AUTOMATIC IMAGE CAPTION GENERATION IN VIETNAMESE USING CNN+LSTM APPROACH

Tin H. Hoang
University of Information Technology
14520956@gm.uit.edu.vn

Duy-Dinh Le
University of Information Technology
duyld@uit.edu.vn

Abstract

The research on image caption generation has mainly explored in English because most available corpora are in this language. The lack of datasets in other languages, especially non-inflected languages such as Vietnamese, is a drawback to exploit the capabilities of image captioning systems. To support the image caption generation task in Vietnamese, we have translated a subset of images taken from the MS COCO caption dataset into Vietnamese captions. The size of our dataset is comparatively admissible with 20,000 captions for 4,000 images. For evaluation performance of an image captioning system on Vietnamese dataset, we use the Show and Tell model [18] which is based on Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM). In addition, to deal with the ambiguity of the space character in Vietnamese grammar, we apply the state-of-the-art Vietnamese word segmentation approach (RDRsegmenter [11]) to the preprocessing phase of training the image captioning model. Extensive experiments are conducted on our dataset and the results show clear improvements when compared to generating in English captions. More remarkably, we obtain CIDEr-D score of 1.148 on test set when evaluating the Vietnamese dataset applied tokenization pre-processing.

1. Introduction

The automatic image caption generation (or image captioning) is a emerging challenge in artificial intelligence. This task takes an image as input and generate a textual sentence that describe the most salient aspects of the image. Research in this area involves various fields such as computer vision, natural language processing. Due to the renewed interest in deep neural networks, especially convolutional neural networks, many resurgent research are conducted on this challenge.

The majority of research on image caption generation have so far been experimented on English since most existing datasets are in this language. It is understandable because English is the most popular language. However, the application of image caption should not be restricted by language. In this paper, we study image captioning model that aims to generate captions in Vietnamese language which is a popular language spoken by more than 90 million people. Furthermore, it has potentials to make good performance in natural language processing side of Image captioning system.

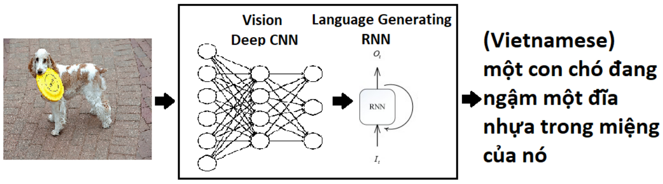

Figure 1. Basic CNN+RNN approach for Image Captioning based on a neural network consisting of a vision Convolutional Neural Network (CNN) followed by a language generating Recurrent Neural Network (RNN). The image captioning system takes image as input and try to generate a sentence in natural language which describe the input image.

The English language is one of the languages most commonly spoken throughout the world but it also has several drawbacks. One of drawbacks, which we think is a disadvantage to natural language processing in a image captioning system, is inflection. For example, the verb "go" can be inflected by tense or subject: go/goes, went, going,... ; or the noun "cat" in single/plural form is processed as two different words: cat/cats. In contrast, Vietnamese is considered as an analytic language. That means it is a language that primarily conveys relationships between words in sentences by way of helper words and word order, as opposed to utilizing inflections. For instance, Vietnamese uses the words "đã" (past tense), "đang" (continuous tense),... before the verb to indicate tense, and the main verb is not changed.

To solve the problem of lacking Vietnamese image captioning datasets, we randomly choose 4,000 images from

MS COCO 2014 caption dataset and translated the captions to Vietnamese (by both automated machine translation and human translation). In detail, each image in dataset has 5 captions, so totally we have 20,000 image-caption pairs. For evaluation performance of an image captioning system on Vietnamese dataset, we use NIC version 2 model [19], which is an improvement from NIC - the winner of MS COCO 2015 image captioning challenge.

Vietnamese language also has its own drawbacks that is the ambiguity of the space character. It's not like English, one Vietnamese word can have two or more syllables separated by space characters. For example, "máy tính xách tay" (laptop) is consider as only one word in Vietnamese. To minimize the effect of this ambiguity, we propose using a state-of-the-art Vietnamese word segmentation module to connect all syllables of the same word.

Lastly, since it needs a lot of time to manually translate captions in large scale dataset, we also propose a strategy to enrich the dataset by combining Human-translated captions and Machine-translated captions.

In summary, our contributions are as follows. First, we present a new Vietnamese caption dataset which is translated from MS COCO 2014 by machine translation and also by human translation in order to get better fluency. Second, we use NIC model to exploit the performance of an image captioning system on Vietnamese to compare with English. Third, we apply a Vietnamese word segmentation module to the pre-processing phase before training the image captioning model, which yeilds the better result. Finally, we propose a combine method between machine-translated captions and human-translated captions to improve performance when the human resources is limited.

2. Related Work

Recently, deep convolutional neural networks (CNN) is adapt in many computer vision tasks. More and more CNN models are created to tackle the task of object recognition or image classification, e.g. AlexNet, VGG, Resnet, Inception-v3 [16]. Similarly, natural language processing domain has seen increased adaptation of deep neural networks. In particular, the performance of machine translation task is improved by adopting sequence-to-sequence training using recurrent neural networks (Cho et al. [3]; Bahdanau et al. [1]). The encoder-decoder framework (Cho et al., 2014 [3]) using for machine translation inspired many model in image caption generation task, as this task is analogous to translating vision image to natural language sentence.

Many research groups have reported a significant improvement in image caption generation since 2014. One of the first methods using neural networks is proposed by Karpathy & Li [6], their model based on computing sentence and image similarity as function of Region-based Convolutional Neural Networks (R-CNN) object detections with outputs of a Bi-directional RNN. In 2015, the "Show and Tell" model introduced by Vinyals et al. [18] with the idea of combining a convolutional neural network for image feature extraction and long short-term memory for generating captions. This work was later followed by Vinyals et al. (2017) [19], which the authors updated CNN to Inception-v3 [16] model. In the same year as Show and Tell model, the now-trending approach using attention mechanism is first introduced in "Show, Attend and Tell" model (Xu et al. [20]) that aligns visual information and sentence generation for improving captions and understanding of model behavior. Different from other approaches, Zhou Ren et al. [14] introduced a novel decision-making framework using Deep Reinforcement Learning-based approach.

In the dataset side, a few caption corpora in languages other than English have been collected, e.g. Chinese (Li et al. 2016 [8]), German (Elliott et al. 2016 [4]), French (Rajendran et al. 2015[13]), Japanese (Miyazaki and Shimizu 2016 [10]; Yoshikawa et al. 2017 [21]). To the best of our knowledge, there are no image caption corpora for Vietnamese. With the creation of the new Vietnamese Captions dataset in this project, we aim to tackle this situation and thereby remove the obstacle to expand the research horizon.

3. Image Caption Generation Model

In this section, we briefly review the caption generation method proposed by Vinyals et al [19], which is used in our experiments (section 4). The model is usually mentioned with alias "NIC" (Neural Image Captioning) in other papers.

This model consists of 2 components: a convolutional neural network and a recurrent neural network. Specifically, NIC-v2 [19] using Inception-v3 as CNN component for extracting image feature, and LSTM as RNN component for word generating. Avoiding overfitting, the weights initialization of the Inception-v3 component is pretrained on very large Imagenet dataset [15]. The choice of LSTM among other RNNs (such as Bidirectional Recurrent Neural Network in [6]) is to deal well with vanishing and exploding gradients, also it has shown state-of-the art performance on sequence tasks such as translation.

In more detail, let I be the input image, and the true caption describing this image be $S = (S_0, \dots, S_N)$. The procedure in [18] as follows:

$$x_{-1} = W_{im}\text{CNN}(I) \quad (1)$$

$$x_t = W_e S_t, \quad t \in \{0 \dots N-1\} \quad (2)$$

$$p_{t+1} = \text{LSTM}(x_t), \quad t \in \{0 \dots N-1\} \quad (3)$$

Where $CNN()$ is a function that return image feature extracted by CNN module, the index $t = (t_0, \dots, t_N)$ to

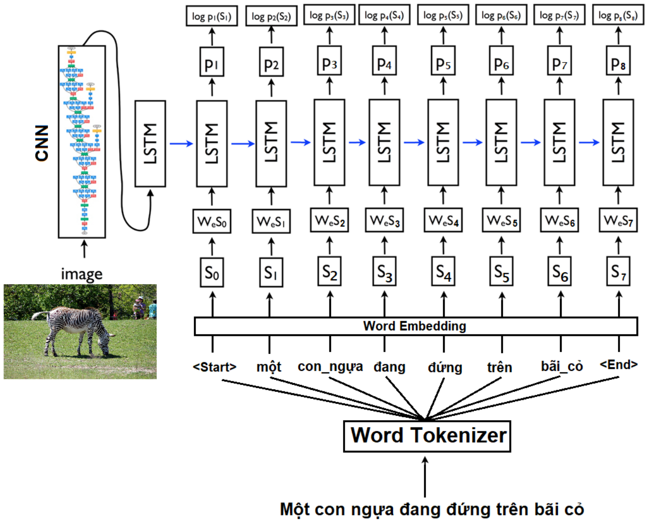

Figure 2. Training CNN+LSTM model for image caption generation: First, the image is put into CNN to extract feature, then the feature vector is set as input for LSTM at step -1. Based on info from previous steps, LSTM generates the caption word-by-word, the preceding predicted word is input for next step. The weight of LSTM network is fit with the groundtruth caption for this image.

denote the position of a word in a sentence, N is the length of caption. The image feature input to LSTM only one time, at $t = -1$, after that LSTM predicts word-by-word based on image content and previous words in sentence. The generation process is repeated until LSTM outputs special end-of-sentence token or reaches maximum length.

4. Experiments

4.1. Dataset

To create Vietnamese caption dataset, we randomly took 4,000 images from MS COCO training set. Each image is annotated with 5 captions, so that is 20,000 image-caption pairs. The translation is obtained by both English-to-Vietnamese machine translation and human translation. For the machine translation we chose Google Translate as it is among the best English-to-Vietnamese tranlation system. The human translation is oriented by some translation guides:

- Omitting specify names (such as New York's Chinatown, Mulholland Drive,...), number (house address, specify time of clock,...), adjectives that expresses personal feeling.
- Allow to keep English words that is commonly used in Vietnamese language (vest, pizza, laptop, tivi,...)
- In the case of conflict between caption and image (e.g. wrong object color, incorrect sexual of person, caption is totally irrelevant to the image,...), the content of image (not the caption) should be the crucial factor to make final decision.

Language of Captions	#Images	#WordClasses	#TotalWords	Avg.Length
English	4000	6672	211690	10.5845
VN - Google Translation	4000	3490	257046	12.8523
VN - Human Translation	4000	2778	260571	13.0286
VN - Tokenized Google	4000	4745	222062	11.1031
VN - Tokenized Human	4000	4065	223708	11.1854

Table 1. Statistics of our 4K Dataset. We use the same 4000 images, only the language of captions is changed to compare performance of image captioning system on different dataset-creation methods.

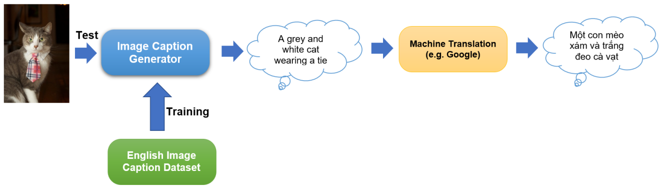
Test
Image Caption Generator
A grey and white cat wearing a tie
Machine Translation (e.g. Google)
Một con mèo xám và trắng đeo cà vạt
Training
English Image Caption Dataset

Figure 3. **Method 1:** Translate result of an English model into Vietnamese

Train/Test Ratio: To make our results comparable, we adapt the same train/test ratio from original paper [18]. In detail, Vinual el.at. took 100% images from MSCOCO train set (82,783 images) and 85% images from MSCOCO validation set (40,504 images), in total they had 117,211 images for training model. For test set, they reserved 4,050 random images from the MSCOCO validation set as test. In this experiment, we take all 4,000 images of our dataset as train set. By that train/test ratio, we take 138 random images from MSCOCO validation set for test. The translation for test set is done the same as 4,000 train test.

4.2. Methods

Given a model for Image Captioning in English, we propose several methods to make it work in Vietnamese. The simplest idea to create a image captioning is just translate the result of an English image caption generator system (see Figure.3). With this method, there are not only need to create new Vietnamese dataset, but also we can take advantage of pre-trained English model on large data. Although, this approach is affected by the Machine Translation errors and fluency. We are not surprised that it yeilds the worst performance in 5.2.

The next idea is training the model on new Vietnamese dataset. This lead to two ways to create a new dataset: translating already published English dataset or crawling new images and annotating them in Vietnamese. Due to restricted time and human annotators, we consider translating popular English dataset MS-COCO is the better option. There are also two approaches for translating dataset:

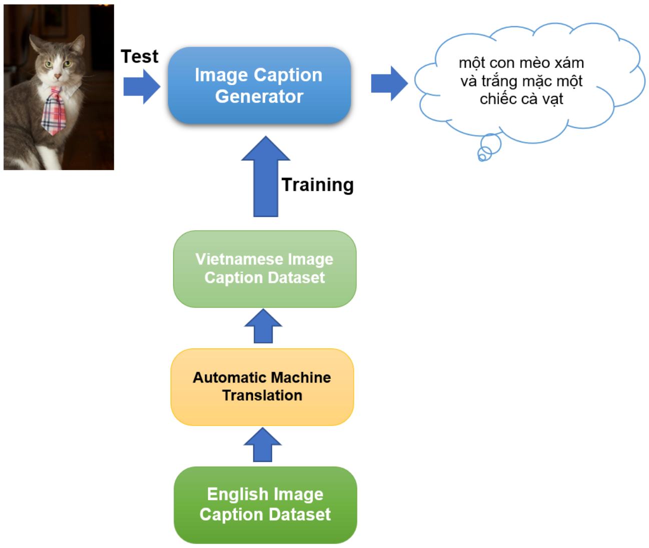
Test
Image Caption Generator
một con mèo xám và trắng mặc một chiếc cà vạt
Training
Vietnamese Image Caption Dataset
Automatic Machine Translation
English Image Caption Dataset

Figure 4. **Method 2:** Train model on Vietnamese dataset translated by Automatic Machine Translation (e.g. Google Translate)

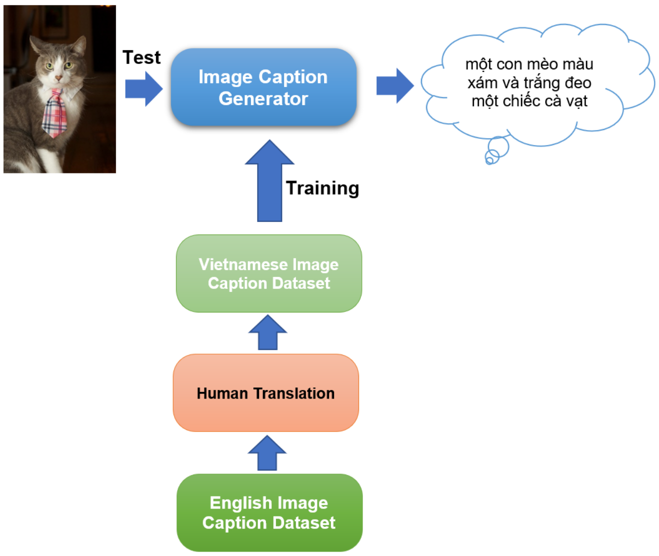
Test
Image Caption Generator
một con mèo màu xám và trắng đeo một chiếc cà vạt
Training
Vietnamese Image Caption Dataset
Human Translation
English Image Caption Dataset

Figure 5. **Method 3:** Train model on Vietnamese dataset translated by Human Translation

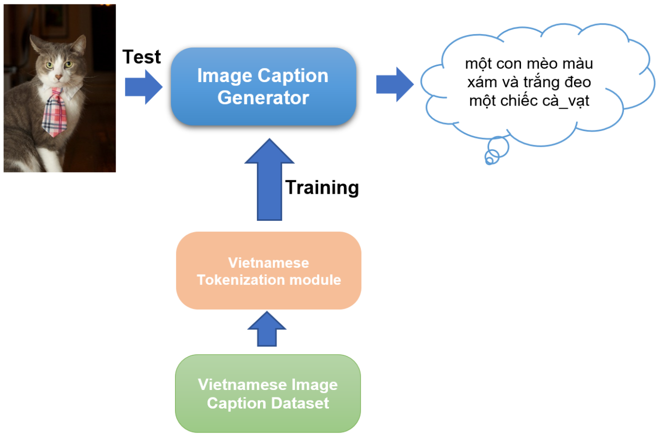
Test
Image Caption Generator
một con mèo màu xám và trắng đeo một chiếc cà_vạt
Training
Vietnamese Tokenization module
Vietnamese Image Caption Dataset

Figure 6. **Method 4:** Train model on Vietnamese dataset (same as method 1 or 2) and use word segmentation (tokenization) module in preprocessing phase

machine translation and human translation. Using machine translation (see Figure.4) is convenient and we can make a full automatic system which is not required human intervention. However, machine translation errors and fluency of translated sentences still make a great effect to performance of systems using this method. To improve the fluency, the translation should be done by human. Needless to say, Human translation (see Figure.5) is very time consuming, but it promises better grammar and fluent generated captions.

In addition, we propose using a tokenization module to preprocess the Vietnamese captions before training the model (see Figure.6). As we know it, the space character in Vietnamese is not used as word delimiter. That means a word in Vietnamese can have two or three syllables which are separated by space characters. The word segmentation (or tokenization) is a task to divide a sentence into its component words. By applying tokenization in preprocessing phase, we hope it will shorten length of Vietnamese captions and deal with the ambiguity of the space character. In this paper, we use the state-of-the-art word segmentation approach RDRsegmenter [11] which is based on the Single Classification Ripple Down Rules methodology. We use '_' character to connect syllables of a word. An example about applying tokenization to a normal Vietnamese caption:

Caption: Hai con voi trng thành đi lang thang xung quanh trong môi trng sng ca chúng

Tokenized Caption: Hai/con/voi/trng_thành/đi/lang_thang/xung_quanh/trong/môi_trng/sng/ca/chúng

Based on the performance in section 5.2, we know the result of model trained on human-translated captions is finest. But the dataset constructed by manual labour consumed a lot of time and human resources. To minimize the human effort, we propose strategies which combine human-translated captions and machine-translated captions. In detail, we assume a scenario: "Time and human resources is only enough for translating 800 images (or 4,000 captions) into Vietnamese, how to improve the performance?", and we come up with two strategies:

- **Strategy 1:** With 4,000 captions can be translated by human, we distributively translate one of the five English captions of each image into Vietnamese, that leaves the other four will be translated by machine. Overall, each image in this 4K dataset will have 1 human-translated caption and 4 machine-translated caption.

- **Strategy 2:** Human translators collectively translate 4,000 captions for all five captions of 800 images. That means this dataset will have 800 human-transled images and 3200 machine-translated images.

We published our experimental code and Vietnamese dataset at [1]

5. Evaluation

5.1. Evaluation Metrics

The researchers usually use machine translation evaluation metrics to measure the quality of generated captions with considering that generating image captions is the same as translating image "language" into natural language. In this paper, we use 6 automatic measures for evaluating: BLEU-1, BLEU-2, BLEU-3, BLEU-4 (Papineni et al., 2002 [12]), ROUGE-L (Lin, 2004 [9]), and CIDEr-D (Vedantam et al., 2014 [17]). All these measures compute a score that indicates the similarity between the system output and one or more human-written reference texts (e.g., ground truth translations or summaries).

- **BLEU** (BiLingual Evaluation Understudy) [12]: It computes the geometric mean of n-gram precision scores by counting n-gram (1 to 4-gram) co-occurrences. The result is multiplied by a brevity penalty in order to avoid too short sentences. BLEU is one of the most popular metrics for machine translation evaluation. However, the correlation between human judgments and unigram BLEU is still debatable [7, 5].
- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) [9]: is a package (or set of metrics) for automatic evaluation of text summaries. A variant of ROUGE called ROUGE-L is usually used in image captioning evaluation, which computes F-measure based on the Longest Common Subsequence.
- **CIDEr** (Consensus-based Image Description Evaluation) [17]: is a metric specifically designed for image captioning evaluation. It measures consensus between candidate image description and the reference sentences by performing a term-frequency inverse document frequency (TF-IDF) weighting for each n-gram. A variant of CIDER called CIDEr-D is preferred, since it more robust to "gaming" problem, which occurs when sentences get high scores from automated metric, yet produce poor results when judged by a human.

Keep in mind that: for all evaluation metrics, higher scores are better. In this paper, we focus more on the CIDEr metric, since it is specifically designed for image captioning evaluation.

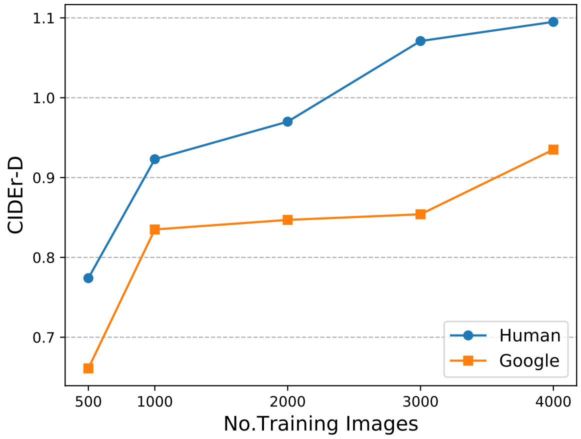

Figure 7. Relationship between the CIDEr-D score and the size of Vietnamese dataset (number of training images). The Human-translated dataset outperformed the machine-translated dataset for all training dataset size. There is tremendous improvement when No.Training Images increases from 500 to 1000.

5.2. Results

In this section, we report the results conducted on our Vietnamese dataset.

We use the official MSCOCO caption evaluation tool [2] to compute the scores. However, all of those metrics in the code are computed with assumption that words are separated by space characters. Since this gives a slight advantage to Vietnamese evaluation, given the space character is not used as word delimiter in Vietnamese language, we run tokenization for both generated caption and reference captions (groundtruth) before using the evaluation code.

Note that: All scores are evaluated on 138-image test set with tokenized Vietnamese captions. Except English score which is computed on the same 138-image test set but with English captions.

Human scores in Table 2 were computed by taking one of the groundtruth human-translated captions to compare against the other four. This score is computed on test set

6. Conclusion

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 2

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[1] https://github.com/Flavius1996/Image-Captioning-in-Vietnamese

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
English	0.677	0.498	0.358	0.249	0.506	0.949
English-to-VN (Method 1)	0.630	0.440	0.300	0.191	0.459	0.827
Google-Translated (Method 2)	0.639	0.475	0.352	0.251	0.493	0.935
Human-Translated (Method 3)	0.683	0.521	0.389	0.284	0.521	1.095
Tokenized Google-Translated (Method 4)	0.616	0.451	0.325	0.231	0.488	0.890
Tokenized Human-Translated (Method 4)	**0.686**	**0.529**	**0.394**	0.281	**0.539**	**1.148**
Human Score	0.691	0.516	0.379	0.269	0.525	1.254

Table 2. Performances of main methods when adapting Image Caption model for Vietnamese language. The scores is evaluated on our 4K image-caption dataset. The first score "English" is performance of the model trained only on English captions.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
800 Human	0.622	0.463	0.326	0.216	0.488	0.889
1 Human cap + 4 Google cap (Strategy 1)	0.634	0.470	0.343	0.244	0.491	0.936
800 Human + 3200 Google (Strategy 2)	0.630	0.469	0.345	0.247	0.501	0.960

Table 3. Performances on strategies combining Human-translated captions and Machine-translated captions. The "800 Human" is evaluated on the dataset using only 800 Human-translated images.

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[4] D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016. 2

[5] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 5

[6] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[7] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 5

[8] X. Li, W. Lan, J. Dong, and H. Liu. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 271–275. ACM, 2016. 2

[9] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 5

[10] T. Miyazaki and N. Shimizu. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1780–1790, 2016. 2

[11] D. Q. Nguyen, D. Q. Nguyen, T. Vu, M. Dras, and M. Johnson. A fast and accurate vietnamese word segmenter. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.* European Language Resources Association (ELRA), 2018. 1, 4

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 5

[13] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran. Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519*, 2015. 2

[14] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li. Deep reinforcement learning-based image captioning with embedding reward. *arXiv preprint arXiv:1704.03899*, 2017. 2

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 2

[17] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[18] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE, 2015. 1, 2, 3

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2017. 2

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 2

[21] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*, 2017. 2