our method in the semi-supervised setup, we compare different backbone architectures [Anderson *et al.*, 2018; Rennie *et al.*, 2017; Vinyals *et al.*, 2015] in our framework, where their methods are developed for fully-supervised methods. We use the same data setup with the above, but we replace CNN (F) and LSTM (H) from our framework with the image encoder and the caption decoder from their image captioning models. Then, these models can be trained in our framework as it is by alternating between the discriminator update and pseudo-labeling. Table 3 shows the comparison. Training with the additional Unlabeled-COCO data via *our* training scheme consistently improves all baselines in all metrics.

4.5 Captioning with Web-Crawled Data

To simulate a scenario involving crawled data from the web, we use the setup suggested by Feng *et al.* [2019]. They collect a sentence corpus by crawling the image descriptions from Shutterstock[1] as unpaired caption data $\mathcal{D}_u^y$, whereby 2.2M sentences are collected. For unpaired image data $\mathcal{D}_u^x$, they use only the images from the MS COCO data, while the captions are not used for training. For training our method, we leverage from 0.5% to 1% of paired MS COCO data as our paired data $\mathcal{D}_p$. The results are shown in Table 4 with the performance obtained by Feng *et*

---

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

	BLEU4	ROUGE-L	SPICE	METEOR	CIDEr
Paired only (0.5% paired)	4.4	33.7	3.6	10.8	8.6
Ours (0.5% paired)	5.4	34.6	4.2	12.0	10.5
Paired only (0.7% paired)	3.5	36.1	3.7	11.4	8.9
Ours (0.7% paired)	8.5	39.0	5.2	13.6	20.2
Paired only (0.8% paired)	8.8	39.1	5.9	13.2	21.9
Ours (0.8% paired)	12.2	41.6	7.6	15.1	29.0
Paired only (1% paired)	13.4	41.9	8.3	15.9	36.0
Ours (1% paired)	15.2	43.3	9.4	16.9	39.7
[Feng *et al.*, 2019]	4.8	27.2	7.0	11.4	23.3

Table 4: Performance comparison with web-crawled data (Shutterstock). On top of unpaired image and caption data, our method is trained with 0.5 – 1% of paired data, while Feng *et al.* use 36M additional images of the OpenImage dataset.

al. Note again that Feng *et al.* exploits external large-scale data, *i.e.*, 36M images in the OpenImages dataset. With 0.5% of paired only data (566 pairs), our baseline shows lower scores in terms of BLEU4 and METEOR than Feng *et al.*, while our proposed model shows comparable or favorable performance in BLEU4, ROUGE-L and METEOR. Our method starts to have higher scores in all metrics from 0.8% of paired data (906 pairs), even without additional 36M images.

5 Conclusion

We introduce a method to train an image captioning model with a large scale unpaired image and caption data, given a small amount of paired data. Our framework achieves favorable performance compared to various methods and setups. Unpaired captions and images are the data that can be easily collected from the web. It can facilitate application specific captioning models, where labeled data is scarce.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

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 vqa. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018.

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

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. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Contextually customized video summaries via natural language. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

LI Chongxuan, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *The workshop on statistical machine translation*, 2014.

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

Zhe Gan, Liqun Chen, Weiyao Wang, Yuchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. Triangle generative adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

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

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

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, Youngjin Yoon, and In So Kweon. Disjoint multi-task learning between heterogeneous human-centric tasks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *European Conference on Computer Vision (ECCV)*, 2018.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2:3, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*, 2018.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. Phrase-based & neural unsupervised machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*. Springer, 2014.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2002.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *European Conference on Computer Vision (ECCV)*, 2018.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based statistical machine translation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2007.

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

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

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

Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *European Conference on Computer Vision (ECCV)*, 2018.

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Hua Wu and Haifeng Wang. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181, 2007.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.

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

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. Joint training for neural machine translation models with monolingual data. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

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