BraIN: A Bidirectional Generative Adversarial Networks for image captions

Yuhui Wang
School of Electrical Engineering & Computer Science,
Washington State University
mark166.wang@wsu.edu

Diane Cook
School of Electrical Engineering & Computer Science,
Washington State University
djcook@wsu.edu

ABSTRACT

Although progress has been made in image captioning, machine-generated captions and human-generated captions are still quite distinct. Machine-generated captions perform well based on automated metrics. However, they lack naturalness, an essential characteristic of human language, because they maximize the likelihood of training samples. We propose a novel model to generate more human-like captions than has been accomplished with prior methods. Our model includes an attention mechanism, a bidirectional language generation model, and a conditional generative adversarial network. Specifically, the attention mechanism captures image details by segmenting important information into smaller pieces. The bidirectional language generation model produces human-like sentences by considering multiple perspectives. Simultaneously, the conditional generative adversarial network increases sentence quality by comparing a set of captions. To evaluate the performance of our model, we compare human preferences for BraIN-generated captions with baseline methods. We also compare results with actual human-generated captions using automated metrics. Results show our model is capable of producing more human-like captions than baseline methods.

CCS CONCEPTS

• ; • **Computing methodologies**; • **Artificial intelligence**; • **Natural language processing**; • **Natural language generation**;

KEYWORDS

LSTM, GAN, BIDIRECTIONAL, IMAGE CAPTION

ACM Reference Format:
Yuhui Wang and Diane Cook. 2020. BraIN: A Bidirectional Generative Adversarial Networks for image captions. In *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2020), December 24–26, 2020, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3446132.3446406

1 INTRODUCTION

Image caption generation has received much attention in recent years. Annotating images with captions provides textual explanations of the picture highlights. The growing popularity of this research topic is due in part to the many varied applications that benefit from this capability, such as image commenting in social chatbots and providing assistance to visually-impaired people. Image caption generation is difficult because it involves both computer vision and natural language processing technologies. To generate a meaningful description, the model is required to recognize objects in images, detect relationships among those objects, and then describe this information using natural language. Over the past few years, many methods have been explored for image captioning [1–3]. These methods combine image representation and sentence generation. Benefitting from recent advancements in deep learning, much of the existing work applies convolutional neural networks (CNN) to the image representation portion of the problem and recurrent neural networks (RNNs) to the sentence generation piece. The performance of these methods has steadily improved based on automated evaluation metrics, such as BLEU [5], CIDEr [6], METEOR [6], and BERTScore [4].

Image captioning contains two important components: one is caption generation, and another is image representation. Automatically-generated captions can often be easily differentiated from human-written captions, which are more diverse. Image caption models learn from human-provided captions, and close matches to the ground truth receive high scores for automatic evaluation metrics. In order to better capture human-like natural and diverse expressiveness, Generative Adversarial Networks (GANs) [7] have been explored. The idea behind image-captioning GANs is to train a discriminator to detect the misalignment between an image and a generated sentence and improve the generator's ability to align the caption with the corresponding image. Recent work applying GANs generally utilizes either reinforcement learning [8] or the Gumbel softmax relaxation [9]. For image representation, many approaches encode the whole image to a global feature vector. Because these methods may suffer from the problems of missing objects and mispredictions, this traditional approach can be improved using an attention mechanism [10]. This method has found some success by automatically searching the parts of a source image that are most relevant to a target word, avoiding the pitfalls of missing objects and mispredictions.

Despite these advances, image captioning is far from being a solved task. Language signals are composed of basic units that are distinct and individuated [28]. These basic units are assembled in varying orders to represent different meanings. While language can thus be considered discrete, generating diverse, natural, and human-like captions is still a big challenge. One shortcoming of existing approaches is that they employ a limited view of language by generating sentences unidirectionally. By selecting words in one direction, from the beginning to the end of each sentence, these

ACAI 2020, December 24–26, 2020, Sanya, China © 2020 Copyright held by the owner/ author(s). ACM ISBN 978-1-4503-8811-5/20/12.
https://doi.org/10.1145/3446132.3446406

methods may lose valuable meaning and naturalistic expression in the sentence as a whole. Language translation scholars have long been aware of the need to consider the sentence in multiple directions [27]. However, this approach has not yet been researched for automated captioning.

In all natural languages, each word is related to words both before and after it, regardless of the adjacency of the words. We hypothesize that to generate human-like captions, a model needs to consider the portions of a sentence that appear before and after each generated word. Inspired by natural language understanding [22], we propose a bidirectional image-captioning model to improve the naturalness of the generated captions. This model builds on the foundation of GAN-based language generation and attention mechanism-enhanced image representation. We demonstrate that enhancing these methods with bidirectional captioning improves the quality of the generated text. We postulate that these improvements will be detected both by traditional automated measures as well as by human evaluators.

2 RELATED WORK

Generating text explanations of images relies on understandable text generation as well as accurate image representation. Most current methods that perform the first step of text generation via image captioning models [11, 12] use an encoder-decoder framework that operates similarly to sequence learning. An encoder-decoder framework will use an encoder to convert the information from one format to another and use a decoder to extract the information we need. Typically, the networks are trained using maximum likelihood estimation (MLE) [13] or reinforcement learning [14]. Although these methods achieve outstanding performance on automated evaluation metrics (BLEU, CIDer, METEOR, etc.), the generated captions usually contain a sequence of commonly-appearing n-gram patterns, spliced together. As a result, these generated sentences lose the natural expression that is found in human-generated text. To minimize this weakness, diverse beam search and ensemble methods [15] have been proposed. The diverse beam search is a graph search algorithm that decodes the list of sentences then enforces diversity between each sentence. To further address the need for natural text expressions, some studies [16] employ variational autoencoders. A variational autoencoder will sample a vector from a Gaussian distribution and feed this vector with the features of the input image into the decoder to generate a meaningful, natural caption. Yet another approach to improve the expressive quality of the captions is Generative Adversarial Networks (GANs) [17]. The GAN generates random samples that enforce the network to produce an output that exhibits greater diversity and makes the sentence more natural. Despite the naturalness of the language, captions generated from these existing methods still exhibit limited diversity. This is an ongoing challenge that we attempt to address through bidirectional sentence generation.

Compounding the challenge of text generation, accurately representing images within the image captioning process presents additional difficulties. Some of the primary image representation hurdles include accounting for missing objects within the image and the potential for producing poor predictions. The above-mentioned approaches encode the whole image, which will lead the generated sentences to describe only some of the features within the image, omitting important details. To address this limitation, we introduce an attention mechanism.

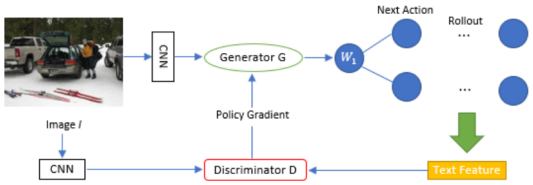

Figure 1: The BraIN framework.

Attention mechanisms have been effective in the field of computer vision [18, 19] and natural language processing [20, 21]. This process segments the important details of input information into smaller pieces. As a result, a model which employs an attention mechanism focuses on features from these smaller, specific details, instead of using a feature vector that represents only global information. In natural language understanding, Bahdanau et al. [22] construct a bi-directional neural network to align a source sentence (the original text) with the corresponding target sentence (a rephrasing or summary of the text). This method automatically searches the parts of a source sentence that are most relevant to a target word. Xu et al. [19] explore two attention-based image captioning methods, soft-attention and hard-attention, and analyze how an attention mechanism works for description generation. The difference between the soft-attention and hard-attention methods is the way they train the attention. Soft-attention is trained by standard back-propagation, while reinforcement learning trains hard-attention. Yao et al. [20] exploit a temporal attention mechanism to capture global temporal structure among video frames based on the Bahdanau's soft-alignment method. The temporal attention mechanism [18] makes the decoder selectively focus on selected key frames which are most relevant to the predicted word. The attended attributes model [20] first utilizes multiple approaches (e.g., k-NN, multi-label ranking and fully convolutional network) to obtain a set of proposed semantic concepts, and then integrates them into one vector, allowing the attention mechanism to guide the language model for description generation. Disregarding such details can further lead to image classification error. To address these problems, our BraIN model will combine global features with object-level features which are generated by the soft-attention method.

3 BRAIN FRAMEWORK

We propose a new framework, BraIN, for image caption generation. BraIN combines the benefits of a bidirectional model with an attention mechanism and a conditional GAN. The proposed framework consists of two networks, a caption generator G, and a caption discriminator D. The framework is illustrated in Figure 1. Given an image I, the goal of the generator is to produce natural and semantically-relevant captions, while the discriminator's goal is to evaluate how well the captions describe the image. The two

networks play a min-max game as follows:

$$\min_{\theta} \max_{\eta} L\left(G_\theta, D_\eta\right) \tag{1}$$

In Equation 1, L represents the overall loss function, and θ and η are trainable parameters guiding the generator and discriminator, respectively. For a reference image I, the generator G_θ outputs a sentence S as the corresponding caption. The discriminator D_η aims to correctly estimate the relevance score of S with respect to a corresponding human-written caption h.

In our framework, the generator G contains three components. These are a convolutional neural network (CNN) that extracts information from the image, a bidirectional LSTM that generates captions, and an attention model that leads the LSTM model to focus on important features of each generated word. The CNN extracts a fixed-dimensional feature $f(I)$ from the input image that captures salient information about the image. The attention model will generate a matrix A for each word in the caption. The attention model calculates $A \times f(I)$ and delivers the results as input to the bidirectional LSTM. The bidirectional LSTM decodes a sentence by taking the output from the attention model and two random vectors z_1 and z_2 and generating a corresponding sequence of words. The role of the two random vectors is to increase the randomness of the sentence, thus improving word diversity and overall sentence expressivity.

The discriminator D contains three components as well. These are a CNN that extracts information from the image (as found in the generator), a bidirectional LSTM that extracts information from the candidate sentences, and a scoring system that calculates the similarity between two vectors output by the LSTM and CNN, respectively. Given an image I and a candidate descriptive sentence $S = (w_1, w_2, \cdots w_T)$, the discriminator will use the CNN to decode the image I into a $m \times n$ dimension vector, $V(I)$. Similarly, the discriminator will employ an LSTM to decode a descriptive sentence S into a same-dimension vector, $V(S)$. Finally, the scoring system will measure the quality of the sentences by calculating the distance between the two vectors, $V(I)$ and $V(S)$, as shown in Equation 2:

$$R_\eta\left(I, S\right) = \sigma\left(\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(V(I)_{ij} - V(S)_{ij}\right)^2}\right) \tag{2}$$

In this equation, variable η represents the discriminator parameters and σ is a logistic function that turns the distance between two matrices into a probability value in [0,1].

4 GENERATOR

BraIN's generator G_θ is designed as a standard encoder-decoder architecture. The generator framework is depicted in Figure 2. Each LSTM block shows in the graph is a single LSTM cell. From image I, the caption model will extract a fixed-dimensional feature vector, f(I), using a CNN as an encoder. The decoder implemented by the bidirectional LSTM network maps image features to a word sequence. We apply an attention mechanism in the bidirectional LSTM network to focus on the most relevant area in the image for a generated word. We use $\overrightarrow{A_{t-1}}$ and $\overleftarrow{A_{t+1}}$ to represent activation values emanating from the word occurring immediately before position t in the sentence and the word occurring immediately after position t, respectively. The arrows indicate the direction of influence, in this

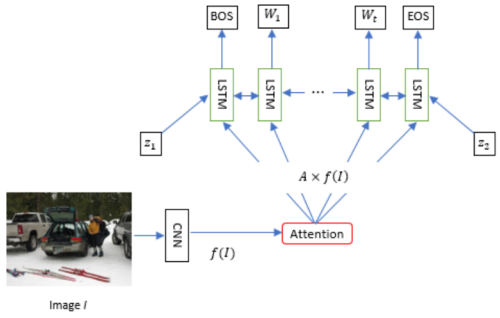

Figure 2: The framework for generator Gθ. The generator selects a word for each position in the sentence from beginning (BOS) to end (EOS).

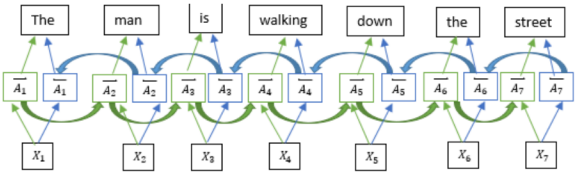

Figure 3: The word generation process.

case from t-1 to the right in the sentence and from t+1 to the left in the sentence. The bidirectional LSTM networks calculate $(\overrightarrow{A_t}, \overleftarrow{A_t})$ based on the likely values for words appearing directly before and after the current position, or $(\overrightarrow{A_{t-1}}, \overleftarrow{A_{t+1}})$. The LSTM selects word W_t based on the value of $(\overrightarrow{A_t}, \overleftarrow{A_t})$, taking advantage of not only the information from potential words that appear earlier in the sentence but also the information from potential words that will appear further along in the sentence. Figure 3 shows how each word in a sentence has been generated from this information. Variable X_t in Figure 3 represents the vector of image features that are calculated by the attention mechanism. Additionally, $(\overrightarrow{A_t}, \overleftarrow{A_t})$ in Figure 3 represents the activation value from two directions at state t. In state t, word W_T is produced by Equation 3:

$$W_t \sim \pi_\theta\left(W_t | I, W_{t-1}, W_{t+1}\right), \quad t \in (1,\ T) \tag{3}$$

In this equation, π_θ is a word distribution over all words in the vocabulary. Variable T represents the maximal length (in number of words) allotted for the caption. A complete caption $S_c = \{W_1, W_2, \cdots\}$ can be produced by generator G_θ, by sequentially sampling words according to π_θ.

5 TRAINING GENERATOR

Using the BraIN algorithm, we cannot directly apply gradient descent for G_θ through backpropagation; thus, we employ a Policy Gradient method. The policy gradient method originates from reinforcement learning [23]. Reinforcement learning (RL) is a learning method which deals with sequential decision-making. The RL problem can be regarded as an agent that has to make decisions in an environment in order to achieve the highest reward. The basic reinforcement parameters are $(S,\ \Delta, R, \pi)$. The S stands for a set of

environment and agent states. The Δ stands for a set of actions of an agent. R is rewarded that agent transit from S to $S^{'}$ by taking action $\Delta^{'}$. π is a policy that contains a sequence of action which try to achieve the maximum reward. The basic idea of this method is to treat a sentence as sequence of actions, wherein each action consists of generating a candidate word W_t. The choice of action is governed by a policy, π_θ which is word distribution. Although the classic policy gradient method can solve the non-differentiable problem, our model still faces some challenges. The bidirectional LSTM network is different from the unidirectional LSTM network as it can capture both past and future information, based on the activation values $(\overrightarrow{A_{t-1}}, \overleftarrow{A_{t+1}})$. Currently, the traditional reinforcement learning-based policy gradient method only handles a one-direction update for an LSTM model. As a result, we cannot directly apply the standard policy gradient. To extend the traditional method, we update the weight from two directions separately. For each single direction, we update the weight only in this direction. Once this step is complete, the policy gradients can be employed to update the weights, when generating the word W_t, for calculating either $\overrightarrow{A_{t-1}}$ or $\overleftarrow{A_{t+1}}$.

In BraIN, the generative procedure works as follows. First, we begin with an empty sentence as the initial state, denoted S_0. We regard each word w_t as an action selected from the set of possibilities, $w_t \in W$. At each step t, the policy π_θ considers features $A(I)$ from the attention model, the activation value $\overleftarrow{A_{t+1}}$ from later in the sentence, and the activation value $\overrightarrow{A_{t-1}}$ from earlier in the sentence as input. Based on the input, BraIN computes a conditional distribution $\pi_\theta\ (w_i\ |A(I), \overrightarrow{A_{t-1}}, \overleftarrow{A_{t+1}})$. W includes all vocabulary, an indicator of sentence start BOS, and an indicator of sentence end EOS . Through this conditional distribution, an action w_t will be sampled. The sentence will be terminated once $W_t = BOS$ or EOS . For all other cases, the sampled word W_t will be inserted at the current point t in the sentence.

The reward of the sequence of actions S is $R_\omega(I, S)$, which is generated by the discriminator D_η and controlled in part by a hyperparameter ω. However, the discriminator can only generate a score based on a complete sentence. Thus, individual rewards will not be assigned for each W_t. In order to solve this problem and observe a reward for each word position, we employ a K-times Monte Carlo rollout process [23] to explore the rest of the unknown words $S_{t+1:T}$ based on the current caption generator, G_θ. The process will rollout K times, which generates K different sentences based on the current caption generator G_θ. The unique sentences result from choosing different actions during each iteration. $S_{t+1:T}$ represents the sequence of words simulated by the Monte Carlo rollout process, which makes the sentence sufficiently complete to obtain a reward value. More specifically, the future reward $r_{\theta,\eta}(W_t|I,\ S_{1:t-1})$ can be approximated by the expected score over K rollout, simulated captions. This is shown in Equation 4:

$$r_{\theta,\eta}\left(W_t|I,\ S_{1:t-1}\right) \cong \ \frac{1}{K}\sum_{k=1}^{K} R_\omega\left(I, S_{1:t} \oplus S_{t+1:T}\right) \tag{4}$$

In Equation 4, $\oplus$represents the concatenation operation. $S_{1:t-1}$ denotes the words that were generated before action W_t is taken, and T represents the maximal length of the sentence.

To train the generator G_θ network, we view maximizing this expected reward $r_{\theta,\eta}$ as the learning object, which leads the generator to create the greatest-reward sentence. We derive the gradient of this objective θ as shown in Equation 5:

$$E[\sum_{t=1}^{T} \nabla_\theta \pi_\theta\left(W_t|I,\ z, S_{1:t-1}\right) \cdot r_{\theta,\eta}\left(W_t|I,\ S_{1:t-1}\right)] \tag{5}$$

where T is the maximal length of the sentence.

6 DISCRIMINATOR

The discriminator evaluates how well a caption $\mathcal{S}$ describes an image I by creating a corresponding score. Naturalness and relevance to the image are two criteria a good textual description, or caption, needs to satisfy. In order to meet the two criteria, we need to consider three types of descriptions for each image I. They are $\mathcal{S}_h$, the set of descriptions provided by a human for image I; $\mathcal{S}_g$, the set of descriptions provided by a generator for image I; and $\mathcal{S}_m$, the set of descriptions provided by a human that is not associated with the specific image I. We create score $\mathcal{S}_h$ using a joint objective formulated in Equation 6:

$$\max_\eta\ L_D\left(\eta\right) = \frac{1}{N}\sum_{j=1}^{N} L_D(I_j;\eta) \tag{6}$$

In this equation, variable N represents the number of training images. For each image I_j,

$$\begin{aligned} L_D\left(I;\eta\right) \ = \ & E_{\mathcal{S}\in\mathcal{S}_h} \log R_\eta\left(I,S\right) + \alpha \cdot E_{\mathcal{S}\in\mathcal{S}_g} \log\left[1 - R_\eta\left(I,S\right)\right] \\ & + \beta \cdot E_{\mathcal{S}\in\mathcal{S}_m} \log\left[1 - R_\eta\left(I,S\right)\right] \end{aligned} \tag{7}$$

The first term in this summation helps the discriminator to value the human caption. The second term is used to distinguish human captions from machine-generated captions, and the third term ensures that the generated description is sufficiently related to the image. The weights α and β are used to balance the contribution of the corresponding terms.

7 EXPERIMENTAL EVALUATION OF BRAIN

We expect that our proposed BraIN bidirectional LSTM model will outperform unidirectional LSTM models in generating natural language captions, since the bidirectional LSTM can capture dependencies both from previous words and words that appear later in the sentence. Therefore, we use both human evaluation and automated evaluation metrics to evaluate our BraIN model. The human evaluation is included to specifically evaluate the naturalness of the caption. At the same time, the automated evaluation metrics reflect the relevance between the caption and image. Though the human evaluation can also reflect the relevance between the caption and image, it is impractical to utilize human experts to evaluate all of the test results. Employing human evaluators is both expensive and time-consuming.

We conduct our experiments in the context of the MS COCO dataset [24], with 123,287 randomly-selected images. Each image is accompanied by 5 ground-truth sentences. Our experiments are based on a public split method [26]: 5000 images are randomly selected for validation and testing, and the rest (118,287 images) are used for training. In order to assess the performance improvement

Table 1: Performance of proposed BraIN method in comparison with baseline methods G-Gan and CAL. Performance is measured using traditional automated scores of BLEU-4, METEOR, and CIDer.

Model	BLEU-4	METEOR	CIDer
G-Gan	0.205	0.221	0.697
CAL	0.208	0.222	0.712
BraIN	**0.212**	**0.224**	**0.719**

offered by the proposed Bidirectional Adversarial Network (BraIN), we comparatively evaluate two baseline models as well, G-Gan [25] and CAL [1]. G-Gan implements adversarial learning for image captioning combined with a traditional, unidirectional LSTM. CAL is similar to G-GAN, in which a comparative score is employed for the discriminator instead of the binary score. To ensure a fair comparison, all image features are extracted by ResNet-152 [14]. All generator and discriminator text-decoders are implemented by LSTMs.

Before adversarial training, we pretrain generator G using standard MLE (maximum likelihood estimation) for 20 epochs, and we pretrain discriminator D for 5 epochs. In the adversarial training stage, two sub-networks are trained jointly, where each iteration consists of one step of G-update followed by one step of D-update. We set the mini batch size to 64, the learning rate to 0.0001, and K=16 for K-times Monte Carlo rollouts. During the testing, the captions are generated based on the learned policy and the final caption is selected based on the sentence that yields the top score.

To evaluate the alternative captioning methods, we employ both automated metrics and human preference. The automated metrics include BLEU-4 [26], METEOR, and CIDer. The results are summarized in Table 1. As we see from these results, BraIN outperforms the other two models based on automated evaluation measures. Combining the three measures, BraIN outperforms CAL by 0.98% and outperforms G-Gan by 3.16%. Even using these traditional metrics, the BraIN architecture offers improvements to the image captioning process. Although the evaluation metrics can represent the accuracy of the description related to the image, however, they overly focus on n-grams matching and pattern matching with ground truth captions and ignore the naturalness of the language.

To evaluate the caption naturalness, we further conduct human evaluation using a survey. The survey includes 50 images selected randomly from the test set. For each image, we present 3 captions that are generated by the 3 alternative methods. Figure 4 shows some captions generated by three different methods. Participants are asked to choose the caption that best describes the corresponding image in their opinion. A total of 32 participants were recruited. We collected a total of 1600 responses, which is the result of all 32 participants completing all of the survey questions.

Additionally, participants optionally provided their own image captions. We collected 250 captions provided by 5 of the participants. To gain insight on the limitations associated with automated evaluation metrics, we applied the BLEU-4, METEOR, and CIDer metrics to these 250 captions. As the results in Table 2 indicate, the human captions do not perform substantially better than the

Figure 4: Image captions generated from three different models: two baseline methods and our BraIN method.

Table 2: Performance of human captions, measured using BLEU-4, METEOR, and CIDer.

Model	BLEU-4	METEOR	CIDer
Human	0.193	0.238	0.857
Automated GAN (average)	0.208	0.222	0.709
MLE	0.298	0.252	0.919

automated GAN methods for the METEOR and CIDer metrics, and actually perform worse than the automated methods according to the BLEU-4 metric. The MLE method shows the best performance in all metrics. It is not surprising. Since metrics primarily focus on n-gram matching with respect to the references, while ignoring other important properties such as naturalness and diversity. These results indicate the need to assess image captions using human judges. They also hint that image captioning methods will benefit from increased naturalness in expression and variation.

The results of this experiment are shown in Figure 5. Overall, 39.62% of the chosen responses are captions generated by BraIN, 32.44% by CAL, and 27.94% by G-Gan. Thus, BraIN exhibits a 2.21% performance improvement over CAL and a 4.18% improvement over G-Gan. These results indicate that BraIN yields more natural and human-liked captions compared with the other two baseline models.

8 CONCLUSION

This paper presents a bidirectional generative adversarial network (BraIN) for image captions. A different generating model has been applied in adversarial learning, which better assesses the quality of captions by taking all the words from captions into consideration. Therefore, the captions model can improve the naturalness and correctness. Experimental results clearly demonstrate that our proposed method generates better captions in terms of both accuracy and naturalness across images. Also, the experimental results show the limitation of the evaluation metrics, caption generates from MLE outperform than human's caption in all metrics. It indicates that auto metrics can not evaluate how naturalness is the sentence

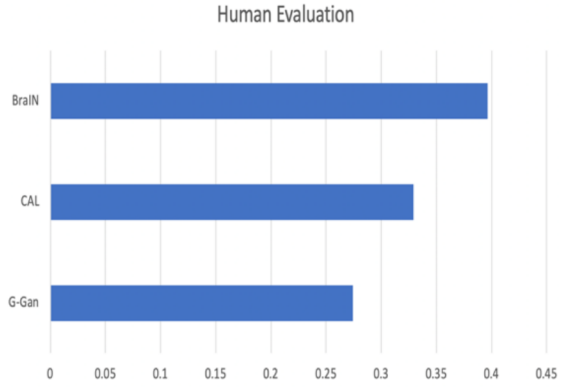

Figure 5: The result of human evaluation.

which is an important characteristic of the language. We would like to investigate some new metrics that can evaluate the naturalness of the sentence in the future. Also, we would like to think about using the knowledge-based method to improve the text generation method and expand our method from image captioning to other kinds of data captioning.

REFERENCES

[1] Kelvin Xu, Jimmy Lei 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. In 32nd International Conference on Machine Learning, ICML 2015, 2048–2057.

[2] Songtao Ding, Shiru Qu, Yuling Xi, and Shaohua Wan. 2020. Stimulus-driven and concept-driven analysis for image caption generation. Neurocomputing 398, (2020), 520–530. DOI:https://doi.org/10.1016/j.neucom.2019.04.095

[3] Songtao Ding, Shiru Qu, Yuling Xi, Arun Kumar Sangaiah, and Shaohua Wan. 2019. Image caption generation with high-level image features. Pattern Recognit. Lett. 123, (2019), 89–95. DOI:https://doi.org/10.1016/j.patrec.2019.03.021

[4] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv: 1904.09675. Retrieved from https://arxiv.org/abs/1904.09675 .

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In Acl, 311–318. DOI:https://doi.org/10.3115/1073083.1073135

[6] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 4566–4575. DOI:https://doi.org/10.1109/CVPR.2015.7299087

[7] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. (2014). arXiv:1411.1784. Retrieved from http://arxiv.org/abs/1411.1784

[8] Tong Che, Yanran Li, Ruixiang Zhang, R. Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. 2017. Maximum-likelihood augmented discrete generative adversarial networks. arXiv:1702.07983. Retrieved from http://arxiv.org/abs/1702.07983

[9] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.

[10] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 4651–4659. DOI:https://doi.org/10.1109/CVPR.2016.503

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog, 1(8), p.9.

[12] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In 34th International Conference on Machine Learning, ICML 2017, 2029–2042.

[13] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3156–3164. DOI:https://doi.org/10.1109/CVPR.2015.7298935

[14] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6964-6974).

[15] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. (2016). arXiv:1610.02424. Retrieved from http://arxiv.org/abs/1610.02424

[16] Moitreya Chatterjee and Alexander G. Schwing. 2018. Diverse and coherent paragraph generation from images. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 729-744).

[17] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In Proceedings of the IEEE International Conference on Computer Vision, 4155–4164. DOI:https://doi.org/10.1109/ICCV.2017.445

[18] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 6274–6283. DOI:https://doi.org/10.1109/CVPR.2019.00644

[19] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 6077–6086. DOI:https://doi.org/10.1109/CVPR.2018.00636

[20] Sean Welleck, Kianté Brantley, Hal Daumé, and Kyunghyun Cho. 2019. Non-monotonic sequential text generation. In 36th International Conference on Machine Learning, ICML 2019, 11656–11676.

[21] N. Komal Kumar, D. Vigneswari, A. Mohan, K. Laxman, and J. Yuvaraj. 2019. Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach. In 2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019, 107–109. DOI:https://doi.org/10.1109/ICACCS.2019.8728516

[22] Bahdanau Dzmitry, Cho Kyunghyun, and Bengio Yoshua. 2014. Neural machine translation by jointly learning to align and translate. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. (2014), 1–15. arXiv:1409.0473. Retrieved from http://arxiv.org/abs/1409.0473

[23] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence generative adversarial nets with policy gradient. In 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 2852–2858.

[24] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 8693 LNCS, PART 5 (2014), 740–755. DOI:https://doi.org/10.1007/978-3-319-10602-1_48

[25] Dianqi Li, Qiuyuan Huang, Xiaodong He, Lei Zhang, and Ming Ting Sun. 2018. Generating diverse and accurate visual captions by comparative adversarial learning. arXiv preprint arXiv:1804.00861. Retrieved from http://arxiv.org/abs/1804.00861

[26] Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 664–676. DOI:https://doi.org/10.1109/TPAMI.2016.2598339

[27] Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. ACL-05 - 43rd Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (2005), 531–540. DOI:https://doi.org/10.3115/1219840.1219906

[28] Harald Hammarström. 2016. Linguistic diversity and language evolution. *J. Lang. Evol.*1, 1 (2016), 19–29. DOI:https://doi.org/10.1093/jole/lzw002