Better Captioning with Sequence-Level Exploration

Jia Chen
Carnegie Mellon University
Language Technology Institute
marshimarocj@gmail.com

Qin Jin*
Renmin University of China
School of Information
qjin@ruc.edu.cn

Abstract

Sequence-level learning objective has been widely used in captioning tasks to achieve the state-of-the-art performance for many models. In this objective, the model is trained by the reward on the quality of its generated captions (sequence-level). In this work, we show the limitation of the current sequence-level learning objective for captioning tasks from both theory and empirical result. In theory, we show that the current objective is equivalent to only optimizing the precision side of the caption set generated by the model and therefore overlooks the recall side. Empirical result shows that the model trained by this objective tends to get lower score on the recall side. We propose to add a sequence-level exploration term to the current objective to boost recall. It guides the model to explore more plausible captions in the training. In this way, the proposed objective takes both the precision and recall sides of generated captions into account. Experiments show the effectiveness of the proposed method on both video and image captioning datasets.

1. Introduction

Captioning is one of the core tasks in vision and language fields. The input is an image or video and the output is a descriptive sentence. In terms of the output structure, the descriptive sentence is actually a sequence, which is more complex than the output of classification and detection tasks and therefore poses a challenge for the learning objective in captioning tasks. Furthermore, there exists multiple correct captions for the same input and it is impossible to enumerate all the correct captions when collecting the groundtruth. The above two unique properties, sequence structure and multiple correct grountruth captions, make captioning tasks difficult and worth special treatment for its own learning/training objective.

Most caption models [32, 24, 2] are based on the encoder-decoder architecture and we will only talk about training objectives associated with this architecture. The original training objective is cross-entropy loss [32], which does word-level supervision. To be specific, the decoder is fed with the word from the groundtruth caption at each step and predicts the word at next step. Thus, the decoder is trained to focus on the correctness of predicting each word separately. However, at each step in the test stage, the decoder is fed with the word predicted from the previous step rather than the groundtruth word. This leads to the gap between training and test and limits the performance in the test. Later, sequence-level learning objective is proposed by researchers to address this gap [23, 24]. In this objective, only after the whole sentence is generated by the decoder, the quality of the caption is evaluated by a score and that score is used to guide the model training. That is, the decoder predicts the word at each step based on the word predicted at last step in both training and test stages. The sequence-level learning objective [23, 24] is shown to im-

a man and a woman sitting on a desk
a man and a woman sitting on a table with a laptop computer
a man and a woman sitting on a desk with a laptop computer
a man and a woman sitting on a desk with a laptop computer
a man and a woman sitting on a desk with a laptop computer

Figure 1: Illustration on limitations of current sequence-level learning: 5 captions randomly sampled from the model [24] are almost identical, which indicates that the model is not likely to have high recall.

*corresponding author

prove performance significantly on most evaluation metrics such as CIDEr[31], METEOR[14] and SPICE[1] compared to the cross-entropy loss.

In this paper, we show the limitations of the current sequence-level learning objective from both theoretical and empirical aspects despite its success in captioning tasks. From theoretical aspect, we show that the current objective is equivalent to optimizing the precision side of the predicted caption set. The standard precision is defined based on the set membership of an element. And the set membership function outputs 0-1 for a caption, which describes whether the caption belongs to a set or not. We relax the 0-1 set membership function used in precision calculation to real-value output within range $[0, 1]$. The relaxed set membership function describes the confidence of a caption belonging to a set. In this way, we show that the current sequence-level learning objective is equivalent to maximizing the generalized precision with the relaxed set membership function and it overlooks the recall side of the problem. From empirical aspect, we show that the model trained by the current sequence-level learning objective tends to cover very few different captions in its predictions and gets low score on recall related metrics. As illustrated in figure 1, we randomly sample 5 sentences from the model and the resulting 5 sentences are almost identical.

To overcome the limitations of the current sequence-level learning objective, we propose to add a sequence-level exploration term to boost recall. In this exploration term, we maximize the difference between the generated captions (sequence-level) of the same input. One example of difference measurement could be edit distance. In the context of captioning task, the proposed exploration term corresponds to maximizing the diversity [26] of generated captions. Furthermore, we show that diversity is a proxy measurement of recall for captioning. In training, this term encourages the model to explore more different captions. Such sequence-level exploration is different from the typical maximum-entropy exploration regularization [20] that is put on the policy in reinforcement learning. In typical maximum-entropy exploration regularization, it maximizes the uncertainty of the policy at each step. That is, given generated words up to step t, it maximizes the uncertainty of the next word. We call this word-level exploration.

In summary, the contributions of this work are:
1) We show the limitations of the current sequence-level learning objective for the captioning task from both theoretical and empirical aspects.
2) We propose a new learning objective for the captioning task which adds a sequence-level exploration term to boost recall.
3) The derived solution from the proposed objective achieves better performance on various standard evaluation metrics of the precision side. It also improves the performance on recall related metrics.

2. Related Work

The dominant neural network architecture of the captioning task is based on the encoder-decoder framework [3]. Early works [32, 19, 29] use convolution neural network as encoder and recurrent neural network with LSTM cell [12] as decoder. In the image captioning task, Xu et al. [34] proposed the spatial attention, which selects relevant image regions to generate image descriptions. In the video captioning task, Yao et al. [35] proposed the temporal attention, which expands the attention mechanism in the temporal direction. After that, different variants of attention mechanism are proposed to further improve the performance, such as attention on semantic concepts [37, 22, 16] and adaptive attention on visual and linguistic contexts [27, 17, 36]. The latest variation on attention mechanism is the up-down attention [2] which enables attention to be calculated at the level of objects and other salient image regions. In addition to attention mechanism, researchers also propose other modification on the neural network architecture. Pan et al. [21] utilized the hierarchical encoder to learn better visual representations.

The original objective function [32, 19] used in the captioning task is cross-entropy loss, which applies word-level supervision. To be specific, in training, the model is fed with the groundtruth word at each step and supervision monitors whether the model outputs the correct next word. We call such supervision as word-level supervision. However, in the test stage, the model is fed with the word predicted by itself at last step rather than the groundtruth word. This is known as the train-test gap in sequence prediction tasks. Bengio et al. [4] proposed scheduled sampling, a curriculum learning approach, to minimize such gap. Later, sequence-level training is proposed by Ranzato et al. [23] to systematically address this issue. Different from word-level supervision, the sequence-level learning evaluates the sentence only after the whole sentence has been generated. The sentence is evaluated by a reward about its semantic coherence with the groundtruth caption. And the reward is usually set to be the evaluation metric that has high correlation with human judgement. Rennie et al. [24] further improves the sequence-level learning by introducing a special baseline in reward, which is the score of the caption greedily decoded from the current model. Sequence-level training objective has been widely used in captioning tasks to achieve state-of-the-art performance [2, 18, 28, 6].

3. Limitations of Current Sequence-level Learning

In this section, we show the limitation of current sequence-level learning for the captioning task from both

theoretical and empirical aspects. Theoretically, we show that the current objective function of sequence-level training is equivalent to optimizing the generalized precision with relaxed set membership function on the predicted captions. Empirically, we show that the model trained by the current sequence-level learning tends to generate very few different captions for the same input and does not get high score on recall related metrics.

3.1. Limitation from theory

We first relax the set membership function in the standard precision measurement for the captioning task. Then we show that the objective of current sequence-level learning is actually optimizing the generalized precision with relaxed set membership function in the context of captioning task.

Suppose that the space of all the possible sentences is $\mathcal{Y}$, the groundtruth sentence set of an input (image / video) x_i is Y and the predicted sentence set of that input by the captioning model is $\widetilde{Y}$. Then the precision is defined by:

$$\begin{aligned} Precision(Y, \widetilde{Y}) &= \frac{|Y \cap \widetilde{Y}|}{|\widetilde{Y}|} \\ &= \frac{\sum_{y \in \mathcal{Y}} \delta[y \in Y]\delta[y \in \widetilde{Y}]}{\sum_{y \in \mathcal{Y}} \delta[y \in \widetilde{Y}]} \\ &= \sum_{y \in \mathcal{Y}} \delta[y \in Y] \underbrace{\frac{\delta[y \in \widetilde{Y}]}{\sum_{y' \in \mathcal{Y}} \delta[y' \in \widetilde{Y}]}}_{p(y \in \widetilde{Y})} \\ &= \sum_{y \in \mathcal{Y}} \delta[y \in Y] p(y \in \widetilde{Y}) \end{aligned} \tag{1}$$

Inside the summation of eq (1), it contains two terms: $\delta[y \in Y]$ and $p(y \in \widetilde{Y}) = \frac{\delta[y \in \widetilde{Y}]}{\sum_{y' \in \mathcal{Y}} \delta[y' \in \widetilde{Y}]}$. In the $\delta[y \in Y]$ term, the δ function checks whether or not caption y belongs to groundtruth sentence set Y. In the $p(y \in \widetilde{Y})$ term, the δ function checks whether or not caption y belongs to the predicted sentence set $\widetilde{Y}$.

For the $\delta[y \in Y]$ term, we relax the binary valued δ function to a real-valued function $\Delta(y, Y)$ with output in the range of $[0, 1]$:

$$\delta[y \in Y] \rightarrow \Delta(y, Y) \tag{2}$$

$\Delta(y, Y)$ indicates the likelihood of each individual y within the set Y and is a relaxed set membership function. One natural choice for $\Delta(y, Y)$ is to use the evaluation metric normalized by its maximum value. As all the current evaluation metrics in the captioning task are bounded, they can be normalized properly. For simplicity, we assume that we are dealing with the evaluation metric $\Delta(y, Y)$ that has already been normalized.

The term $p(y \in \widetilde{Y})$ can be interpreted as the chance of the sentence y within set $\widetilde{Y}$. Note that the value of $\delta[y \in \widetilde{Y}]$ is 0-1, which represents whether the captioning model considers sentence y as correct or not. Correspondingly, $p(y \in \widetilde{Y})$ can only take values eithor 0 if $y \notin \widetilde{Y}$ or $\frac{1}{|\widetilde{Y}|}$ if $y \in \widetilde{Y}$. It does not cover the whole range $[0, 1]$ of a probability. If we again relax the 0-1 membership function $\delta[y \in \widetilde{Y}]$ to a real-valued confidence, $p(y \in \widetilde{Y})$ can cover the whole range $[0, 1]$ of a probability. After the relaxation, $p(y \in \widetilde{Y})$ is actually the probability of caption y from the captioning model. Thus by using the relaxed set membership function, we replace $p(y \in \widetilde{Y}) = \frac{\delta[y \in \widetilde{Y}]}{\sum_{y' \in \mathcal{Y}} \delta[y' \in \widetilde{Y}]}$ with $p_\theta(y|x_i)$, which is the probability from the captioning model:

$$p(y \in \widetilde{Y}) = \frac{\delta[y \in \widetilde{Y}]}{\sum_{y' \in \mathcal{Y}} \delta[y' \in \widetilde{Y}]} \rightarrow p_\theta(y|x_i) \tag{3}$$

Substituting $\delta[y \in Y]$ and $p(y \in \widetilde{Y})$ in eq (1) by (2) and (3) respectively, we get the generalized precision (GP) for the captioning task:

$$GP(Y, \theta|x_i) = \sum_{y \in \mathcal{Y}} \Delta(y, Y) p_\theta(y|x_i) \tag{4}$$

We could use generalized precision GP to rewrite the original sequence-level learning objective for the captioning task. Setting $\Delta(y, Y)$ as reward, the original objective is to maximize the expected return:

$$J(\theta) = \sum_{i=1}^{n} \mathbb{E}_{p_\theta(y|x_i)} \Delta(y, Y) \tag{5}$$

By comparing eq (5) with the generalized precision measurement defined in eq (4), we see that they are exactly the same:

$$\begin{aligned} J(\theta) &= \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \Delta(y, Y) p_\theta(y|x_i) \\ &= \sum_{i=1}^{n} GP(Y, \theta|x_i) \end{aligned} \tag{6}$$

This means that sequence-level learning objective only optimizes the precision side of the captions predicted by the captioning model. However, as there exist multiple correct answers for the same input x_i, which means that the recall side should also be taken into account when training the captioning model. On the contrary, the original objective totally overlooks the recall side of the problem.

3.2. Limitation from empirical results

Complementary to the theoretical analysis above, we also measure the precision and recall side of the model

Table 1: Comparison between word-level cross-entropy loss (XE) and sequence-level learning (SLL) on precision and recall sides

Method	Precision	Recall		
	CIDEr (↑)	Div1 (↑)	Div2 (↑)	mBleu4 (↓)
XE	74.2	0.57	0.78	0.06
SLL	114.6	0.25	0.32	0.81

trained by current sequence-level learning objective. The precision side could be measured by the standard evaluation metrics in captioning tasks such as METEOR[14] and SPICE[1]. As it is not possible to collect all the correct answers for an input x_i, directly computing recall is not feasible. Instead, we use set level diversity metrics [26] *Div-1*, *Div-2* and *mBleu* as a proxy measurement of the recall. The set level diversity metrics are defined on a set of captions, $\widetilde{Y}$, corresponding to the same input x_i.

- *Div-1* ratio of the number of unique unigrams in $\widetilde{Y}$ to the number of words in $\widetilde{Y}$. Higher is more diverse.
- *Div-2* ratio of the number of unique bigrams in $\widetilde{Y}$ to the number of words in $\widetilde{Y}$. Higher is more diverse.
- *mBleu* Bleu score is computed between each caption in $\widetilde{Y}$ against the rest. Mean of these Bleu scores is the mBleu score. Lower is more diverse.

To report set level diversity metrics, we sample 5 captions from the model for each input. Correspondingly, when calculating the precision metric CIDEr, we average the CIDEr scores of the 5 sampled captions.

Here is the reasoning of why the above diversity metrics is related to recall. Standard recall is defined by:

$$\begin{aligned} Recall(Y, Y) &= \frac{|Y \cap \widetilde{Y}|}{Y} \\ &\propto |Y \cap \widetilde{Y}| \\ &\propto |\widetilde{Y}| Precision(Y, \widetilde{Y}) \end{aligned} \tag{7}$$

When the precision is fixed, we see that the recall is proportional to the size of the predicted set $\widetilde{Y}$. To compare the recall at the same precision level, we could instead compare the size of the predicted caption set from the model. In this way, any measurement on the size of set $\widetilde{Y}$ could be considered as a proxy measurement of recall. Directly measuring the size of $\widetilde{Y}$ by the number of captions is not meaningful if we are allowed to sample infinite times from the model. A more meaningful way to measure the size of $\widetilde{Y}$ is: *given fixed number of sampling times, calculating the difference between sampled captions*. And this is exactly the quantity defined in set level diversity metrics.

As shown in table 1 compared to word-level cross-entropy (XE) loss, sequence-level learning (SLL) leads to a large performance drop on the recall side though it improves the metrics on the precision side significantly. This could be further illustrated by the examples shown in figure 2. In this example, 5 randomly sampled captions are almost identical for the model trained by sequence-level learning (SLL) objective while this is not an issue for the model trained by the word-level cross-entropy (XE) objective. We explain this observation by the peak width of the distribution. As

XE:
a couple of men standing in the ocean holding surfboards
a surfer walking through the ocean with his surfboard
a couple of people walking through the water
two men in a beach holding surfboards in the water
a surfer carrying his surfboard while another surfer walks into the water
SLL:
a couple of people standing in the ocean with surfboards
a couple of people standing in the ocean with surfboards
a couple of people standing in the ocean with surfboards
a couple of people standing in the ocean with surfboards
a couple of people standing in the ocean holding surfboards

Figure 2: Illustration of 5 captions sampled from models given the same input: XE is the model trained by cross-entropy objective and SLL is the model trained by sequence-level learning objective.

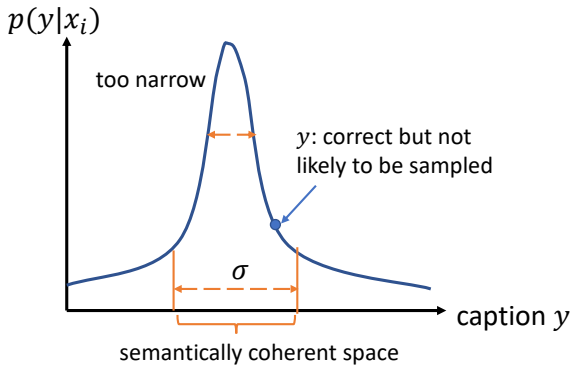

Figure 3: Illustration of the peak width of caption distribution $p(y|x)$ based on empirical results of the sequence-level learning objective

illustrated in figure 3, suppose we project the captions to a one-dimensional space and the width of the line segment containing semantic coherent captions for an input x_i is σ. Based on the empirical result observed in this section, the peak width of the model trained by SLL objective should be much smaller than σ so that most sampled sentences for input x_i are almost identical. However, the peak width of an ideal model should be similar to σ. In this case, the samples from the model is likely to cover the semantically coherent space and get high score on recall as a result.

4. Solution

We first propose a new objective function to address the limitations of current sequence-level learning objective shown in the last section. Then we derive the optimization procedure for this new objective function. Finally, we describe the network architecture and training details in implementation.

4.1. Objective Function

As we have shown that diversity is a proxy measurement of recall, we introduce an additional diversity term to the original sequence-level learning objective function to cover the recall side of the problem:

$$\begin{aligned}\max_\theta :&\underbrace{\alpha \sum_{y\in\mathcal{Y}} \Delta(y, y_i)p_\theta(y|x_i)}_{\text{precision}} + \\ &(1-\alpha)\underbrace{\sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{Y}} d(y, y')p_\theta(y|x_i)p_\theta(y'|x_i)}_{\text{diversity}}\end{aligned} \tag{8}$$

In this objective function, x_i is the input image or video, y_i is the groundtruth caption, y and y' are any two captions in the caption space $\mathcal{Y}$ that can be sampled from the caption model. $p_\theta(y|x_i)$ is the conditional probability given by the caption model.

• $\Delta(y, y_i)$ in precision term measures semantic coherence between caption y and the groundtruth caption y_i. It is equivalent to $\Delta(y, Y)$ when there is only one groundtruth caption y_i of input x_i. It encourages the model to put more probability mass $p_\theta(y|x_i)$ on captions that is semantically coherent with the groundtruth. Example choices for $\Delta(y, y_i)$ could be METEOR, CIDEr, SPICE, which are shown to have good correlation with human judgements.

• $d(y, y')$ in diversity term measures the syntactic difference between two captions. It encourages the model to explore more different ways to express the same semantic meaning. Example choices for $d(y, y')$ could be edit distance or BLEU3/4, which measures the difference in sentence structure.

The diversity term is different from the standard maximum-entropy regularization used in reinforcement learning [20], which is put on the *policy* by $\mathbb{H}(p_\theta(w_j|w_{<j}, x_i))$ and maximizes the uncertainty of the next step word w_j given the past words $w_{<j}$. The diversity term introduced here is directly put on captions, which are *trajectories* in the reinforcement learning. Furthermore, we use distance d rather than entropy of captions to avoid the intractable estimation of denominator Z that involves summing over the probability of all captions. Using distance d also offers us more flexibility to plug-in any measurement of difference in sentence structure. Thus, compared to standard maximum-entropy regularization, the diversity term has more direct effect on encouraging the model to explore more different captions and is more flexible for more syntactic difference measurements.

Putting both precision term and diversity term together, the meaning of the proposed objective function is to encourage the model to *explore more captions different in syntax but are semantically coherent with the groundtruth caption y_i of input x_i*. Hyper-parameter α is introduced to balance between precision and diversity terms.

4.2. Optimization

We first show that the precision term in the objective function could be directly solved using REINFORCE algorithm [30]. Then we show that the diversity term could be solved with some variation on the technique used in the REINFORCE algorithm. Finally, we derive the surrogate loss and a complete algorithm for our objective function.

In optimization convention, we always minimize the objective function. Thus, we take negation of the objective function in eq (8) and decompose it into two parts:

$$\begin{aligned}L(\theta) &= \alpha L_1(\theta) + (1-\alpha)L_2(\theta) \\ L_1(\theta) &= -\sum_{y\in\mathcal{Y}} \Delta(y, y_i)p_\theta(y|x_i) \\ L_2(\theta) &= -\sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{Y}} d(y, y')p_\theta(y|x_i)p_\theta(y'|x_i)\end{aligned} \tag{9}$$

1. Solution to $L_1(\theta)$: We could rewrite L_1 as expectation:

$$\begin{aligned}L_1(\theta) &= -\sum_{y\in\mathcal{Y}} \Delta(y, y_i)p_\theta(y|x_i) \\ &= -\mathbb{E}_{p_\theta(y|x_i)}[\Delta(y, y_i)]\end{aligned} \tag{10}$$

We could use REINFORCE [30] to calculate its gradient:

$$\begin{aligned}\nabla L_1(\theta) &= -\mathbb{E}_{p_\theta(y|x_i)}[\Delta(y, y_i)\nabla \log p_\theta(y|x_i)] \\ &\approx -\Delta(\tilde{y}, y_i)\nabla \log p_\theta(\tilde{y}|x_i)\end{aligned} \tag{11}$$

The second line is Monte Carlo sampling with just one sample caption $\tilde{y}$ from the model.

2. Solution to $L_2(\theta)$: we could also rewrite L_2 as expecta-

tion:

$$L_2(\theta) = -\sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{Y}} d(y,y')p_\theta(y|x_i)p_\theta(y'|x_i) \\ = -\mathbb{E}_{p_\theta(y|x_i)}\mathbb{E}_{p_\theta(y'|x_i)}d(y,y') \tag{12}$$

We see that there are two expectations involved. We could still apply REINFORCE to the outer expectation and inner expectation respectively and get:

$$\nabla L_2(\theta) = -\mathbb{E}_{p_\theta(y'|x_i)}\Big[\mathbb{E}_{p_\theta(y|x_i)}[d(y,y')]\nabla\log p_\theta(y|x_i)\Big] \\ -\mathbb{E}_{p_\theta(y|x_i)}\Big[\mathbb{E}_{p_\theta(y'|x_i)}\big[d(y,y')\nabla\log p_\theta(y'|x_i)\big]\Big] \tag{13}$$

Approximating it by Monte Carlo sampling leads to the following solution: we sample s captions $\widetilde{y}_1,\ldots,\widetilde{y}_s$ and calculate pairwise distances. For each sample $\widetilde{y}_j$, its corresponding gradient is:

$$\nabla L_2(\theta) = -\frac{2}{s^2}\sum_{j=1}^{s}\Big(\sum_{k=1}^{s} d(\widetilde{y}_j,\widetilde{y}_k)\nabla\log p_\theta(\widetilde{y}_j|x_i)\Big) \tag{14}$$

3. Complete solution: In standard policy gradient of reinforcement learning, the multiplier before $\nabla\log p_\theta(\widetilde{y}_j|x_i)$ represents the reward. In the gradient of L_2, the multiplier is $\sum_{k=1}^{s} d(\widetilde{y}_j,\widetilde{y}_k)$ for each sample $\widetilde{y}_j$. It is the sum of sample $\widetilde{y}_j$'s distance to other samples of input x_i. This aligns exactly with our formulation of L_2, which is the diversity term. This multiplier could be further considered as "reward" that involves multiple samples of the input x_i jointly in calculation while calculating standard reward only uses each sample separately.

Finally, we wrap up all the gradients of $L(\theta)$ in the following surrogate loss of the entire stochastic computation graph [25]:

$$\mathcal{L}(\theta) = \frac{1}{s}\sum_{j=1}^{s}\mathcal{L}^j(\theta) \tag{15}$$

$$\mathcal{L}^j(\theta) = -\alpha\Delta(\widetilde{y}_j,y_i)\log p_\theta(\widetilde{y}_j|x_i) \\ -(1-\alpha)\frac{2}{s}\sum_{k=1}^{s} d(\widetilde{y}_j,\widetilde{y}_k)\log p_\theta(\widetilde{y}_j|x_i) \tag{16}$$

Following the standard procedure in sequence-level learning of the captioning task, we first train the model by the word-level cross-entropy loss and then switch to this surrogate loss for training. Algorithm 1 summarizes the entire training process.

4.3. Network Architecture and Training Details

Our proposed objective and solution is compatible with any captioning model that follows the encoder-decoder architecture [32]. The encoder depends on the input (image or video) and will be specified in the experiment section. The decoder is an RNN model of LSTM cell with hidden dimension set to 512. We add one full connection layer after the encoder to reduce the dimension to 512. In step 0, the hidden state is initialized by the output of this full connection layer.

We use CIDEr metric to calculate $\Delta(y,y_i)$ and we use BLEU3 + BLEU4 to calculate $d(y,y')$ in eq (15). We set the number of samples s to 5. To reduce the variance introduced in the Monte Carlo sampling step when estimating the gradient in optimization, we follow the standard practice of using baseline. For the gradient of precision term, we set its baseline to the CIDEr score of greedily decoded caption from the model following work [10]. For the gradient of diversity term, we set it to $\frac{1}{s^2}\sum_{k=1}^{s}\sum_{j=1}^{s} d(\widetilde{y}_j,\widetilde{y}_k)$, the average of all the pairwise distances between sampled captions. We use ADAM optimizer in optimization.

Algorithm 1 Training algorithm of sequence-level exploration

```
for epoch in [0, M) do
    train by cross-entropy loss
end for
for epoch in [M, N) do
    for each instance x_i do
        sample s captions ỹ_1, ..., ỹ_s
        for each sample ỹ_j do
            calculate L^j(θ) as in eq (16)
        end for
        calculate surrogate loss L(θ) as in eq (15)
        update parameter θ by stochastic gradient descent
    end for
end for
```

5. Experiment

In this section, we first introduce the experiment setup. Then we report the performance of the model trained by our proposed objective on standard evaluation metrics of precision side in the image captioning task and video captioning task respectively. Finally, we discuss the model behavior on both precision and recall sides.

5.1. Experiment Setup

For the image captioning task, we use the MSCOCO dataset [8], which is one of the largest image caption datasets that contains more than 120K images crawled from Flickr. Each image is annotated with 5 reference captions. We use the public split [13] for experiments. For the video captioning task, we use the TGIF dataset [15], which is one of the largest video caption datasets that contains 100K animated GIFs collected from Tumblr and $120K$ caption sentences. We use the official split [15] for experiments.

For image, we use Resnet152 [11] pretrained on ImageNet [9] and apply spatial mean pooling to get a 2048-dim feature vector. For video, we also use Resnet152 [11] for fair comparison to other works rather than use a stronger CNN such as I3D [5]. We apply spatial-temporal mean pooling to get a 2048-dim feature vector. For simplicity, we don't finetune the feature on the caption datasets. We tune the hyper-parameter α in eq (8) among .25, .5 and .75 on the validation set and set it to .75. We find that .75 is a quite stable value to reach the best performance across different datasets.

5.2. Image Captioning

We first study the contribution of our proposed objective by comparing it to training our model with the original sequence-level learning loss (SLL) and sequence-level learning with maximum entropy regularization (SLL-ME) [20]. The weight of the maximum-entropy regularization in SLL-ME is tuned among 10^{-1}, 10^{-2}, 10^{-3} and set to 10^{-2} for the best performance. Both the network architecture and input feature are the same across SLL, SLL-ME and SLL-SLE (ours). We use beam search in test stage with width of 5. As shown in the middle block from table 2, we can see that our model SLL-SLE improves over SLL and SLL-ME significantly on all metrics. The improvement of SLL-SLE over SLL-ME on all metrics (Meteor: 0.2, CIDEr: 1.8, SPICE: 0.2) is much larger than the improvement of SLL-ME over SLL on all metrics (Meteor: 0.0, CIDEr: 0.6, SPICE: 0.1). This shows that the typical maximum-entropy regularization doesn't help to solve the issue of original sequence-level objective in the captioning task. Our proposed sequence-level exploration is effective in guiding the model to explore more plausible captions in training and consequently SLL-SLE generates more accurate captions in test. In the last block of table 2, we also include results of SLL, SLL-ME, SLL-SLE objectives when combined with attention architecture. Again the similar trend is observed: SLL-SLE improves over SLL and SLL-ME significantly.

We also compare our proposed model to various state-of-the-art (SOTA) models with different network architectures trained by either word-level cross-entropy loss or sequence-level learning objective. For word-level XE loss, we compare to NIC model [32], Adaptive [17], Top-down attention [2]. For sequence-level learning objective (SLL), we compare to self-critical learning (SCST:FC & SCST:Att2in) [24] and Top-Down attention [2]. As shown in table 2, we see that the proposed objective leads to better performance on all metrics over all SOTA models.

5.3. Video Captioning

Similarly, we first compare our proposed objective with original sequence-level learning loss (SLL) and sequence-level learning with maximum entropy regularization (SLL-ME). As we fix the hyper-parameter across datasets for our method (SLL-SLE), we also fix the hyper-parameter (weight before maximum-entropy regularization) in SLL-ME and set it to 10^{-2}, same as that on MSCOCO dataset. We use beam search with width of 5 in test stage. As shown in the last three rows from table 3, we can see that our model, SLL-SLE, again improves over SLL and SLL-ME significantly on all metrics. Actually, SLL-ME performs worse than SLL on all metrics, which indicates that the maximum-entropy regularization is not stable across datasets and may even deteriorate the performance in some captioning task. Our model, SLL-SLE improves over SLL by 0.6 on Meteor, 2.7 on CIDEr and 0.6 on SPICE with the same hyper-parameter setting as that on MSCOCO. This shows that the proposed sequence-level exploration term is stable and robust across datasets and are helpful to the model performance in general.

We also compare our proposed model to various state-of-the-art (SOTA) models on the video captioning task. The TGIF dataset comes with an official baseline (Official) [15] trained by word-level cross-entropy loss. Show-adapt [7] leverages both TGIF and other datasets in training. By comparing our implementation of baseline model SLL to these

Table 2: Performance improvement on the image captioning: * means bottom-up region features are used with attention architecture

Method	Meteor	CIDEr	Spice
NIC [32]	23.7	85.5	NA
Adaptive [17]	26.6	108.5	NA
SCST:FC [24]	25.5	106.3	NA
SCST:Att2in [24]	26.3	111.4	NA
Top-Down-XE [2]	26.1	105.4	19.2
Top-Down-SLL [2]	26.5	111.1	20.2
SLL	26.8	115.0	20.0
SLL-ME	26.8	115.6	20.1
SLL-SLE (ours)	**27.0**	**117.2**	**20.3**
SLL*	26.6	117.2	19.4
SLL-ME*	26.7	117.9	19.5
SLL-SLE* (ours)	**27.0**	**119.6**	**19.9**

Table 3: Performance improvement on the video captioning

Method	METEOR	CIDEr	SPICE
Official[15]	16.7	31.6	NA
Show-adapt[7]	16.2	29.8	NA
SLL	17.8	45.9	15.9
SLL-ME	18.2	48.1	16.0
SLL-SLE (ours)	**18.8**	**50.8**	**16.6**

Table 4: Comparison of models trained by XE, SLL, SLL-ME, our SLL-SLE on both precision and diversity sides (MSCOCO dataset): (rs) denotes random sampling decoding and (bs) denotes beam search decoding

Method	precision	recall		
	CIDEr	Div1 (↑)	Div2 (↑)	mBleu4 (↓)
XE (rs)	74.2	0.57	0.78	0.06
SLL (rs)	114.6	0.25	0.32	0.81
SLL-ME (rs)	115.1	0.25	0.33	0.80
SLL-SLE (rs)	115.9	0.29	0.40	0.68
XE (bs)	102.5	0.27	0.35	0.80
SLL (bs)	115.0	0.26	0.35	0.78
SLL-ME (bs)	115.6	0.26	0.34	0.79
SLL-SLE (bs)	**117.2**	**0.27**	**0.36**	**0.76**
VAE[33] (bs)	100.0	NA	NA	NA
GAN[26] (rs)	NA	0.41	0.55	0.51
GAN[26] (bs)	NA	0.34	0.44	0.70

models, we see that it performs better than them, which indicates that SLL is already a very strong baseline. This further suggests that the improvement over SLL is not trivial.

5.4. Discussion of Model Behavior on Precision and Recall

We study the model behavior on precision and recall sides for these objectives: cross-entropy (XE), sequence-level learning (SLL), sequence-level learning with maximum-entropy (SLL-ME), our SLL-SLE. On the precision side, we use CIDEr metric as it is shown to have good correlation with human judgement. On the recall side, we use diversity metrics Div1, Div2, mBleu[26] as proxy measurements. To calculate the diversity metrics, we adopt two decoding straties as [26]. The first decoding strategy is to sample 5 captions from the model for each image (rs). The second decoding strategy is to beam search top 5 captions from the model for each image (bs). The reported CIDEr is the average of CIDEr scores of the 5 sampled captions. As shown in table 4, compared to SLL and SLL-ME, the proposed objective, SLL-SLE, performs not only better on the precision side and but also better on the recall side under both random sampling and beam search decoding strategies. Compared to XE, SLL-SLE improves on both precision and recall aspects under beam search decoding strategies. We also list VAE and GAN's performance on precision and recall aspects for reference.

Figure 4 shows that the proposed objective can generate diverse and high quality captions with sampling strategy. The quality of captions generated by the XE model is not good. The SLL model with sampling strategy has limited diversity and keeps generating almost the same caption with sampling strategy.

XE:
a person standing in a bathroom holding a book
a man is standing next to an open toilet
a man is standing in front of a toilet
a man is standing in front of a toilet
a man sitting on a chair with his feet up
SLL:
a man standing in a bathroom with a toilet
a man standing in a bathroom with a toilet
a man standing in a bathroom with a toilet
a man standing in a bathroom with a toilet
a man standing in a bathroom with a toilet
SLL-SLE:
a man that is holding a swim in a toilet
a man sitting next to a toilet reading a book
a man standing on top of a toilet reading a book
a man sitting in a toilet reading a book
a man reading a newspaper next to a toilet paper

Figure 4: Case study of model behavior on precision and recall by sampling strategy in decoding

6. Conclusion

In this work, we show the limitation of current sequence-level learning objective in captioning tasks from both theoretical and empirical aspects. From the theoretical aspect, this objective is equivalent to maximizing the generalized precision of the predicted caption set, which ignores the recall side. From the empirical aspect, models trained by this objective receive low score on proxy measurements of recall. To overcome the above limitations, we propose adding a sequence-level exploration term to maximize the diversity, a proxy measurement of recall, on generated captions. It encourages the model to explore more captions that are different in syntax but are semantically coherent with the groundtruth in training. Extensive experiments on both image and video captioning tasks show that the proposed objective leads to a win-win solution that consistently performs better on both precision and recall.

7. Acknowledgement

We would like to express our great appreciation to Shiwan Zhao for insightful discussions and valuable suggestions. This work was partially supported by National Natural Science Foundation of China (No. 61772535) and Beijing Natural Science Foundation (No. 4192028).

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 2, 4

[2] 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. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086, 2018. 1, 2, 7

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 2

[4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1171–1179, 2015. 2

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733, 2017. 7

[6] Shizhe Chen, Jia Chen, Qin Jin, and Alexander Hauptmann. Video captioning with guidance of multimodal latent topics. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1838–1846, 2017. 2

[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. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 521–530, 2017. 7

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. 7

[10] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. *arXiv preprint arXiv:1611.08002*, 2016. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[13] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137, 2015. 6

[14] Michael Denkowski Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. *ACL*, page 376, 2014. 2, 4

[15] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, pages 4641–4650, 2016. 6, 7

[16] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 338–354, 2018. 2

[17] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, 2017. 2, 7

[18] Ruotian Luo, Brian L. Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6964–6974, 2018. 2

[19] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632, 2014. 2

[20] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1928–1937, 2016. 2, 5, 7

[21] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, pages 1029–1038, 2016. 2

[22] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017. 2

[23] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732, 2015. 1, 2

[24] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016. 1, 2, 7

[25] John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3528–3536, 2015. 6

[26] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4155–4164, 2017. 2, 4, 8

[27] Jingkuan Song, Lianli Gao, Zhao Guo, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. Hierarchical LSTM with adjusted temporal attention for video captioning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2737–2743, 2017. 2

[28] Yuqing Song, Shizhe Chen, Yida Zhao, and Qin Jin. Unpaired cross-lingual image caption generation with self-supervised rewards. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 784–792, 2019. 2

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014. 2

[30] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. 5

[31] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 2

[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 1, 2, 6, 7

[33] Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NIPS*, 2017. 8

[34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044*, 2015. 2

[35] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015. 2

[36] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 2

[37] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *arXiv:1603.03925*, 2016. 2