

Diverse Image Captioning via GroupTalk

Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, Yueting Zhuang

College of Computer Science, Zhejiang University, China

{zhuhaow, wufei, luwm, junx, xilizju, ztzhang, yzhuang}@zju.edu.cn

Abstract

Generally speaking, different persons tend to describe images from various aspects due to their individually subjective perception. As a result, generating the appropriate descriptions of images with both diversity and high quality is of great importance. In this paper, we propose a framework called *GroupTalk* to learn multiple image caption distributions simultaneously and effectively mimic the diversity of the image captions written by human beings. In particular, a novel iterative update strategy is proposed to separate training sentence samples into groups and learn their distributions at the same time. Furthermore, we introduce an efficient classifier to solve the problem brought about by the non-linear and discontinuous nature of language distributions which will impair performance. Experiments on several benchmark datasets show that *GroupTalk* naturally diversifies the generated captions of each image without sacrificing the accuracy.

1 Introduction

Generating descriptions for images automatically is a fundamental problem in machine learning which involves both computer vision and natural language processing. There is much recent advance focusing on developing models that generate image captions, and the performances have been significantly improved thanks to the development of deep learning. Among them, a recent trend is to combine neural networks to build up an end-to-end model that extracts visual features and generates sentences, these models are trained based on learning a model to capture the conditional distribution of the sentences given the image features [Karpathy and Li, 2015; Vinyals *et al.*, 2014; Donahue *et al.*, 2014; Mao *et al.*, 2014; Jin *et al.*, 2015; Xu *et al.*, 2015]. This framework is demonstrated to be easy to train and performs quite well. However, the generated image-caption result is vastly constrained (e.g. sentences describing the same image are almost identical) if compared to the rich-semantic descriptions written by human beings.

Corresponding authors: Fei Wu, Weiming Lu, Jun Xiao and Yueting Zhuang.

In general, persons tend to describe images in their preferred ways with various words, phrases, structures and perspectives, even a same person may give the same image totally different descriptions if asked twice. However, the sentences generated by the existing models are usually following a similar pattern since these models are deliberately trained to capture the most common routine of the training sentences to maximize the likelihood function for all training samples. Given an image, it is more useful and practical that we are able to generate several different sentences describing it from different perspectives in various ways.

Usually, each image in the image-caption datasets widely used comes with several sentences collected from a crowdsourcing platform [Young *et al.*, 2014; Chen *et al.*, 2015], which means the sentences in these datasets are written by a wide variety of people. Thus the sentences in the datasets naturally reflect the diverse nature of human generated captions. Since we already have the data, the only question left is how to learn to generate sentences with diversity just as in both the reality and the training sets.

In this paper, we seek to explore the approach to generate much more diverse image captions compared to previous methods by training several language models simultaneously, we refer to these language models as **describing models**¹ where each model shares the same structure but effectively captures a different distribution. These describing models are trained to mimic diverse descriptions generated by human beings. We call this framework **GroupTalk** which acts as if a group of people are describing the same image at the same time with different preferences.

The main difficulty in this framework is how to cluster the training sentence samples into several groups for each describing model based on some type of similarity since no group information is available in training set; this seemingly easy task must be dealt with elaborately as we will show in the following sections. Furthermore, as shown in [Hessel *et*

¹Usually, in an image related context, a generating model is a model representing the joint distribution of $P(x, y, \dots)$ where x is the representation of the image and y, \dots are the label and/or some other relevant variables. To obviate misunderstanding, we here use the describing model instead of generating model (since we are not interested in the joint distribution) to refer to the model generating sentences given images which can be formulated as $P(s|x)$ where s is the sentence.

al., 2015], training with fewer sentences is a even more severe obstacle than training with fewer images; we must handle it carefully to prevent the accuracy deterioration of each describing model as we do not train each of them with the whole dataset as the traditional ways do.

The main contributions can be summarized as:

- We deploy a framework called GroupTalk to deal with aforementioned problems, which trains multiple describing models simultaneously to generate diverse and high quality image descriptions.
- We propose a classification-by-generation scheme, which employs a classifier to separate the training samples into groups which (1) prevents overfitting; (2) allows the sharing of general sentences among different groups to improve the performance; (3) is highly efficient to classify a training sample.

The experiments on several benchmark datasets show that our framework can generate much more diverse captions without sacrificing the accuracy.

2 Related Work

2.1 Image Captioning

Most of current image captioning approaches bridge a deep convolutional network and a language network together to construct an end-to-end model to generate image captions from images. Karpathy et al. [Karpathy and Li, 2015] proposed a model composed of a VGG Net for image feature extraction and a basic recurrent neural network (RNN) for caption generation. Vinyals et al. [Vinyals et al., 2014] deployed a more powerful Long Short Term Memory (LSTM) model. Some other work also focused on applying more powerful language model, such as multi-layer LSTM [Donahue et al., 2014], Factored 3-way Log-bilinear model [Kiros et al., 2014]. Some other studies explored the structure of the whole network to effectively model the correlation between image and its corresponding sentence, such as multimodal Recurrent Neural Networks [Mao et al., 2014] and its extension [Mao et al., 2015]. Recently, it is discovered that it helps to input an attentional visual zone instead of the whole image in each tick of the language sequential network by learning a separated attention model [Jin et al., 2015; Xu et al., 2015].

Some other studies designed a pipeline instead of an end-to-end network for image caption generation. Fang et al. [Fang et al., 2014] detected candidate words by multiple-instance learning first and then learned a statistical model for sentence generation and finally reranked the output sentences based on semantic similarity. In [Wu et al., 2015], they trained several SVM classifiers for some visual attributes and took the outputs of these classifiers as inputs for the LSTM which generates image descriptions. Lebre et al. [Lebre et al., 2015] used phrases instead of words as basic sentence building blocks and concatenated candidate phrases by a constrained language model. In [Jiang et al., 2015], Jiang et al. explored the correlations among images and sentences by enhancing the local and global semantic alignments between them.

All above methods have been mainly focused on improving the accuracy of the descriptions with respect to the groundtruth but none of them pay special attention to the diversity of the generated sentences.

2.2 Unsupervised Sentence Clustering

In order to cluster the training sentences into several groups, we explore unsupervised clustering methods. Though various methods exist, it is hard to define a proper grouping gauge that the grouping result is advantageous in our scenario.

There are some general unsupervised clustering algorithms such as k-means and Gaussian Mixture Model (GMM). However, most of them are designed for much simpler distributions which do not fit well with the complex distribution of natural language.

Some other models such as Latent Dirichlet Allocation (LDA) [Blei et al., 2003] and its extensions can cluster sentences in a more sophisticated way. However, various previous research found that LDA does not work well in short corpus [Hong and Davison, 2010]. Furthermore, sentences clustered by LDA are actually clustered in terms of some latent topics; if we train one model based on some group with a latent topic which is actually sport-related, then this describing model will produce sport-related captions no matter the input images are sport-relevant or not. Moreover, the describing models are not necessarily capable of representing the distributions of the unsupervised clustering results.

The proposed model in this paper is similar to the GMM, where the gaussian model is replaced with a language model and the target likelihood function is modified to deal with the high non-linearly and discontinuity of the language model. Different from GMM, our model is not proposed for clustering, we focus more on the generating part. Furthermore, we propose a classifier to replace the E-step in the EM algorithm to prevent the optimization phase trapped in highly unsatisfactory local minimum quickly.

3 Algorithm

Assuming we are given a set of images $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$, where each image I_i comes with several sentences describing it as $\mathcal{S}(I_i) = \{s_1, s_2, \dots, s_m\}$, and each s_j consists of a sequence of words. Our goal is to generate several sentences describing a new image I in a diverse manner.

Traditionally, we train a sequential model (usually some type of recurrent neural network) that maximizes the likelihood function of

$$P(\mathcal{S}|\mathcal{I}) = \prod_{I_i \in \mathcal{I}} \prod_{s_j \in \mathcal{S}(I_i)} P(s_j|I_i). \quad (1)$$

which assumes there is only one model that captures the conditional distribution of descriptions with respect to the given images. Given test image I_t , we can generate image captions via $\arg \max_s P(s|I_t)$. Since human generated captions are highly diverse, there should be several sentences with high generating probability that vary significantly from each other.

From a model capacity perspective, though theoretically possible, but merely using one model to capture the intricate distribution of various human descriptions requires a powerful and complex model which is hard to train and demands

a lot of training samples to avoid overfitting. Moreover, in practice, due to the discontinuous nature of language distribution, finding all the sentences with high probability is not computationally feasible; usually, beam search strategy is used to find several candidates [Karpathy and Li, 2015; Vinyals *et al.*, 2014]. While it is demonstrated to be effective and performs well in accuracy, the candidates generated via beam search strategy are all very similar. Another generating strategy is to sample the next word in every tick, however, this strategy impairs the accuracy significantly comparing to beam search.

In a word, from the perspectives on model capacity and computational complexity, it is difficult to use one model to generate sentences with diversity. It is natural to assume that human beings describe images with various preferences which leads to several conditional distributions; different persons may describe images in different ways and each distribution captures some specific one. By this way, we can generate diverse descriptions with several models efficiently.

We assume that there is K intrinsic conditional distributions of sentences given one image.² The set of sentences generated from the k th distribution is denoted as \mathcal{S}_k , and the set of sentences from the k th distribution describing image I_i is denoted as $\mathcal{S}_k(I_i)$. Every sentence comes from some model, so we have: $\bigcup \mathcal{S}_k(I_i) = \mathcal{S}_i$. However, there are general descriptions that may come with several models, i.e., $\mathcal{S}_k \cap \mathcal{S}_l$ may not be empty.

Now we can separate the original distribution to several distributions by:

$$P(\mathcal{S}|\mathcal{I}) = \prod_{I_i \in \mathcal{I}} \prod_{s_j \in \mathcal{S}(I_i)} \prod_{k=1}^K P(s_j \in \mathcal{S}_k | I_i) P(s_j | I_i, s_j \in \mathcal{S}_k) \quad (2)$$

where each $P(s|I, s \in \mathcal{S}_k)$ is modeled by the k th describing model.

Though eq. (1) and eq. (2) are theoretically equivalent, in practice, the model we use has a representing capacity that is far below the requirements to capture the genuine distribution of human language, so using multiple distributions allow us to relax the requirements of the representing power of one model. Training several models with different distributions allows us to generate diverse captions efficiently with beam search.

In this paper, we use a moderate complex but still powerful language model, LSTM, for sentence generation³. The LSTM generator can be formulated as:

$$\begin{aligned} i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}), \\ f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \\ o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \\ m_t &= o_t \odot c_t, \quad p_{t+1} = \text{Softmax}(m_t) \\ x_0 &= W_{if}I_{feat}, \quad x_t = X_{p_t} \end{aligned}$$

where I_{feat} is the feature of input image I extracted by convolutional network, X is the encoding matrix of input words and p is the index of the word, W s are the connection weights in the network, σ is the sigmoid and h is the hyperbolic tangent function. We extract features of images with VGG Net [Simonyan and Zisserman, 2014] trained on image dataset in ILSVRC-2014 image classification competition. Please refer to [Hochreiter and Schmidhuber, 1997] for more details and explanations of LSTM.

Generating captions with GroupTalk is illustrated in Figure 1 with reference to the **generating path**: image feature is first extracted by CNN then feed into different LSTMs to generate diverse descriptions.

However, given the set of sentences \mathcal{S} , the group information $P(s \in \mathcal{S}_k | I)$ is not available, which makes it impossible to directly train our model by eq. (2).

In order to compute $P(s \in \mathcal{S}_k | I)$, we need to cluster sentences. Equation (2) is quite similar to GMM which can be solved with EM algorithm [Dempster *et al.*, 1977]. However, a language neural network (e.g., LSTM we used) represents a highly non-linear and discontinuous distribution that will easily overfit to several sentences.

Considering if we directly use the E-step in EM algorithm to compute $P(s \in \mathcal{S}_k | I)$ by

$$P(s \in \mathcal{S}_k | I) = \frac{P(s|I, s \in \mathcal{S}_k)}{\sum_l P(s|I, s \in \mathcal{S}_l)}. \quad (3)$$

This will suffer from two major problems:

1. We need to forward this sentence in all K LSTMs with full process to compute $P(s|I, s \in \mathcal{S}_l)$, which is inefficient. In every tick, we have to compute the probability of next word given the current one, which requires us to compute the matrix multiplication involving an $H * N$ matrix, where N is the dictionary size and H is the size of the hidden units. This is the most expensive part of the forward phase which takes more than 70% computing time, and the computation overload will grow proportionally with respect to both the number of the describing models and the size of the dictionary.

2. As the distribution represented by LSTM is highly non-linear and discontinuous, these describing models can overfit to some local minimum very soon that only generates some specific sentences with very high probability for any input images (i.e., for some sentences, $P(s \in \mathcal{S}_k | I)$ is very large for a specific model for any input images and almost zero for some other sentences), which means they converge to an unsatisfying solution quickly. Consider an extreme case where we have two models, where an impenetrable one only generates sentence ‘‘a dog is chasing a man’’ for any input images, and another generating sensible sentences based on the visual information, then it’s very obvious that only the sentences very

²Method to determine K is discussed in the Speeding Up section.

³It does not really matter what kind of recurrent neural network is used here as the language sequential model since what we proposed in this paper is a framework to generate diverse sentences based on any kinds of such model. We choose LSTM due to its simplicity and the satisfactory results.

similar to “a dog is chasing a man” will have a high probability to be generated by this model, and further training with these sentences makes this model more confident about the correctness of itself. This problem will become more serious when we enlarge the number of models.

To deal with aforementioned problems, we need a way to compute $P(s \in \mathcal{S}_k|I)$ in an efficient way while weaken the accuracy of it in order to have subtle distinctions not yield large probability difference.

In this paper, we devise several important methods to overcome these problems:

1. We explicitly design a weak classifier to compute $P(s \in \mathcal{S}_k|I)$ that avoids overfitting to the distributions of the describing models. The sentence classifier is a LSTM without the word decoding part plus a softmax classifier. Given one image and its corresponding description, we forward them in the LSTM and obtain the hidden representation (m_t). Then we input these features into the softmax classifier. In practice, we can use the hidden state of the last tick (m_{last}) or the average of all ticks ($\frac{1}{T} \sum_T m_t$) as input into the softmax classifier. We find they have similar performance while the averaging way is faster to converge, so in experiments, we use the latter. As the number of hidden units represents the power of the language model and increasing the size of hidden units brings corresponding performance improvements, we purposely set the size of hidden units of the classifier to a small value compared to the describing model, which makes the classifier can only distinguish the distinct differences between the sentences generated by the describing models. Moreover, more describing models makes it harder for the classifier to give an accurate result, which further prevents some specific models quickly overfitting to a few sentences.

2. The LSTM classifier is much more efficient. No matter how many describing models we have, we only need to forward the input sentence in the classifier once, and we do not need to compute the decoding part which allow us to skip the most expensive part in forwarding an LSTM. So the classification speed is constant with respect to either the number of the describing models or the size of the dictionary. Further, training the classifier could be done with several thousands of training samples in several epochs efficiently.

3. We use sampling instead of classification when selecting training samples for each model, i.e. each sentence has probability $P(s \in \mathcal{S}_k|I)$ to be a training sample for the k th model in one iteration. This allows us to use one sentence to train several different models; these sentences can be seen as general sentences with no specific preferences. In this way, we train each model with as many sentences as possible to improve the accuracy.

3.1 Optimization of GroupTalk

In GroupTalk, we propose a classification-by-generation strategy to train our models, the overall framework of GroupTalk is illustrated in Figure 1. Formally, in each iteration (usually hundreds of batches), we train a classifier based on the sentences generated by different describing models following **generating path** to compute $P(s \in \mathcal{S}_k|I)$, and then use this classifier to separate the sentences in the training set into different groups for each model by sampling, which is

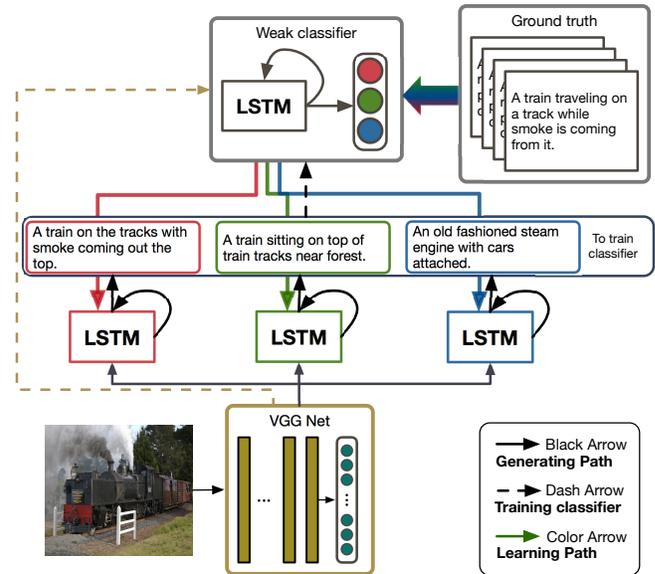


Figure 1: The intuitive structure of GroupTalk. During training, a classifier is trained based on sentences generated by several LSTM models (we use three describing models here) following the generating path, then the sentences in training set are classified by this classifier as training samples for respective models as in the learning path; and these steps are repeated until convergence.

shown as the **learning path**. This allows us to learn the group information based on the distribution of each describing model represented by the sentences they generated.

3.2 Speeding Up with Pre-training

Training several models following the previous algorithm from scratch may be slow, even with the efficient classifier; usually, training two describing models requires at least double training time. However, when the models have not learnt the distribution of natural language the trained classifier is not meaningful, which suggests that we do not need a classifier in the early stage. As widely applied in convolutional network, usually we use a model well trained for some related task as a start point and further fine-tune it to improve the performance for our target task to speed up the training. Here, we borrow this idea and use the model trained following the traditional way based on eq. (1) as a start point denoted as **base model**. However, if we use weights of this base model as the initial weights for all describing models, the generated sentences will be same for all models, thus the classifier can not learn any information. Therefore, we have to make some modifications to the weights of the base model. There are many possible ways, such as randomly selecting a subset of training samples to train the model a few more epochs or altering some weights of the network. Here we use a more sophisticated way, we select a set of sentences with low generating probability with reference to the base model as training samples. These samples can be seen as the non-general sentences that we need our models to learn to increase diversity,

Table 1: The image captioning performance comparisons in terms of accuracy.

	Flickr8K				Flickr30K				MS COCO			
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
Nearest Neighbor	–	–	–	–	–	–	–	–	48.0	28.1	16.6	10.0
Mao et al.	58	28	23	–	55	24	20	–	–	–	–	–
LRCN	–	–	–	–	58.8	39.1	25.1	16.5	62.8	44.2	30.4	–
Chen and Zitnick	–	–	–	14.1	–	–	–	12.6	–	–	–	19.0
NeuralTalk	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0
Base model (S)	45.8	22.6	11.5	6.5	43.4	19.9	9.3	4.6	51.3	28.1	15.7	8.9
Base model (B)	62.6	40.5	26.9	18.0	63.5	39.9	26.0	17.0	68.1	47.4	33.9	24.8
GroupTalk												
Model 1	61.5	39.7	26.1	17.1	64.3	40.0	25.9	16.7	65.1	44.3	31.3	22.7
Model 2	62.7	39.5	25.5	16.6	59.6	37.4	24.1	15.8	68.5	46.9	32.8	23.5
Model 3	–	–	–	–	–	–	–	–	68.4	46.3	31.8	22.2

thus this gives us a reasonable start point. The further trained model is referred to as the seed model.

After we obtain the base model and the seed model, we take these two as two initial models and train them following the algorithm we discussed in previous section. After some epochs, when these two models both have learnt their specific distributions, we add another seed model to this model set and further train these three models. We can monitor the performance of the new model and the difference among these models to decide whether we can add another model based on the requirements of the specific application. This step is repeated until we get the desired number of describing models. We find in our experiments that this method is effective and we can train three models in less than a week on a large dataset (MS COCO) with an implementation in Python on CPU⁴.

4 Experiments

4.1 Datasets, Data Preprocessing and Network Settings

We run our experiments on Flickr8K [Hodosh *et al.*, 2013], Flickr30K [Hessel *et al.*, 2015] and MS COCO [Chen *et al.*, 2015] which are three benchmark datasets widely used in image caption related tasks. These three datasets contain 8,000, 31,000 and 123,000 images respectively and each image is annotated with 5 sentences on crowd-sourcing platform. The crowd-sourcing data naturally provided sentences with great diversity for us.

For Flickr8K and Flickr30K, we use 1,000 images for validation and 1,000 for testing, and for MS COCO, we use 5,000 for validation and 5,000 for testing.

Since many image-caption related studies have discovered that not back-propagating the gradients into the convolutional network yields better performance [Karpathy and Li, 2015; Vinyals *et al.*, 2014], we extract features of all images with VGG Net trained on the image dataset in ILSVRC-2014 image classification competition before our training.

⁴This can be further speed up several times by using more efficient implementations such as Lua with Torch based packages and training on GPU.

Following [Karpathy and Li, 2015], we convert all sentences to lowercase, discard non-alphanumeric characters and then filter words to those that occur at least 5 times in the training set, which results in 2,538, 7,414, and 8,791 words for Flickr8k, Flickr30K, and MS COCO datasets respectively.

There are many image caption generating framework, and most of them can be integrated with our framework, we decided to use one simple but still powerful model to generate sentences as our base model.

As shown in Figure 1, we use a convolutional network (VggNet) as image feature extractor and use a single-layer LSTM as a language generation model. Using a more powerful language model such as multi-layer LSTM, bidirectional LSTM or LSTM combined with attention model are demonstrated to give better performance based on the BLEU score which measure the similarity of sentences between generated sentences and reference sentences (groundtruth).

We set the size of hidden units in the LSTM of describing models to 512 and classifier to 128. We use two describing models (Model 1/2) for Flickr8K and Flickr30K and three describing models (Model 1/2/3) for MS COCO. The number of describing model is set based on the size and the degree of diversity of the training data.

4.2 Accuracy

Before we evaluate the diversity of the generated sentences, we need to validate that the sentences generated by these describing models of GroupTalk are reasonable and accurate, i.e., GroupTalk does not make the captions more variable by generating incorrect captions.

First we need to evaluate the performance of the **base model** (in the Speeding Up section which is trained following the traditional way). To show that the base model we choose bears both expression power and simplicity, we first compare it with other state-of-the-art end-to-end image captioning models with similar complexities:

- Nearest Neighbor: This is the baseline where a retrieval-based strategy is applied based on the similarity of image features.
- Mao et al. [Mao *et al.*, 2014]: In this model, image network and RNN are fused in a multimodal layer.

Table 2: Diversity evaluation. \downarrow/\uparrow depicts that a larger value indicates a less/more diverse result.

Dataset	Model	mB-1 \downarrow	mB-2 \downarrow	mB-3 \downarrow	mB-4 \downarrow	DIV-1 \uparrow	DIV-2 \uparrow	DIV-3 \uparrow	DIV-4 \uparrow
Flickr 8K	Base Model (B)	0.83	0.77	0.72	0.69	0.49	0.63	0.68	0.72
	GroupTalk	0.53	0.37	0.27	0.20	0.62	0.85	0.91	0.94
Flickr 30K	Base Model (B)	0.85	0.79	0.75	0.71	0.48	0.62	0.66	0.70
	GroupTalk	0.60	0.44	0.35	0.28	0.58	0.81	0.87	0.92
MS COCO	Base Model (B)	0.87	0.80	0.74	0.70	0.38	0.54	0.61	0.67
	GroupTalk	0.73	0.59	0.49	0.41	0.47	0.71	0.81	0.86

Table 3: Statistics of word usage preferences.

Model	#plane	#airplane	#jetliner	#jet
Model 1	0	14	64	10
Model 2	13	84	0	0
Model 3	72	25	0	0

- LRCN [Donahue *et al.*, 2014]: A model with a 2-layer factored LSTM with a structure similar to Mao *et al.*
 - Chen and Zitnick [Chen and Zitnick, 2014]: This model explores the bi-directional mapping between images and sentences with RNN.
 - Neurtalk [Karpathy and Li, 2015]: This model uses a structure similar to our base model except the LSTM is replaced with a basic RNN.
- The evaluation results are from [Karpathy and Li, 2015].

It is worth noting that the structure of our base model is quite similar to Google NIC [Vinyals *et al.*, 2014], in which the VGG Net in our base model is replaced with GoogLeNet, and model ensembling is applied⁵. Since none of the other methods use this convolutional network and this training technique, we do not compare with the original Google NIC in our paper for a fair comparison.

We generate sentences with base model following two different strategies, with beam search (B) or with sampling (S). While beam search is widely used in previous works, randomly sampling next word based on the probability distribution of the LSTM output clearly has the potential to promote the diversity.

We report BLEU- $\{1/2/3/4\}$ [Papineni *et al.*, 2001] which measures the similarity between the generated sentences and the groundtruth in table 1. Sentences are generated with beam size 5 (except for Base Model (S)).

From table 1, we can observe that base model with beam search (B) has a very satisfactory performance, even outperforms other models that are more complex and theoretically more powerful than LSTM (this may due to the more powerful convolutional network). We incline to the opinion that the model structure and training settings have a fundamental impact on the performance of the model. However, how the network settings and model designs influence the performance of the language model is beyond the scope of this paper. In

⁵Ensembling works best with small datasets but is much slower in both training and testing. We find on large dataset such as MS COCO, we can obtain similar performance without ensembling.

order to evaluate the effectiveness of our framework, we argue that the following comparison should be limited within the the base model and all describing models with the exactly same settings.

We can also observe that generating image captions with sampling (S) is not viable at all; the result shows that sampling makes the captions false of the image.

The accuracy of the describing models (Model 1/2/3) trained by GroupTalk following the Optimization and Speed Up section is also given in table 1.

It can be observed that the accuracy of our describing models is only slightly worse than the base model (even slightly better in BLEU-1), which shows the outputs of the multiple describing models are accurate and GroupTalk does not make the captions false of the image.

As shown in [Hessel *et al.*, 2015], training with more captions is more important to boost the accuracy than training with more images. With GroupTalk, we manage to train two or three models simultaneously using the same amount of training samples by sharing common sentences in multiple models.

4.3 Diversity

Metrics

In order to evaluate the diversity of the models, we use two metrics: 1) mBLEU based on BLEU proposed by ourselves and 2) degree of diversity (DIV) used in [Li *et al.*, 2015]

For simplicity, We denote $B_i(cand, ref)$ as the score of BLEU- i where the candidate set is $cand$ and reference set is ref . Then the mBLEU- i of sentence set $S(I_j)$ containing generated sentences describing the j th image in testing set is defined as:

$$mB_i(S(I_j)) = \frac{1}{|S(I_j)|} \sum_{s \in S(I_j)} B_i(\{s\}, S(I_j) \setminus \{s\}). \quad (4)$$

DIV- n is computed as the proportion of unique n -grams in all n -grams in set $S(I_j)$.

We report the mean of mBLEU- $\{1,2,3,4\}$ and DIV- $\{1,2,3,4\}$ over all images in testing set in Table 2. For each image, we generate one sentence with each model in GroupTalk as $S(I)$ (GroupTalk in Table 2) and the same amount of sentences by base model with beam search as $S(I)$ (Base Model (B) in Table 2)⁶.

⁶Sampling (S) is not viable as it generates captions with low quality which is shown in previous section.

Table 4: The demonstration of generated sentences by base model and GroupTalk on MS COCO.

	Base Model with beam search size 3:	GroupTalk
	<p>A large jetliner sitting on top of an airport tarmac.</p> <p>A large jetliner sitting on top of an airport runway.</p> <p>A large passenger jet sitting on top of an airport runway.</p>	<p>Model 1: A large jetliner sitting on top of an airport runway.</p> <p>Model 2: An airplane sitting on the tarmac at an airport.</p> <p>Model 3: A large air plane on a run way.</p>
	<p>A train traveling down tracks next to a forest.</p> <p>A train traveling down train tracks next to a forest.</p> <p>A train traveling down tracks next to a lush green field .</p>	<p>Model 1: A train traveling down tracks next to a forest.</p> <p>Model 2: A train on a train track with trees in the background.</p> <p>Model 3: A train is traveling down the railroad tracks.</p>

Since BLEU is a measurement of literal similarity between sentences, the lower the BLEU score, the more diverse the sentences generated by different describing models. The BLEU-1 score (comparing with 1-gram) represents the different preferences with words while the BLEU-4 score (comparing with 4-gram) indicates the usage of different phrases and structures. DIV follows a similar strategy but focuses on the dissimilarity instead of the similarity.

Performance

We can clearly observe the mBLEU scores of the base model are high, while GroupTalk obtains much lower mBLEU scores, which shows that the sentences generated for one image by the base model are all very similar while our framework allows us to train several models generating image captions with high diversity. The mBLEU-1 scores show our describing models use words with different preferences while the mBLEU-4 scores shows the describing models tend to describe images with different phrases and structures. DIV supports the same conclusion from the perspective of dissimilarity.

In order to demonstrate the models do learn different describing preferences, we choose a set of words for aircraft, and count the occurrences of them in the sentences generated by the describing models in testing phase on MS COCO and the results are given in Table 3. We also give an example in Table 4. It is clear that the three models learn different preferences while Model 1 learns to distinguish the type of planes and Model 2 and 3 simply use ‘plane’ and ‘airplane’ with different biases.

Inspecting the classifier

To a deeper understanding of GroupTalk, we further inspect the classifier by the learnt features. We randomly sample 30 images from the testing set of MS COCO and generate sentences with three describing models. Then we forward these sentences through the classifier and visualize the feature extracted by the classifier (the activations of hidden units of the LSTM) with PCA in figure 2.

The figure shows that each model learns their specific distribution. Moreover, the sentences generated by these models

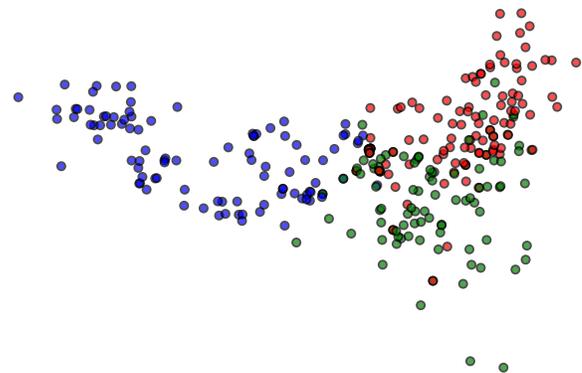


Figure 2: The embedding of sentences generated by three different models in the weak classifier.

mix in the center, which indicates that there are sentences shared in the training phase as what we design. This is consistent with the fact that people with difference preferences still give some basic descriptions that are similar.

5 Conclusion

In this paper, we propose a framework called GroupTalk to efficiently learn multiple language describing distributions simultaneously to mimic the diversity of the image captions written by human beings. Experiments on several benchmark datasets show that we diversify the captions describing the same image significantly without sacrificing the accuracy.

Acknowledgments

This work was supported in part by the National Basic Research Program of China under Grant 2015CB352300, NSFC(U1509206, 61572431), in part by the China Knowledge Centre for Engineering Sciences and Technology, and in part by the Fundamental Research Funds for the Central Universities.

References

- [Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, March 2003.
- [Chen and Zitnick, 2014] Xinlei Chen and C Lawrence Zitnick. Learning a Recurrent Visual Representation for Image Caption Generation. *arXiv.org*, November 2014.
- [Chen *et al.*, 2015] 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. *arXiv.org*, April 2015.
- [Dempster *et al.*, 1977] A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977.
- [Donahue *et al.*, 2014] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *arXiv.org*, November 2014.
- [Fang *et al.*, 2014] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, C Lawrence Zitnick, and Geoffrey Zweig. From Captions to Visual Concepts and Back. *arXiv.org*, November 2014.
- [Hessel *et al.*, 2015] Jack Hessel, Nicolas Savva, and Michael J Wilber. Image Representations and New Domains in Neural Image Captioning. In *Conference on Empirical Methods in Natural Language Processing*, August 2015.
- [Hochreiter and Schmidhuber, 1997] S Hochreiter and J Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [Hodosh *et al.*, 2013] M Hodosh, P Young, and J Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 2013.
- [Hong and Davison, 2010] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88, New York, New York, USA, July 2010. ACM.
- [Jiang *et al.*, 2015] Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yueting Zhuang. Deep compositional cross-modal learning to rank via local-global alignment. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 69–78, New York, NY, USA, 2015. ACM.
- [Jin *et al.*, 2015] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv.org*, June 2015.
- [Karpathy and Li, 2015] 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.
- [Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal Neural Language Models. In *International Conference on Machine Learning*, pages 595–603, 2014.
- [Lebret *et al.*, 2015] Rémi Lebret, Pedro O Pinheiro, and Roman Collobert. Phrase-based Image Captioning. In *International Conference on Machine Learning*, February 2015.
- [Li *et al.*, 2015] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *CoRR*, abs/1510.03055, 2015.
- [Mao *et al.*, 2014] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain Images with Multimodal Recurrent Neural Networks. *arXiv.org*, October 2014.
- [Mao *et al.*, 2015] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images. *arXiv.org*, April 2015.
- [Papineni *et al.*, 2001] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *the Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2001. Association for Computational Linguistics.
- [Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv.org*, September 2014.
- [Vinyals *et al.*, 2014] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, November 2014.
- [Wu *et al.*, 2015] Qi Wu, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, and Anthony Dick. Image Captioning with an Intermediate Attributes Layer. *arXiv.org*, June 2015.
- [Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, February 2015.
- [Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2(0):67–78, February 2014.