Language Resource Efficient Learning for Captioning

Jia Chen[*,†], **Yike Wu**[1,*], **Shiwan Zhao, Qin Jin**[2,‡]
[1]College of Computer Science, Nankai University, Tianjin, China
[2]School of Information, Renmin University of China, Beijing, China
sjtu_chenjia@163.com, wuyike@dbis.nankai.edu.cn
zhaosw@gmail.com, qjin@ruc.edu.cn

Abstract

Due to complex cognitive and inferential efforts involved in the manual generation of one caption per image/video input, the human annotation resources are very limited for captioning tasks. We define language resource efficient as reaching the same performance with fewer annotated captions per input. We first study the performance degradation of caption models in different language resource settings. Our analysis of caption models with SC loss shows that the performance degradation is caused by the increasingly noisy estimation of reward and baseline with fewer language resources. To mitigate this issue, we propose to reduce the variance of noise in the baseline by generalizing the single pairwise comparison in SC loss and using multiple generalized pairwise comparisons. The generalized pairwise comparison (GPC) measures the difference between the evaluation scores of two captions with respect to an input. Empirically, we show that the model trained with the proposed GPC loss is efficient on language resource and achieves similar performance with the state-of-the-art models on MSCOCO by using only half of the language resources. Furthermore, our model significantly outperforms the state-of-the-art models on a video caption dataset that has only one labeled caption per input in the training set.

1 Introduction

Generating natural language descriptions for images and videos (Vinyals et al., 2015; Chen et al., 2015; Yao et al., 2015; Li et al., 2016) is one of the core steps towards ultimate image and video understanding. However, the cost of collecting a caption dataset is nontrivial. Actually, it is much higher than the cost of collecting a detection/classification dataset with the same number of images/videos, since annotating an image/video with a caption involves more complex cognitive and inferential efforts for human beings. That means the manual labeling effort is a very limited resource that could not be neglected in collecting a caption dataset. This issue becomes even more critical as researchers move on to new domains and need more data to train caption models to cover new scenarios or tasks.

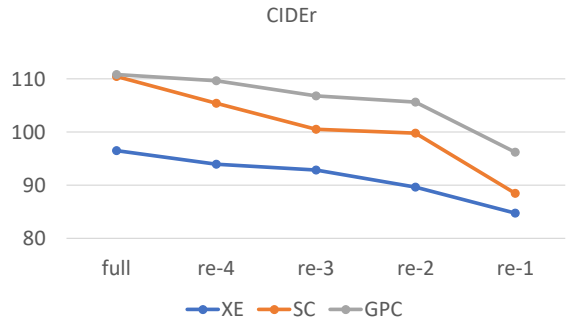


Figure 1: Behavior of models in different language resource settings: "re-K" means K labeled captions per input are used in training and "full" means using all 5 labeled captions per input in training. XE loss is supervised learning (Vinyals et al., 2015); SC loss is reinforcement learning (Rennie et al., 2017); GPC loss is the proposed method.

The scale of a caption dataset can be defined by the product of the number of images/videos and the number of captions per input. Thus there are two ways to reduce the labeling resources: reducing the number of images/videos (image/video resource) or reducing the number of captions per input (language resource). In this paper, we focus on "language resource efficient" that aims to reach the same performance with fewer language resources. As shown in figure 1, with fewer number of captions provided per input from full setting to re-1 setting in training, the performance of the model degrades in general. Slower degrading curve means that the corresponding model requires fewer number of captions per input to reach the

[*]Equal contribution.
[†]Work performed at Carnegie Mellon University.
[‡]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1887–1895
November 7–11, 2021.

performance of other models and therefore is more resource-efficient.

In figure 1, the performance degradation of models with XE loss trained by supervised learning is easy to understand as there are fewer supervisions with fewer captions per input. However, the performance degradation of models with SC loss trained by reinforcement learning is more complex. In SC loss, we sample one caption from the model for each input in each epoch of training. Thus, the number of captions sampled from the model in training is irrelevant to the size of language resource. On the other hand, when calculating the reward in SC loss, we first calculate the evaluation score based on each groundtruth caption and then average the evaluation scores across multiple groundtruth captions of the same input, which is related to the size of language resource. Similar things happen when we calculate the baseline value in the SC approach, which evaluates the score of the caption decoded greedily from the model. Thus, the calculation of reward and baseline is affected by the size of language resource. As the evaluation score, e.g., CIDEr (Vedantam et al., 2015), still has some gap from human judgement, we consider that both the reward and baseline in the SC model have noise. Such noise is amplified when the language resource is small, which contributes to the performance degradation of models with SC loss.

To solve this issue, we introduce generalized pairwise comparison (GPC) to reduce the noise in the baseline. GPC measures the difference between the evaluation scores of two captions with respect to an input. For example, neither *"a white cat plays with a ball"* nor *"a dog plays with a ball"* is a correct caption for the input image with a groundtruth caption *"a brown cat plays with a brown ball"*. However, the former is much closer to the groundtruth. Such subtlety comes from the fact that a caption is a complex object. We use the pairwise comparison to quantify the difference rather than the absolute value. We propose to combine multiple GPCs by comparing the sampled caption to multiple other captions, which results in GPC loss. The GPC loss can be decomposed into two items. The first item is the reward of the sampled caption, same as the reward in SC. The second item is the average evaluation score of multiple captions other than the sampled caption, which works as a new baseline with smaller noise in estimation.

We show theoretically and empirically that GPC loss has a less noisy baseline compared to SC loss. Experimental results show that our proposed GPC can achieve the same performance as the state-of-the-art SC loss (Rennie et al., 2017) while using only half of the captions per input on MSCOCO (Chen et al., 2015). We also test GPC on a video caption dataset TGIF (Li et al., 2016), which has only one caption per video in the training set (low language resource), and achieve significant performance improvement over the state-of-the-art models on all metrics.

In summary, the main contributions of this work are as follows:

- We propose to optimize language resource efficiency in captioning tasks.
- We study and analyze the behavior of models trained by supervised learning and reinforcement learning in terms of language resource efficiency.
- We propose generalized pairwise comparison (GPC) to reduce noise in the baseline.
- Extensive experiments are conducted to assess the language resource efficiency of the model trained by the proposed GPC. We achieve the state-of-the-art performance by using only *half* of the captions per input on MSCOCO, and improve performance significantly on all metrics on TGIF.

2 Related Work

With the success of the encoder-decoder architecture in machine translation (Bahdanau et al., 2014), researchers begin to apply the encoder-decoder architecture to directly generate image/video descriptions in an end-to-end way (Vinyals et al., 2015; Mao et al., 2014; Sutskever et al., 2014). Convolutional neural networks are utilized as the encoder to encode visual contents as distributional vector representations, and recurrent neural networks are widely used as the decoder to produce natural and meaningful description sentences. The success of encoder-decoder architectures in the captioning task has attracted more research interest on this topic. Many research works have been proposed to improve the basic architecture. For example, the spatial (Xu et al., 2015; Li et al., 2017) and temporal (Yao et al., 2015) attention mechanisms have been proposed in image and video captioning respectively to dynamically select relevant visual content for generating future words. Semantic

concepts produced in object detection and action recognition tasks are also beneficial to caption generation and have been encoded into the decoder in different approaches (You et al., 2016; Pan et al., 2017). All these research works focus on improving the network architecture under the supervised problem formulation with cross-entropy loss.

Recently, researchers (Ranzato et al., 2015; Rennie et al., 2017; Liu et al., 2017) have proposed to use reinforcement learning to bridge the gap between training and testing in the captioning task, which considerably boosts the performance. Dai and Lin (Dai and Lin, 2017) combine the contrast loss with the cross-entropy loss to generate more discriminating captions. Luo et al. (Luo et al., 2018) take a different approach to generate more discriminating captions by adding the contrastive loss as reward. All these works improve the caption performance by changing the loss functions.

Meanwhile, many caption datasets (Chen et al., 2015; Xu et al., 2016; Li et al., 2016; Sigurdsson et al., 2016) covering different media such as images, GIFs and videos have been proposed to promote the captioning research. Among them, MSCOCO is one of the largest one in both the number of instances and the number of instance-caption pairs. Many of the caption datasets are either small in the number of instances such as MSRVTT (Xu et al., 2016) or the number of captions per instance such as TGIF (Li et al., 2016) and Charade (Sigurdsson et al., 2016) due to the limited budget for collecting data. Surprisingly, the huge cost of data collecting has been neglected and most research works are trained and evaluated on rich resource datasets such as MSCOCO. To the best of our knowledge, we are the first to take the labeling resource into account at the beginning to develop the model.

3 Study of Language Resource Efficiency

3.1 Language Resource Efficiency

The resource we talk about here is the human labeling resource, which is the major bottleneck in collecting a large scale caption dataset. The scale of a dataset can be measured, for example, by the number of images/videos N times the number of captions per input K, which results in $N \times K$ in total. Correspondingly, the cost of labeling a caption dataset could be roughly estimated by its scale $N \times K$. The major labeling cost comes from the training set. To save labeling efforts for datasets that contain many images/videos, researchers cut down the number of labeled captions per input such as the TGIF (Li et al., 2016) dataset. Given a fixed amount of images/videos in the training set, if one model achieves the same performance as another model using fewer number of labeled captions per input in training, this model is more language efficient as fewer labeling efforts are needed for training. In this way, the *language resource efficiency* is defined as the number of labeled captions per input in the training set. Note that the language resource efficiency doesn't apply to the test set as multiple captions per input are helpful for stable and robust evaluation (Vedantam et al., 2015). Furthermore, the labeling cost of the test set is usually not the focus as the number of images/videos in the test set constitutes only a small fraction of the whole set.

We construct a series of different language resource settings from the caption dataset MSCOCO (Chen et al., 2015) for a systematic study of current models. To be specific, MSCOCO contains 5 captions per image in the training set. For each image, we randomly preserve only one caption and construct the re-1 training setting. Similarly, we could randomly preserve K captions for each image and get corresponding re-K training setting, where K ranges from 1 to 4. Together with the full setting containing all the available captions, we have 5 settings in total: re-1, re-2, re-3, re-4, full. We only apply the 5 settings on the training set. The test set still contains 5 captions per input to guarantee the stable evaluation result. We use the standard split (Karpathy and Li, 2015) for all experiments.

3.2 Current Model Behavior on Language Resource Efficiency

Under the five resource settings constructed in the above subsection, we study models trained by two widely used objective functions in captioning tasks: cross-entropy (XE) loss (Vinyals et al., 2015; Mao et al., 2014) and self-critical (SC) loss (Rennie et al., 2017). The objective function of XE loss is:

$$\min : -\sum_{i=1}^{N}\sum_{j=1}^{K} \log p(y_i^j|x_i) \quad (1)$$

where x_i is the input image, y_i^j is the j-th groundtruth caption for image x_i. It does word-level supervision and is limited by the train-to-test gap for sequence prediction (Ranzato et al., 2015). In contrast, SC loss (Rennie et al., 2017) doesn't have the train-to-test gap and reaches better perfor-

mance than XE loss. Its objective function maximizes the expected return of sampled caption y^s:

$$\max : \sum_{i=1}^{N} \mathbb{E}_{y^s \sim p(y^s|x_i)} \left[r(y^s) - b \right] \tag{2}$$

As shown in figure 1, we observe that for both learning objective functions XE and SC, the performance drops almost linearly on CIDEr when the number of captions per input decreases. From full setting to re-1 setting, the model trained by XE loss drops by 11.8 on CIDEr and the model trained by SC loss drops by 21.9 on CIDEr. The experimental setup is elaborated in section 5.1.

It is easy to understand the performance drop of models with XE loss as fewer labels are provided for supervision from full setting to re-1 setting. However, the performance drop of models with SC loss requires more complex reasoning. In this approach, we sample one caption from the model for each input in each epoch of training. The amount of captions sampled from the model in the training process is proportional to the number of images/videos in the training set and is therefore *fixed* across different resource settings. Thus, the performance degradation is related to the calculation of reward and baseline.

Next we analyze how the number of captions per input influences the calculation of reward and baseline, which further leads to the performance degradation. The reward $r(y^s)$ and baseline b in eq (2) are calculated via a evaluation score ϕ:

$$\begin{aligned} r(y^s) &= \frac{1}{K} \sum_{j=1}^{K} \phi(y^s, y_i^j) \\ b &= \frac{1}{K} \sum_{j=1}^{K} \phi(y^g, y_i^j) \end{aligned} \tag{3}$$

In the above equation, calculating reward $r(y^s)$ for the sample caption y^s involves K calls of the evaluation score $\phi(y^s, y_i^j)$ and each call uses one of the groundtruth caption to calculate the score. Similarly, calculating baseline b also involves K calls of the evaluation score $\phi(y^g, y_i^j)$ on the caption y^g which is decoded greedily from the model. Ideally, the evaluation score ϕ should be exactly equivalent to the human judgement, and we denote this "perfect" evaluation score as $\widetilde{\phi}$. Intuitively, the evaluation score $\widetilde{\phi}(y)$ of any caption y should be the same no matter which groundtruth caption is used for evaluation:

$$\widetilde{\phi}(y) = \widetilde{\phi}(y, y_i^1) = \cdots = \widetilde{\phi}(y, y_i^K) \tag{4}$$

Therefore, in an ideal situation, the calculation of reward and baseline is independent of the number of captions:

$$\begin{aligned} \widetilde{r}(y^s) &= \frac{1}{K} \sum^{K} \widetilde{\phi}(y^s) = \widetilde{\phi}(y^s) \\ \widetilde{b} &= \frac{1}{K} \sum^{K} \widetilde{\phi}(y^g) = \widetilde{\phi}(y^g) \end{aligned} \tag{5}$$

where $\widetilde{r}(y^s)$ and $\widetilde{b}$ denote the "prefect" reward and baseline respectively. However, the evaluation score ϕ that we use in practice is usually based on n-gram matching (e.g., CIDEr) which correlates well with $\widetilde{\phi}$ but is not perfect. For any caption y and the groundtruth caption y_i^j, we introduce an additional random noise ϵ_j to describe such noisy relation between $\phi(y, y_i^j)$ and $\widetilde{\phi}(y)$:

$$\epsilon_j = \phi(y, y_i^j) - \widetilde{\phi}(y) \tag{6}$$

Thus we can measure the difference $\mathcal{E}$ between the baseline b in practice and the perfect baseline $\widetilde{b}$:

$$\mathcal{E} = b - \widetilde{b} = \frac{1}{K} \sum_{j=1}^{K} (\phi(y^g, y_i^j) - \widetilde{\phi}(y^g)) = \frac{1}{K} \sum_{j=1}^{K} \epsilon_j \tag{7}$$

For simplicity, we assume that the random noises $\epsilon_{j=\{1,\ldots,K\}}$ are i.i.d. with variance σ^2, and the variance of the difference $\mathcal{E}$ can be calculated as:

$$\mathrm{Var}(\mathcal{E}) = \frac{1}{K} \sigma^2 \tag{8}$$

According to eq (8), when K becomes smaller (i.e., fewer groundtruth captions per input), the variance of the difference $\mathcal{E}$ is amplified, which means the estimation of the perfect baseline $\widetilde{b}$ becomes less accurate. Similar argument could be also applied to the reward. Training with less accurate reward and baseline leads to the performance degradation of models with SC loss as shown in figure 1.

4 Reducing Variance by Generalized Pairwise Comparison

We propose to reduce the variance $\mathrm{Var}(\mathcal{E})$ in eq (8) and make the estimation of the prefect baseline $\widetilde{b}$ more accurate by generalized pairwise comparison (GPC). **Using re-1 as an example**, we have only one groundtruth caption y_i for input x_i, which results in only one call of the evaluation score ϕ in the calculation of the baseline. GPC enables us to add more independent calls of ϕ even in the re-1 setting to reduce the noise in the baseline and still keeps the difference between the reward and the baseline meaningful, which is the merit of SC loss.

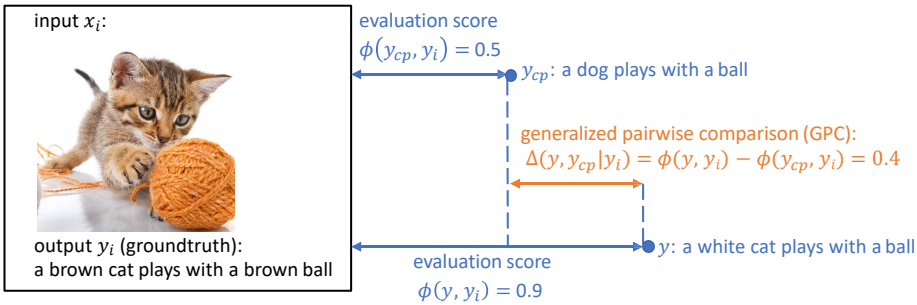


Figure 2: Illustration of generalized pairwise comparison (GPC)

4.1 Generalized Pairwise Comparison

The generalized pairwise comparison (GPC) $\Delta(y, y_{cp}|y_i)$ is defined on the triplet (y, y_{cp}, y_i), where y_i is the groundtruth caption used as the reference, and y, y_{cp} could be any two captions, e.g., captions sampled from the model or captions associated with other input in the dataset. As illustrated in figure 2, the generalized pairwise comparison measures the difference between the evaluation scores of $\phi(y, y_i)$ and $\phi(y_{cp}, y_i)$ with the reference caption y_i as follows:

$$\Delta(y, y_{cp}|y_i) = \phi(y, y_i) - \phi(y_{cp}, y_i) \tag{9}$$

We show that SC is a special case of GPC if we substitute y and y_{cp} by the sampled caption y^s and the greedily decoded caption y^g respectively:

$$r(y^s) - b = \phi(y^s, y_i) - \phi(y^g, y_i) = \Delta(y^s, y^g|y_i) \tag{10}$$

In GPC view, the meaning of $r(y^s) - b$ is that how much better the sampled caption y^s is compared to the greedily decoded caption y^g on the evaluation score ϕ with the reference y_i. Actually, we could substitute y_{cp} with any other caption in GPC and the corresponding meaning is the comparison of the sampled caption y^s with any other caption, including the greedily decoded caption. Furthermore, we could combine multiple generalized pairwise comparisons instead of only using single one.

For m multiple GPCs, we average them and get:

$$\begin{aligned}\frac{1}{m}\sum_{n=1}^{m}\Delta(y^s, y^n_{cp}|y_i) &= \phi(y^s, y_i) - \frac{1}{m}\sum_{n=1}^{m}\phi(y^n_{cp}, y_i) \\ &= r(y^s) - b_{GPC} \\ r(y^s) &= \phi(y^s, y_i) \\ b_{GPC} &= \frac{1}{m}\sum_{n=1}^{m}\phi(y^n_{cp}, y_i)\end{aligned} \tag{11}$$

where y^n_{cp} denotes the n-th in the m captions for comparison. As a result, we get the same reward as that in SC but a different baseline b_{GPC}. The variance of the difference $\mathcal{E}_{GPC}$ between the perfect baseline $\widetilde{b} = \frac{1}{m}\sum_{n=1}^{m}\widetilde{\phi}(y^n_{cp})$ and the new baseline b_{GPC} is related to m, which changes by our choice, rather than K, which is the fixed number of groundtruth captions:

$$\mathrm{Var}(\mathcal{E}_{GPC}) = \mathrm{Var}(b_{GPC} - \widetilde{b}) = \frac{1}{m}\sigma^2 \tag{12}$$

Thus, comparing eq (12) with eq (8), we could reduce the noise in the baseline even in the re-1 setting ($K = 1$) by introducing more generalized pairwise comparisons ($\frac{1}{m}\sigma^2 < \sigma^2$ when $m > 1$). We leave making the estimation of the reward more accurate in the future work.

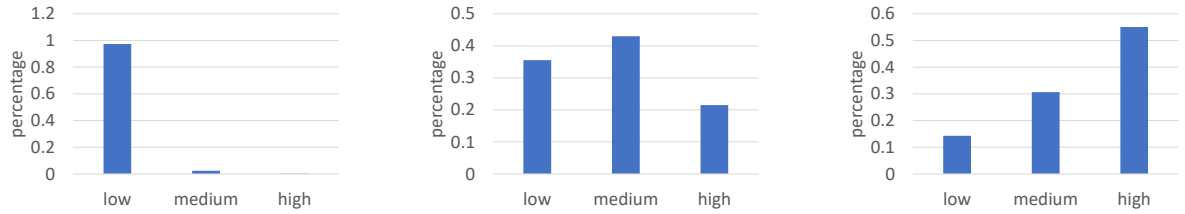
4.2 Learning with GPC by Mixed Distribution Sampling

Generalizing from re-1 setting to re-K setting, we finally obtain the objective function of multiple GPCs on the entire dataset as follows:

$$\begin{aligned}L = -\sum_{i=1}^{N}\frac{1}{K}\sum_{j=1}^{K}\big(\phi(y^s, y^j_i) - \\ \frac{1}{m}\sum_{n=1}^{m}\phi(y^n_{cp}, y^j_i)\big)\log p(y^s|x_i)\end{aligned} \tag{13}$$

It could be optimized by the standard policy gradient algorithm REINFORCE if we have m different y_{cp} for each x_i. Note that we assume that the random noise ϵ_j associated with $\phi(y^n_{cp}, y^j_i)$ is independent in the variance reduction analysis. To ensure this, we need to sample m captions y_{cp} independently to cover the whole caption space, which is almost intractable.

Mixed Distribution Sampling Instead of trying to cover the whole caption space, we turn to cover the whole value range of evaluation score ϕ. First, We categorize the evaluation scores into three groups, low, medium and high based on the pre-defined thresholds. Then we sample from a



(a) CIDEr score distribution of captions sampled from the whole dataset

(b) CIDEr score distribution of captions sampled from the model

(c) CIDEr score distribution of captions greedily decoded from the model

Figure 3: Mixed distribution sampling: the x-axis refers to the three groups and y-axis refers to proportion of each group. CIDEr thresholds for low, medium and high groups are set to 0.2 and 0.7 for illustration purpose.

Algorithm 1 Learning with GPC by Mixed Distribution Sampling

for epoch in [0, M) **do** ▷ standard warm-up stage before reinforcement learning for captioning tasks
  train by cross-entropy loss
end for
for epoch in [M, N) **do** ▷ reinforcement learning with GPC loss
  for each instance x_i **do**
    sample y^s from current model
    $y_{cp}^1, \ldots, y_{cp}^m = \text{MIXDIS\_SAMPLING}(m_l, m_m, m_h)$, where $m = m_l + m_m + m_h$
    calculate loss L by eq (13)
    update model by gradient $-\frac{1}{K}\sum_{j=1}^{K}\left(\phi(y^s, y_i^j) - \frac{1}{m}\sum_{n=1}^{m}\phi(y_{cp}^n, y_i^j)\right)\nabla \log p(y^s|x_i)$
  end for
end for
procedure MIXDIS_SAMPLING(m_l, m_m, m_h) ▷ mixed distribution sampling from three distributions $\mathcal{D}^l, \mathcal{D}^m, \mathcal{D}^h$
  for $\mathcal{D}^l$, sample m_l captions $y^1, \ldots, y^{m_l}$ from the captions in the whole dataset
  for $\mathcal{D}^m$, sample m_m captions $y^{m_l+1}, \ldots, y^{m_l+m_m}$ from the current model
  if $m_h == 1$ **then**
    for $\mathcal{D}^h$, greedily decode caption y^g from the current model
    return $y^1, \ldots, y^{m_l+m_m}, y^g$
  else
    return $y^1, \ldots, y^{m_l+m_m}$
  end if
end procedure

mixture of different distributions which concentrate on different groups of evaluation score:

low-score distribution $\mathcal{D}^l$. We randomly sample m_l captions from the dataset. The score of such samples is usually low from the statistics shown in figure 3a.

medium-score distribution $\mathcal{D}^m$. We randomly sample m_m captions from the model. The score of such samples is usually medium based on the statistics shown in figure 3b.

high-score distribution $\mathcal{D}^h$. We use greedy decoding to generate a caption from the model. The score is usually high based on the statistics shown in figure 3c. Since at most one caption could be greedily decoded, the number of captions sampled from this distribution, m_h is 1 or 0. $m_h = 0$ means that we do not sample from $\mathcal{D}^h$.

Finally, we combine the sampled results into the m captions y_{cp} in eq (13). In the experiment section, we will show empirically that the captions sampled by the procedure of mixed distribution sampling turns out to be a quite good approximation of the whole caption space.

Following the standard procedure of reinforcement learning in captioning tasks, we first run a warm-up stage of training with XE loss. Then we switch to the reinforcement learning stage with objective function defined in eq (13). In each evaluation of the objective function, we need to run the sub-procedure MIXDIS_SAMPING (mixed distribution sampling) to get $y_{cp}^1, \ldots, y_{cp}^m$. The entire learning algorithm is summarized in algorithm 1.

5 Experiments

5.1 Experiment Setup

We use the image caption dataset MSCOCO (Chen et al., 2015) and the video caption dataset TGIF (Li et al., 2016). **MSCOCO**, one of the largest image caption datasets, contains more than 120K images crawled from Flickr. Each image is annotated with 5 reference captions. We use the public split (Karpathy and Li, 2015) to evaluate our model as most image caption researches (Xu et al., 2015;

Rennie et al., 2017) are evaluated on this split. We follow the practice in section 3 to synthesize 4 language resource settings that are different in the caption number per image. Combined with the original setting, we have 5 settings in total to do extensive evaluations. **TGIF** is one of the largest video caption datasets in terms of video numbers, which contains $100K$ animated GIFs collected from Tumblr and $120K$ caption sentences. We use the official split (Li et al., 2016) to evaluate the generation task. For videos in the training and validation set, it contains one caption per video. For videos in the test set, it contains three captions per video.

For image, we use Resnet101 (He et al., 2016) which was pre-trained on ImageNet and apply spatial mean pooling to generate a feature vector of dimension 2048. We resize the larger side of the image to 450. For video, we use I3D (Carreira and Zisserman, 2017) pre-trained on Kinetics400 and apply spatial-temporal mean pooling to generate a feature vector of dimension 1024. We resize the larger side of the video to 224.

We use the vanilla encoder-decoder architecture for simplicity. For the encoder, we use a full connection layer to reduce the dimension of input feature to 512. For the decoder, we use standard RNN with LSTM cell. The dimension of hidden unit is set to 512. In step 0 the hidden state is initialized by the output of the encoder. We use ADAM (Kingma and Ba, 2014) optimizer with batch size 64 and set the learning rate to 10^{-5} to run algorithm 1. The model is selected based on CIDEr score on the validation set. In MIXDIS_SAMPLING procedure, we set m_l, m_m, m_h to $2, 2, 1$ across different language resource settings and datasets. Detailed ablation study of tuning m_l, m_m, m_h will be discussed in section 5.2.1.

5.2 Evaluation on Synthesized Language Resource Settings of MSCOCO

We compare the performance of the proposed generalized pairwise comparison (GPC) loss to cross-entropy (XE) loss and self-critical (SC) loss under different resource settings. The construction of different language resource settings, re-1, re-2, re-3, re-4 and full, are the same as those in section 3.

As shown in table 1, we see that the performance improvement of GPC loss over XE and SC loss is very significant in re-1 and re-2 settings. It improves over XE on CIDEr by 11.5 absolute points (13.6% relatively) in re-1 and 16.1 absolute points (18.0% relatively) in re-2. Compared to SC loss, it improves on CIDEr by 7.8 absolute points (8.8% relatively) in re-1 and 9.4 absolute points (9.8% relatively) in re-2. Furthermore, the model trained by GPC loss converges quickly to the full setting performance on most metrics (BLEU4, METEOR) with very few captions per instance such as re-2 setting. The improvement is not significant in the full setting as the variance of baseline is already very small given 5 groundtruth captions per input. This aligns well with our motivation and variance reduction analysis in section 3.

Table 1: Performance comparison on different language resource settings: SC* refers to the performance reported in (Rennie et al., 2017)

model	setting	BLEU4	METEOR	CIDEr
XE	re-1	26.5	24.4	84.7
XE	re-2	27.7	25.0	89.6
XE	re-3	29.2	25.2	92.8
XE	re-4	28.7	25.4	94.0
XE	full	29.6	25.6	96.5
SC	re-1	27.9	23.6	88.4
SC	re-2	29.7	24.5	96.3
SC	re-3	30.7	24.7	100.5
SC	re-4	31.8	25.4	105.4
SC	full	33.1	26.0	110.4
SC*	full	31.9	25.5	106.3
GPC	re-1	30.0	24.8	96.2
GPC	re-2	31.9	25.4	105.7
GPC	re-3	32.1	25.6	106.8
GPC	re-4	32.3	25.5	109.6
GPC	full	33.2	25.8	110.8

We also list the performance of SC loss reported in the original work (Rennie et al., 2017) for reference in the table. SC loss implemented by us performs better than the one reported by (Rennie et al., 2017) and we attribute the difference to the preprocessing as we resize the extracted feature of images to a larger size 450[1]. The comparison to the results in the original paper shows that the SC model implemented by us is a strong baseline. Thus, we can conclude that the model trained by GPC loss works in all resource-efficient levels.

We further compare the labeling resource required by different methods when the performance is fixed. As highlighted in red, GPC loss reaches almost the same performance of XE with only $1/5$ of training data as re-1 setting has 1 caption per image and full setting has 5 captions per image. Furthermore, GPC loss reaches almost the same performance of SC loss with only $1/2$ of training

[1]This is related to both the receptive field size of the CNN and the size of object in the image, which is out of the scope of this paper.

Table 2: Ablation study of each distribution in MIXDIS_SAMPLING procedure from algorithm 1 on re-1 setting

	m_l	m_m	m_h	BLEU4	METEOR	CIDEr
1	1	0	0	28.9	24.4	93.0
2	2	0	0	29.1	24.4	93.3
3	3	0	0	29.1	24.4	92.8
4	4	0	0	29.1	24.4	93.2
5	5	0	0	**29.5**	**24.6**	**94.0**
6	0	1	0	29.2	24.4	93.6
7	0	2	0	29.3	24.5	93.8
8	0	3	0	**29.4**	**24.6**	94.2
9	0	4	0	29.3	**24.6**	**94.4**
10	0	5	0	29.1	24.5	94.1
11	0	0	1	27.9	23.6	88.4

data as highlighted in blue. This shows that GPC loss is more resource efficient and works particularly well with very few captions per input.

5.2.1 Ablation Study on Mixed Distribution Sampling

We first study only sampling from one distribution. As shown in table 2, the first, second, and third blocks correspond to only sampling from $\mathcal{D}^l$, $\mathcal{D}^m$ and $\mathcal{D}^h$ respectively. We see that in general the performance improves mildly when we sample more captions from the distribution. Sampling from only the distribution $\mathcal{D}^h$ actually degenerates to self-critical loss (Rennie et al., 2017) based on the MIXDIS_SAMPLING procedure in algorithm 1. Comparing different distributions with m fixed to 1, we see that sampling from $\mathcal{D}^l$ (row *1*) and $\mathcal{D}^m$ (row *6*) both outperforms that using only greedily decoded samples $\mathcal{D}^h$ (row *11*) on CIDEr by 4.6 and 5.2 respectively. This shows that the proposed GPC loss is not only a generalization of the self-critical loss but also performs much better for variance reduction of baseline in different language resource settings.

We also study sampling from the combination of different distributions. As shown in table 3, we set the total number of samples m to 5. Among all the distribution combinations (altogether 4) under the same quota $m = 5$, we see that sampling from all the three distributions ($m_l = 2$, $m_m = 2$, $m_h = 1$) performs best on all the three metrics. This shows that covering the whole score range of ϕ is beneficial for the variance reduction. Furthermore, the setting $m_l = 2$, $m_m = 2$, $m_h = 1$ turns out to be a good and stable approximation of the whole caption space across different language resource settings and datasets.

Table 3: Ablation study of distribution combination in MIXDIS_SAMPLING procedure from algorithm 1 on re-1 setting

m_l	m_m	m_h	BLEU4	METEOR	CIDEr
2	3	0	29.5	24.6	94.1
4	0	1	29.5	24.5	94.0
0	4	1	29.4	24.5	94.4
2	2	1	**30.0**	**24.8**	**96.2**

Table 4: Evaluation on TGIF dataset

method	BLEU4	METEOR	CIDEr
Official	12.7	16.7	31.6
Show-adapt	11.8	16.2	29.8
XE	15.7	18.4	45.6
SC	15.7	18.5	49.8
GPC	**16.1**	**19.0**	**52.1**

5.3 Evaluation on TGIF

To show the general resource-efficiency of GPC loss, we further run experiments on TGIF. TGIF is a GIF dataset in which only one caption per input is provided for training. It is different from the above experiments from two aspects. First, it is a video dataset. Second, the language resource setting of one labeled caption per input is not synthesized. From table 4, we see that GPC loss performs significantly better than both XE and SC loss, i.e., boosting 6.5 points (14.3% relatively) and 2.3 points (4.6% relatively) on CIDEr over XE and SC loss respectively. It is interesting to compare the performance boost of SC loss over XE and the performance boost of GPC loss over SC loss. SC loss achieves almost no boost on BLEU4 and METEOR over XE loss. But GPC loss boosts all metrics over SC loss. This shows that GPC loss is effective on the real-world language resource efficient setting with one labeled caption per input.

6 Conclusion

In this paper, we propose the language resource efficient concept for captioning tasks in terms of the number of captions per input. Our analysis shows that in captioning tasks, fewer captions per input lead to larger noise in estimating the reward and baseline for self-critical loss of reinforcement learning. We propose to reduce the noise in the baseline by multiple generalized pairwise comparisons, which results in the GPC loss. Experimental results show that our proposed model is efficient on language resource and achieves similar performance with the state-of-the-art models by using only half of the captions per input. Furthermore,

the proposed model performs significantly better than the state-of-the-art models on a video caption dataset that has only one labeled caption per input.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 61772535) and Beijing Natural Science Foundation (No. 4192028). We also sincerely appreciate the reviewers for their valuable comments and suggestions.

References

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

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

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

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *NIPS 2017*.

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

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Linghui Li, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. 2017. Gla: Global-local attention for image description. *IEEE Transactions on Multimedia*.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *ICCV*.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *CVPR*.

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

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

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

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

Gunnar A. Sigurdsson, Gul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

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

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

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