Semi-Supervised Learning for Video Captioning

Ke Lin[1,2] **Zhuoxin Gan**[1] **Liwei Wang**[2]
[1]Samsung Research China - Beijing, China. [2]Peking University, China
[1]{ke17.lin,zhuoxin1.gan}@samsung.com [2]wanglw@pku.edu.cn

Abstract

Deep neural networks have made great success on video captioning in supervised learning setting. However, annotating videos with descriptions is very expensive and time-consuming. If the video captioning algorithm can benefit from a large number of unlabeled videos, the cost of annotation can be reduced. In the proposed study, we make the first attempt to train the video captioning model on labeled data and unlabeled data jointly, in a semi-supervised learning manner. For labeled data, we train them with the traditional cross-entropy loss. For unlabeled data, we leverage a self-critical policy gradient method with the difference between the scores obtained by Monte-Carlo sampling and greedy decoding as the reward function, while the scores are the K-L divergence between output distributions of original video data and augmented video data. The final loss is the weighted sum of losses obtained by labeled data and unlabeled data. Experiments conducted on VATEX, MSR-VTT and MSVD dataset demonstrate that the introduction of unlabeled data can improve the performance of the video captioning model. The proposed semi-supervised learning algorithm also outperforms several state-of-the-art semi-supervised learning approaches.

1 Introduction

Video captioning refers to the task that generating a description of a given video automatically and it combines computer vision and Natural Language Processing (NLP) in a unified framework. It can be widely used in video retrieval, video recommendation, disabled supporting and scene understanding (Yao et al., 2015), (Venugopalan et al., 2015). With the rapid development of deep learning, deep neural networks have dominated the video captioning task. Venugopalan et al. (Venugopalan et al., 2015) extend encoder-decoder framework to video captioning which employs a CNN as the encoder and an RNN as the decoder and the following video captioning algorithms almost use this architecture.

Although recent video captioning algorithms have made great success, they are heavily dependent on supervised training data consisting of video-caption pairs. It is expensive to take long hours of laboring to collect such labeled data, thus there is a strong interest to develop the algorithm which does not need a lot of annotated examples. Some studies embed the visual feature and text imformation into a mutual space and design unsupervised learning algorithm to reduce the requirement for annotated data (Gu et al., 2018), (Gu et al., 2019), (Laina et al., 2019), (Feng et al., 2019). However, the performances of such algorithms are poor because they do not use pairs of labeled examples at all. Semi-supervised learning, leveraging a small number of labeled examples and a large number of unlabeled examples at the same time, provides another solution to solve the problem of strong dependency on labeled examples. Chen et al. (Chen et al., 2016) proposed a Semi-Supervised Learning (SSL) image captioning strategy which using unsupervised out-of-domain textual data to boost the captioning performance. Kim et al. (Kim et al., 2019) proposed another semi-supervised image captioning algorithm which jointly using the labeled and unlabeled data and assigning pseudo-labels to unlabeled data via Generative Adversarial Networks to learn the joint distribution of image and text.

Recently, some semi-supervised learning works which use the consistency of the output probability distribution between original data and augmented data have achieved excellent performances with the help of some latest data augmentation methods on several classification problems of computer vision and NLP (Berthelot et al., 2019), (Xie et al., 2019a). Although these data augmentation based methods have great potential, it is still challenging when applied to video captioning task, because the input

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1096–1106

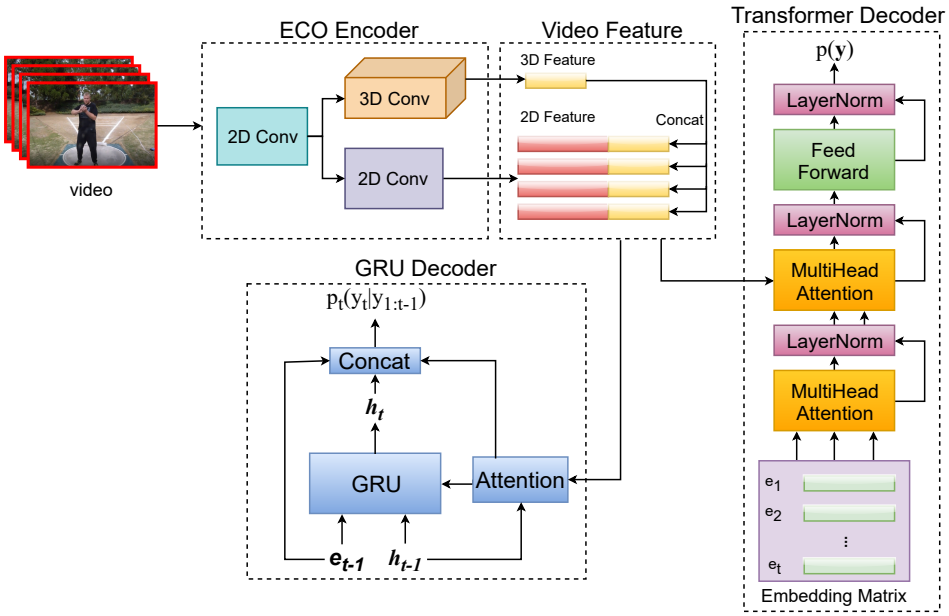


Figure 1: The architecture of the video captioning model, which includes an ECO encoder and a GRU decoder (or a Transformer Decoder).

and output complexity of video captioning is much higher than that of image or text classification. In this paper, we design our algorithm based on output consistency of data augmentation, and apply a self-critical training strategy (Rennie et al., 2017) to exploit large amounts of unsupervised video data without requiring the corresponding caption annotations. First, we get a pseudo-label by Monte-Carlo sampling. Then a reward score is obtained, while the score is the K-L divergence between the output distribution of augmented examples and real examples. Based on the observation of Figure 3, K-L divergence is positive related with the quality of the sentence, thus K-L divergence can be used as the reward score in policy gradient. We use greedy decoding to get another pseudo-label and get another reward score as the baseline. Then we use the difference between these two reward scores as the final reward. Finally, we combine the reward with the log probability to get the policy gradient loss. In the meantime, we compute the cross-entropy loss for labeled data and train the model with these two losses jointly.

In summary, the main contributions of the proposed algorithm are three-fold:

- To the best of our knowledge, this is the first attempt to use semi-supervised learning on video captioning task.
- We apply a self-critical policy gradient learning algorithm and consistency regularization to leverage the unlabeled data to improve the model performance.
- Our proposed approach is robust for different captioning tasks, different datasets and different models and outperforms several state-of-the-art semi-supervised learning algorithms.

2 Related Work

2.1 Video Captioning

The development of video captioning algorithms comes from researchers' unremitting efforts to find better video features, stronger model architectures and better optimization strategies. For feature extraction, 3D-CNN spatial-temporal feature (Yao et al., 2015), (Aafaq et al., 2019), transferred semantic attributes (Pan et al., 2017), external semantic information (Venugopalan et al., 2016), audio features (Wang et al., 2018c) and Part-of-Speech (POS) information (Hou et al., 2019), (Wang et al., 2019a) are used to enhance the representation ability of features. For model architecture, attention mechanism (Yao et al., 2015), (Wang et al., 2018c), (Song et al., 2017) and strong decoders (Pasunuru and Bansal, 2017), (Wang et al., 2018b), (Zhou et al., 2018) are proposed to enhance the decoding ability. For optimization strategy, Rennie et al. (Rennie et al., 2017) propose a self-critical sequence training strategy for image captioning and

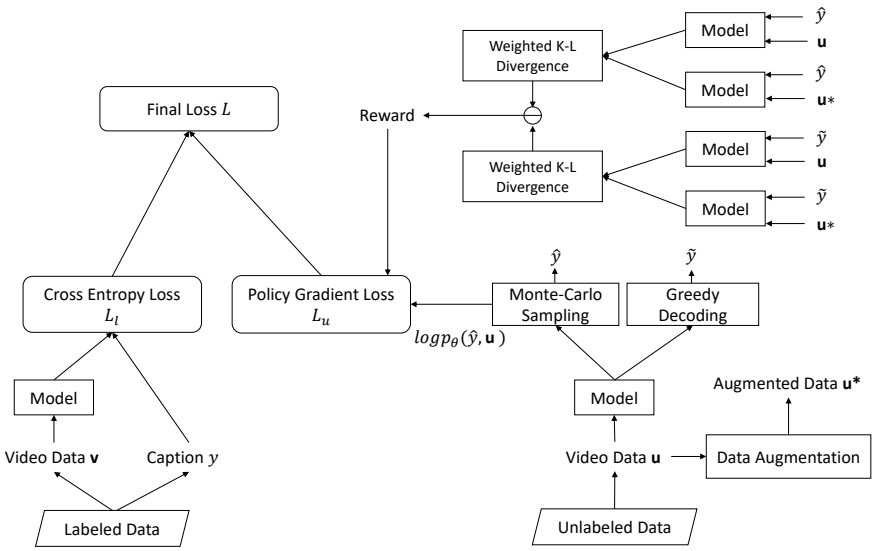


Figure 2: The proposed Self-Critical Semi-Supervised (SC-SSL) Learning algorithm.

Want et al. (Wang et al., 2017) extend this method to video captioning by introducing a hierarchical Reinforcement Learning (RL) algorithm.

2.2 Semi-supervised learning

Semi-supervised learning has been proposed for a long time as a solution to reduce the dependency on supervised data. SSL can be roughly divided into the following four types: transductive models (Joachims, 1999), (Joachims, 2003), graph-based approaches (Zhu et al., 2003), generative models (Pu et al., 2016), (Salimans et al., 2016) and consistency regularization (Laine and Aila, 2017), (Tarvainen and Valpola, 2017), (Miyato et al., 2017), (Xie et al., 2019b) based methods. Because our method belongs to consistency regularization, we pay more attention to discuss this kind of method. Consistency regularization applies data augmentation to unlabeled data and this operation is based on the insight that for an unlabeled example even after it has been augmented, the output distribution of a classifier should be similar with the original data."Π -Model" (Laine and Aila, 2017) computes the Mean-Squared Error (MSE) of the class distribution between two different augmented examples from one unlabeled data. "Mean Teacher" (Tarvainen and Valpola, 2017) replaces one of the terms in "Π -Model" with the output of the model using an exponential moving average of model parameters. Virtual Adversarial Training (Miyato et al., 2017) (VAT) proposes a novel virtual adversarial loss to measure the local smoothness of the conditional label distribution given input which can address the domain-specific data augmentation problem. Some recent works (Berthelot et al., 2019), (Xie et al., 2019a) utilize latest data augmentation methods such as MixUp (Zhang et al., 2017) or AutoAugment (Cubuk et al., 2018) to improve the performance of SSL.

3 Methods

3.1 Video Captioning Model

The main purpose of this study is to verify the effectiveness of SSL in video captioning instead of proposing a strong video captioning model, so we only use a simple video captioning model. Besides, we apply two candidate decoders (GRU and Transformer) to verify the robustness of the proposed algorithm.

Video Encoder. For video captioning, an input video $\mathbf{v}$ is given and we are required to generate a caption with a sequence of words $\mathbf{y} = [y_1, \dots y_t, \dots, y_T], y_t \in \mathcal{Y}$ to describe the video, where T is the maximum length of a sentence and $\mathcal{Y}$ is the vocabulary set. To encode the visual feature of the given video, we use an Efficient Convolutional Network (ECO) (Zolfaghari et al., 2018) pre-trained on Kinetics-400 dataset (Kay et al., 2017) as the encoder.

GRU Captioning Decoder. Our GRU captioning decoder model is similar to (Yao et al., 2015) which utilizes a variant LSTM as the base model, but in our implementation, a GRU in which visual feature is added into the inputs with an attention module is used to replace the original LSTM.

Algorithm 1 Self-Critical Semi-Supervised Learning (SC-SSL) algorithm for video captioning

Require: Batch of labeled data $\mathbf{v} = [\boldsymbol{v}_b], b \in (1, ..., B_l)$ and their caption labels $\mathbf{y} = [y^{(b)}], b \in (1, ..., B_l)$, batch of unlabeled data $\mathbf{u} = [\boldsymbol{u}_b], b \in (1, ..., B_u)$, number of augmented data for one unlabeled data K, maximum length of a sentence T, temporal weights $\mathbf{W} = [W_t], t \in (1, ...T)$, weight parameter for unlabeled loss λ.

$L_u = 0$ // Loss of unlabeled data
$L_l = 0$ // Loss of labeled data
for $b = 1$ **to** B_u **do**
  $\hat{\mathbf{y}} = [\hat{y}_1 \dots \hat{y}_T], where\ \hat{y}_t = \underset{\hat{y}_t}{Sample}\ p_\theta(\hat{y}_{1:t-1}, \boldsymbol{u}_b)$
  // Monte-Carlo Sampling
  $\tilde{\mathbf{y}} = [\tilde{y}_1 \dots \tilde{y}_T], where\ \tilde{y}_t = \arg\max_{\tilde{y}_t} p_\theta(\tilde{y}_{1:t-1}, \boldsymbol{u}_b)$
  // Greedy Decoding
  for $k = 1$ **to** K **do**
    $\boldsymbol{u}_b^* = DataAugmentation(\boldsymbol{u}_b)$ // AutoAugment or RandomDrop
    $\hat{d}_t = D_{KL}(p_\theta(\hat{y}_t|\hat{y}_{1:t-1}, \boldsymbol{u}_b)||p_\theta(\hat{y}_t|\hat{y}_{1:t-1}, \boldsymbol{u}_b^*))$
    $\tilde{d}_t = D_{KL}(p_\theta(\tilde{y}_t|\tilde{y}_{1:t-1}, \boldsymbol{u}_b)||p_\theta(\tilde{y}_t|\tilde{y}_{1:t-1}, \boldsymbol{u}_b^*))$
    $r = \sum_{t=1}^{T}(\hat{d}_t - \tilde{d}_t) \times W_t$ // Reward
    $\nabla_\theta l_u(\theta) = -\sum_{t=1}^{T} r \nabla_\theta log p_\theta(\hat{y}_t|\hat{y}_{1:t-1}, \boldsymbol{u}_b)$ // Policy Gradient
    $L_u = L_u + l_u$
  end for
  $L_u = L_u / K$
end for
for $b = 1$ **to** B_l **do**
  $L_l = L_l + -\sum_{t=1}^{T} log(p_\theta(y_t^{(b)}|y_{1:t-1}^{(b)}, \boldsymbol{v}_b))$
end for
$L = L_l + \lambda * L_u$ // Final loss
return L

Transformer Captioning Decoder. In order to demonstrate the validity of our SSL method, besides the recurrent captioning decoder, we also take experiments on Transformer (Vaswani et al., 2017) captioning decoder. Since video captioning is a video-to-text task rather than a text-to-text task, our decoder model only consists of transformer decoder. The whole architecture of the proposed video captioning model is illustrated as Figure. 1.

3.2 Self-Critical Semi-Supervised Learning

3.2.1 Algorithm

We have some labeled video data $\mathbf{v}$ with caption annotations $\mathbf{y}$ and unlabeled video data $\mathbf{u}$. To train the labeled data, we use the traditional cross-entropy loss:

$$L_l(\theta) = -\sum_{t=1}^{T} log(p_\theta(y_t|y_{1:t-1}, \mathbf{v})) \quad (1)$$

For unlabeled data $\mathbf{u}$, we generate a sentence as pseudo-label $\hat{\mathbf{y}} = [\hat{y}_1 \dots \hat{y}_T]$ by Monte-Carlo sampling using the current model parameters. We apply data augmentation on the unlabeled example by K times to get K augmented video examples. We denote one augmented data as $\mathbf{u}^*$. Here, we use two data augmentation methods: AutoAugment (Cubuk et al., 2018) and randomly dropping some frames of the video with probability σ. Some other video data augmentation methods can also be used in our algorithm. We use AutoAugment as the default data augmentation method. Then we compute the K-L divergence of the output class distribution between $\mathbf{u}$ and $\mathbf{u}^*$ given $\hat{\mathbf{y}}$. Thus, we can obtain a consistency distance $\hat{d}$:

$$\hat{d} = \sum_{t=1}^{T} \hat{d}_t = \sum_{t=1}^{T} D_{KL}(p_\theta(\hat{y}_t|\hat{y}_{1:t-1}, \boldsymbol{u})||p_\theta(\hat{y}_t|\hat{y}_{1:t-1}, \boldsymbol{u}^*)) \quad (2)$$

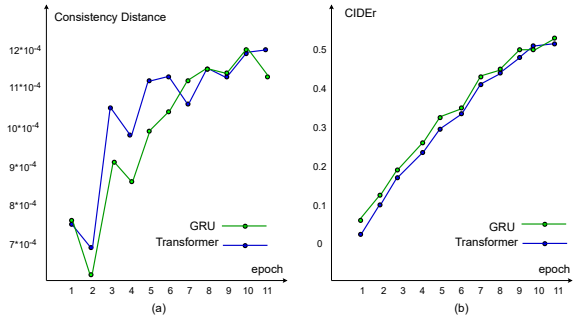


Figure 3: (a) Consistency distance. and (b) CIDEr. using the sentence generated by captioning model trained with different epochs.

Next, we will demonstrate that the consistency distance $\hat{d}$ is positive correlated with the quality of $\hat{\mathbf{y}}$. We perform a experiment on VATEX dataset using GRU and Transformer decoders. We generate several captions for all the validation samples using the model trained with different epochs. We denote these captions as $\{\mathbf{Y}_e\}, e \in (1, \dots, E)$, E is the maximum epoch. Then we compute the CIDEr score which is often regarded as the best metric to measure the quality of sentence between the generated captions and ground truth. Figure 3. (b) shows CIDEr increases with the epoch increasing. We also use the model of the last epoch to compute the average consistency distance over all validation data given different $\mathbf{Y}_e$. The trend of the change of consistency distance is consistent with that of CIDEr. The correlation coefficient between consistency distance and CIDEr are $0.92, 0.86$ for GRU and transformer.

Based on the result of Figure 3., if the quality of the generated sentence is high (i.e. with high CIDEr score), the consistency distance between $\mathbf{u}$

and $\mathbf{u}^*$ might be higher. Otherwise, the consistency distance will be lower. The consistency distance is positively correlated to evaluate the quality of sentence and can be regarded as the reward function of policy gradient algorithm (Williams, 1992), which is a specific type of Reinforcement Learning. The reward and policy gradient are:

$$\hat{r} = \hat{d} \tag{3}$$

$$\nabla_\theta L_u(\theta) = -\sum_{t=1}^{T} \hat{r} \nabla_\theta log p_\theta(\hat{y}_t | \hat{y}_{1:t-1}, \boldsymbol{u}) \tag{4}$$

Inspired by self-critical sequence training (SCST) (Rennie et al., 2017), we also add a baseline term on the reward function. The baseline term is the reward obtained using the pseudo-label $\tilde{\mathbf{y}} = [\tilde{y}_1 \dots \tilde{y}_T]$ which is generated by greedy decoding, where

$$\tilde{y}_t = \arg\max_{\tilde{y}_t} p_\theta(\tilde{y}_{1:t-1}, \boldsymbol{u}) \tag{5}$$

The consistency distance and the reward obtained by $\tilde{y}$ are:

$$\tilde{d}_t = D_{KL}(p_\theta(\tilde{y}_t | \tilde{y}_{1:t-1}, \boldsymbol{u}) || p_\theta(\tilde{y}_t | \tilde{y}_{1:t-1}, \boldsymbol{u}^*)) \tag{6}$$

$$\tilde{r} = \sum_{t=1}^{T} \tilde{d}_t \tag{7}$$

And the policy gradient is replaced to:

$$\nabla_\theta L_u(\theta) = -\sum_{t=1}^{T} (\hat{r} - \tilde{r}) \nabla_\theta log p_\theta(\hat{y}_t | \hat{y}_{1:t-1}, \boldsymbol{u}) \tag{8}$$

L_u is averaged over K augmented examples. We jointly train the labeled and unlabeled data using the weighted sum of the losses from labeled and unlabeled data:

$$L = L_l + \lambda * L_u \tag{9}$$

where λ is a hyper-parameter to control the weight of each component.

3.2.2 Training Techniques

For the pseudo label mentioned above, words occur later in a sentence have lower confidence due to the problem of error accumulation (Ranzato M A, 2015). To address this issue, we add a temporal weight on the reward function to decrease the weights of losses of later words. The temporal weight is $W_t = T/t$, and the equation (3) and equation (7) are replaced by:

$$\hat{r} = \sum_{t=1}^{T} \hat{d}_t \times W_t \tag{10}$$

$$\tilde{r} = \sum_{t=1}^{T} \tilde{d}_t \times W_t \tag{11}$$

To overcome the problem of overfitting of labeled data, following (Xie et al., 2019a), we add a training signal annealing on the calculation of the loss of labeled data. Equation (1) is changed to the following equation:

$$L_l = -\sum_{t=1}^{T} \log(p_\theta(y_t | y_{1:t-1}, \mathbf{v})) I(p_\theta(y_t | y_{1:t-1}, \mathbf{v}) < \eta_\tau) \tag{12}$$

where $I(\cdot)$ is the indicator function. We use linear-schedule annealing signal:

$$\eta_\tau = \frac{\tau}{M} \times (1 - \frac{1}{C}) + \frac{1}{C} \tag{13}$$

where C is the vocabulary size, M is the total training steps.

The proposed Self-Critical Semi-Supervised Learning (SC-SSL) algorithm is summarized as Algorithm 1 and illustrated as Figure. 2.

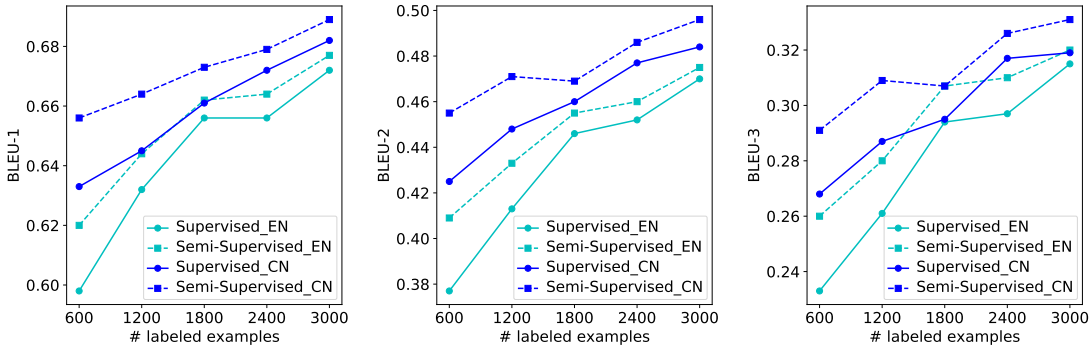
4 Results

4.1 Datasets and Implementation Details

We conduct experiments on three benchmark datasets Video And TEXt (VATEX) (Wang et al., 2019b), Microsoft Research video to text (MSR-VTT) and Microsoft Research Video Description Corpus (MSVD) (Chen and Dolan, 2011).

VATEX. VATEX contains over 41250 video clips in 10 seconds and each video clip depicts a single activity. Each video clip has 10 English descriptions and 10 Chinese descriptions. We use the official 25991 training examples as labeled and unlabeled training data and 3000 validation examples for testing. For labeled and unlabeled partition, we randomly select 600, 1200, 1800, 2400, 3000 as labeled data and use the rest training data as unlabeled data.

MSR-VTT. We use the initial version of MSR-VTT, referred as MSR-VTT-10K which has 10k video clips and each video clip has 20 descriptions annotated by 1327 workers from Amazon Mechanical Turk. MSR-VTT has 200k video-caption pairs and 29316 unique words. We take 7010 video clips



(a) BLEU-1 (b) BLEU-2 (c) BLEU-3

(d) BLEU-4 (e) ROUGE-L (f) METEOR

Figure 4: (a) BLEU-1, (b) BLEU-2, (c) BLEU-3, (d) BLEU-4, (e) ROUGE-L and (f) METEOR on VATEX English and Chinese captioning tasks with different number of labeled examples. "EN" is short for English and "CN" is short for Chinese.

as labeled training data and 2990 clips for testing. We use the training data of VATEX as unlabeled training data.

MSVD. MSVD dataset contains 1970 YouTube short video clips in 10 seconds to 25 seconds and each video clip depicts a single activity. Each video clip has about 40 English descriptions. We use the public splits which take 1200 video clips for training, 100 clips for validation and 670 clips for testing. We use the training data of VATEX and MSR-VTT as unlabeled training data.

We follow the standard caption pre-processing procedure including converting all words to lower cases, tokenizing on white space, clipping sentences over 24 words and filtering words which occur at least five times. We use open source Jieba [1] toolbox to segment the Chinese words. The final vocabulary sizes are 10260 for VATEX English task, 12776 for VATEX Chinese task, 8784 for MSR-VTT dataset and 5663 for MSVD dataset. We use standard automatic evaluation metrics including BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2014).

We uniformly sample 32 frames for each video clip. The embedding dimension 512. For GRU decoder, the model size and all hidden size are 512. For transformer decoder, the layer number is 6, the number of head is 8 and the model dimension is 512. We train the captioning model using an Adam optimizer. At first 10 epochs, we only train labeled data. The learning rate is 5×10^{-4}, batch size is 100 and dropout rate is 0.1. Then we train the labeled data and unlabeled data jointly with learning rate of 1×10^{-4}, labeled batch size of 100 and unlabeled batch size of 400. We set hyper-parameters by $K = 10$, $\lambda = 1 \times 10^3$, and $\sigma = 0.1$.

4.2 Evaluation and Comparison

Figure. 4. shows the results of BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L and METEOR on VATEX English and Chinese captioning tasks with different number of labeled examples using GRU decoder. It can be seen that the proposed semi-supervised learning algorithm outperforms supervised learning algorithm for all metrics. As

[1] https://github.com/fxsjy/jieba

Table 1: Comparison between supervised and semi-supervised learning using GRU or Transformer decoder on VATEX English Captioning task.

#Labeled Examples	Model Type	Training Type	BLEU-4	METEOR	ROUGE-L	CIDEr
600	GRU	Supervised	0.142	0.166	0.332	0.177
		Semi-Supervised	**0.163**	**0.175**	**0.346**	**0.181**
	Transformer	Supervised	0.141	0.162	0.325	0.157
		Semi-Supervised	**0.153**	**0.178**	**0.342**	**0.174**
1200	GRU	Supervised	0.163	0.177	0.344	0.224
		Semi-Supervised	**0.179**	**0.186**	**0.362**	**0.229**
	Transformer	Supervised	0.157	0.174	0.345	0.221
		Semi-Supervised	**0.169**	**0.186**	**0.360**	**0.234**

Table 2: Comparison with state-of-the-arts on MSR-VTT dataset. Unlabeled data comes from VATEX.

	BLEU-4	METEOR	ROUGE-L	CIDEr
MGSA (Chen and Jiang, 2019)	0.424	0.276	-	0.475
Hierarchical (Song et al., 2017)	0.383	0.263	-	-
M3 (Wang et al., 2018b)	0.381	0.266	-	-
GRU-EVE (Aafaq et al., 2019)	0.383	0.284	0.607	0.481
PickNet (Chen et al., 2018)	0.413	0.277	0.598	0.441
Reconstruction (Wang et al., 2018a)	0.391	0.266	0.593	0.427
MARN (Pei et al., 2019)	0.404	0.281	0.607	0.471
XGating (Wang et al., 2019a)	0.420	0.282	0.616	0.487
OA-BTG (Zhang and Peng, 2019)	0.414	0.282	-	0.469
JSRL+VCT (Hou et al., 2019)	0.423	0.297	0.628	0.491
Ours: Supervised	0.419	0.294	0.621	0.489
Ours: SC-SSL w VATEX	**0.427**	**0.300**	**0.632**	**0.498**

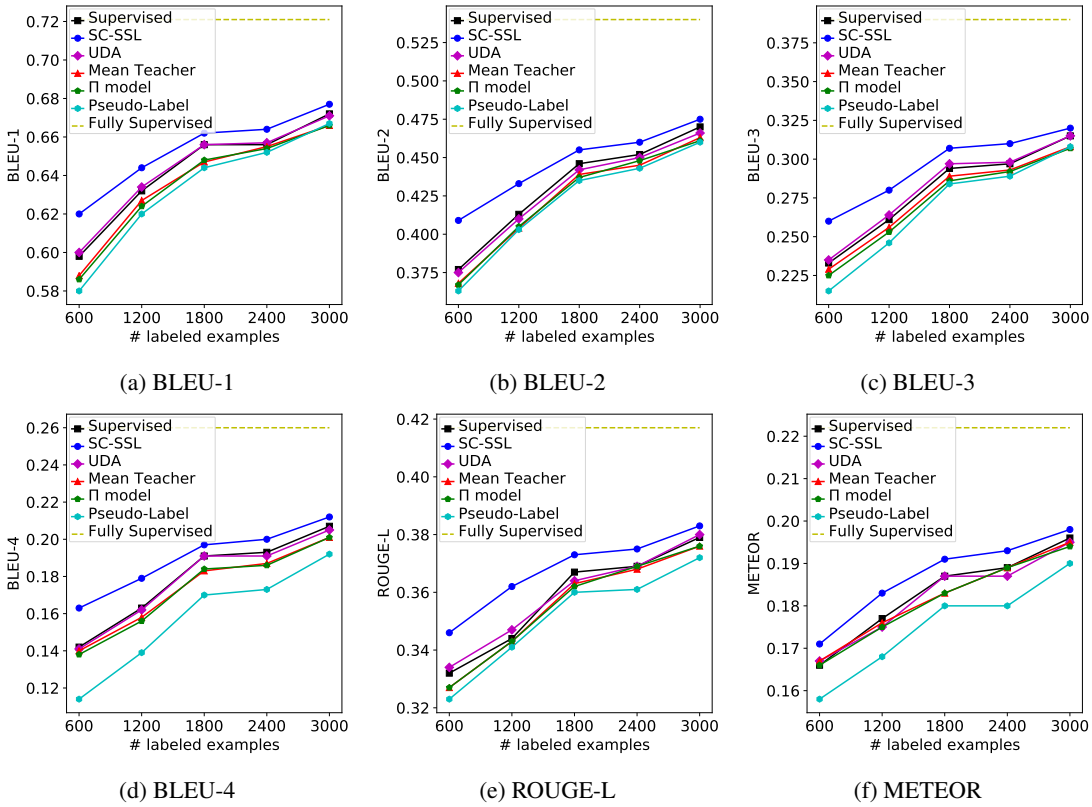
Table 3: Comparison with state-of-the-arts on MSVD dataset. Unlabeled data comes from VATEX and MSR-VTT.

	BLEU-4	METEOR	ROUGE-L	CIDEr
FCVC-CF (Fang et al., 2019)	0.531	0.348	0.718	0.798
MGSA (Chen and Jiang, 2019)	0.534	0.350	-	0.867
LSTM-TVAIV (Pan et al., 2017)	0.528	0.335	-	0.740
Hierarchical (Song et al., 2017)	0.530	0.336	-	0.738
M3 (Wang et al., 2018b)	0.520	0.322	-	-
GRU-EVE (Aafaq et al., 2019)	0.479	0.350	0.715	0.781
PickNet (Chen et al., 2018)	0.523	0.333	0.696	0.765
Reconstruction (Wang et al., 2018a)	0.523	0.341	0.698	0.803
ECO (Zolfaghari et al., 2018)	0.535	0.350	-	0.858
XGating (Wang et al., 2019a)	0.525	0.341	0.713	0.887
JSRL+VCT (Hou et al., 2019)	0.528	0.361	0.718	0.878
Ours: Supervised	0.556	0.347	0.711	0.857
Ours: SC-SSL w VATEX	0.567	0.353	0.715	0.870
Ours: SC-SSL w VATEX & MSR-VTT	**0.572**	**0.364**	**0.725**	**0.888**

the number of labeled example increasing, which means the ratio between labeled and unlabeled examples is getting larger, the gap between semi-supervised and supervised decreases. The gaps for most metrics are between 0.01 and 0.02. This result demonstrates that by introducing unlabeled data, the performance can be boosted. Combining the results of English and Chinese captioning tasks, we can see that the proposed SC-SSL algorithm is robust for different captioning tasks.

The proposed SC-SSL is effective for different models as well. From Table 1. we can see that the results of semi-supervised learning are higher than supervised learning on all metrics for both GRU and Transformer based decoder using 600 and 1200 labeled examples on VATEX English captioning task. The above results demonstrate that SC-SSL will not overfit to a certain model or a certain task, and it is a robust and general algorithm. Another interesting result in Table 2. is that some metrics of semi-supervised learning using 600 labeled examples are comparable with that of supervised learning using 1200 labeled examples (e.g. 0.346 vs. 0.344 of ROUGE-L using GRU decoder). This result shows that the proposed SC-SSL algorithm can reduce the requirement of annotating videos by half with the help of a large number of unlabeled data under certain circumstances.

Figure. 5. shows the comparison with other state-of-the-art semi-supervised learning algorithms of BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L and METEOR on VATEX English captioning task with different number of labeled examples using GRU decoder. Here, as baselines for comparison, we consider four other methods: Pseudo-Label (Lee, 2013), Π Model (Laine and Aila, 2017), Mean Teacher (Tarvainen and Valpola, 2017) and UDA (Xie et al., 2019a). Other semi-supervised learning algorithms such as VAT (Miyato et al., 2017) or MixMatch (Berthelot et al., 2019) require single label output and can only be used in classification tasks, so we do not compare with these methods. For Pseudo-Label, Π Model, Mean Teacher and UDA, we use the sentence generated by greedy decoding as the pseudo-label. The result shows that because video captioning task is much harder than classification task, Pseudo-Label, Π Model, Mean Teacher and UDA all fail to beat the supervised learning baseline. Among these methods, UDA



(a) BLEU-1 (b) BLEU-2 (c) BLEU-3

(d) BLEU-4 (e) ROUGE-L (f) METEOR

Figure 5: Comparison with other state-of-the-art semi-supervised learning algorithms of (a) BLEU-1, (b) BLEU-2, (c) BLUE-3, (d) BLEU-4, (e) ROUGE-L and (f) METEOR on VATEX English captioning task with different number of labeled examples.

achieves the best performance and this result is consistent with other prior arts (Xie et al., 2019a). SC-SSL outperforms other four algorithms especially UDA with a significant gap, because it fully considers the sequential property of captioning task and uses policy gradient to update the model parameter instead of using the K-L divergence as loss directly.

Table 2. shows the results on MSR-VTT dataset using GRU decoder. For fair comparison, we only show the results of prior arts trained with cross-entropy loss, because we train labeled data using only cross-entropy loss without RL optimization. It can be seen that SC-SSL outperforms supervised learning method which demonstrates that unlabeled data from another dataset can also help to boost the captioning performance even the distributions of labeled and unlabeled data are not consistent. Thanks to the unlabeled data, the proposed SC-SSL outperforms other state-of-the-art video captioning models for all metrics, even the decoder used in our method is quite simple.

Table 3. shows the results on MSVD dataset using GRU decoder. It has similar results with MSR-VTT that unlabeled data from VATEX can help to boost the captioning performance. Jointly using unlabeled data from VATEX and MSR-VTT, performances are enhanced further. The proposed SC-SSL outperforms other state-of-the-art video captioning models using cross-entropy loss for most metrics. Because MSVD is much smaller than MSR-VTT and VATEX, the gap between supervised learning and SC-SSL is more significant than that in Table 3. It is worth mentioning that our supervised result is comparable with that of ECO (Zolfaghari et al., 2018) because the backbone are identical. While our SC-SSL outperforms ECO with a significant gap, this result demonstrate the superiority of SC-SSL.

Table 4. shows the result of an ablation study on different data augmentation methods, temporal weights and baseline reward using 1200 labeled examples on VATEX English captioning task using GRU decoder. It can be seen that SC-SSL using AutoAugment is slightly better than SC-SSL using RandomDrop. SC-SSL w / o temporal weights has lower performances on all metrics than SC-SSL which verifies the temporal weights can decrease

Table 4: An ablation study of the influence of temporal weights, baseline reward and RL training using #1200 labeled examples on VATEX English captioning task.

	BLEU-4	METEOR	ROUGE-L	CIDEr
SC-SSL w AutoAugment	**0.179**	**0.186**	**0.362**	0.229
SC-SSL w RandomDrop	0.176	0.180	0.358	**0.230**
SC-SSL w / o temporal weights	0.172	0.174	0.353	0.222
SC-SSL w / o baseline reward	0.169	0.170	0.350	0.210

the influence of error accumulation. SC-SSL w / o baseline reward means only using reward obtained by Monte-Carlo sampling, i.e. the loss of unlabeled data is $L_u(\theta) = -\sum_{t=1}^{T} \hat{r} \cdot logp_\theta(\hat{y}_t|\hat{y}_{1:t-1}, \boldsymbol{u})$. The performance drops significantly. This result verifies that self-critical training has better performance than traditional policy gradient training.

5 Conclusion

In this paper, we make the first attempt to train the video captioning model in a semi-supervised learning manner. We train labeled data with the traditional cross-entropy loss. For unlabeled data, we leverage a self-critical policy gradient method to train the data. The reward function is the difference between the scores obtained by Monte-Carlo sampling and greedy decoding and the scores are the K-L divergences between output distribution of original video data and augmented video data. The final loss is the weighted sum of two losses mentioned above. Experiments conducted on VATEX, MSR-VTT and MSVD dataset demonstrate that the introduction of unlabeled data can improve the performance of the video captioning model significantly. The proposed method is robust for different tasks (English captioning task and Chinese captioning task), different datasets (VATEX, MSR-VTT and MSVD) and different models (GRU and Transformer). The proposed semi-supervised learning algorithm also outperforms several state-of-the-art semi-supervised learning approaches.

References

Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. 2019. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. CVPR.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Avital Oliver, Nicolas Papernot, and Colin Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. NeurIPS.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. ACL.

Shaoxiang Chen and Yu-Gang Jiang. 2019. Motion guided spatial attention for video captioning. AAAI.

Wenhu Chen, Aurelien Lucchi, and Thomas Hofmann. 2016. A semi-supervised framework for image captioning. *arXiv preprint arXiv:1611.05321*.

Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2018. Less is more: Picking informative frames for video captioning. ECCV.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805:09501*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. EACL 2014 Workshop on Statistical Machine Translation.

Kuncheng Fang, Lian Zhou, Cheng Jin, Yuejie Zhang, Kangnian Weng, Tao Zhang, and Weiguo Fan. 2019. Fully convolutional video captioning with coarse-to-fine and inherited attention. AAAI.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. CVPR.

Jiuxiang Gu, Shafig Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired image captioning by language pivoting. ECCV.

Jiuxiang Gu, Shafig Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. ICCV.

Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, and Yunde Jia. 2019. Joint syntax representation learning and visual cue translation for video captioning. ICCV.

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. ICML.

Thorsten Joachims. 2003. Transductive learning via spectral graph partitioning. ICML.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. EMNLP.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. ICCV.

Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. ICLR.

Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. ICML Workshop on Challenges in Representation Learning.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. Proceedings of Workshop on Text Summarization Branches Out.

Takeru Miyato, Shin ichi Maeda, Shin Ishii, and Masanori Koyama. 2017. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *PAMI*.

Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. CVPR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. ACL.

Ramakanth Pasunuru and Mohit Bansal. 2017. Multi-task video captioning with video and entailment generation. ACL.

Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. 2019. Memory-attended recurrent network for video captioning. CVPR.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. NeurIPS.

Auli M Ranzato M A, Chopra S. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. CVPR.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. NeurIPS.

Jingkuan Song, Lianli Gao, Zhao Guo, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. 2017. Hierarchical lstm with adjusted temporal attention for video captioning. IJCAI.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems.

Ramakrishna Vedantam, Lawrence C. Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. CVPR.

Subhashina Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence - video to text. ICCV.

Subhashini Venugopalan, Lisa Anne Hendricks, Raymond Mooney, and Kate Saenko. 2016. Improving lstm-based video description with linguistic knowledge mined from text. EMNLP.

Bairui Wang, Lin Ma, Wei Zhang, Wenhang Jiang, Jingwen Wang, and Wei Liu. 2019a. Controllable video captioning with pos sequence guidance based on gated fusion network. ICCV.

Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018a. Reconstruction network for video captioning. CVPR.

Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. 2018b. Multimodal memory modelling for video captioning. CVPR.

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2017. Video captioning via hierarchical reinforcement learning. *arXiv preprint arXiv:1711.11135*.

Xin Wang, Yuan-Fang Wang, and William Yang Wang. 2018c. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. *arXiv preprint arXiv:1804.05448*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. ICCV.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2019a. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2019b. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. ICCV.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Junchao Zhang and Yuxin Peng. 2019. Object-aware aggregation with bidirectional temporal graph for video captioning. CVPR.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. CVPR.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. ICML.

Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. 2018. Eco: Efficient convolutional network for online video understanding. Proceedings of the European misc on Computer Vision (ECCV).