

Article

Variational Autoencoder-Based Multiple Image Captioning Using a Caption Attention Map

Boeun Kim , Saim Shin and Hyedong Jung

Artificial Intelligence Research Center, Korea Electronics Technology Institute, Seongnam-si, Gyeonggi-do 13509, Korea

* Correspondence: kbe36@keti.re.kr

Received: 31 May 2019; Accepted: 29 June 2019; Published: 2 July 2019



Abstract: Image captioning is a promising research topic that is applicable to services that search for desired content in a large amount of video data and a situation explanation service for visually impaired people. Previous research on image captioning has been focused on generating one caption per image. However, to increase usability in applications, it is necessary to generate several different captions that contain various representations for an image. We propose a method to generate multiple captions using a variational autoencoder, which is one of the generative models. Because an image feature plays an important role when generating captions, a method to extract a Caption Attention Map (CAM) of the image is proposed, and CAMs are projected to a latent distribution. In addition, methods for the evaluation of multiple image captioning tasks are proposed that have not yet been actively researched. The proposed model outperforms in the aspect of diversity compared with the base model when the accuracy is comparable. Moreover, it is verified that the model using CAM generates detailed captions describing various content in the image.

Keywords: image captioning; multiple image captioning; visual attention; caption attention map; variational autoencoder

1. Introduction

An image captioning system able to describe an input image as a sentence has many potential applications. It could be used for services that explain situations to visually impaired people and to search for specific scenes within a large amount of image or video content. Recently, image captioning using deep learning has been actively studied in the field of computer vision. These studies have focused on producing one caption per image. Most popular studies have used a convolutional neural network (CNN) as an image encoder and a recurrent neural network (RNN) as a module for generating sentences [1–4]. There has been an effort to improve the accuracy of captioning models and the approaches can be grouped into two paradigms. Bottom-up methods use semantic concepts [3,5,6] and top-down methods include attention based models [4,7,8].

In commercial services, it is important to make multiple captions for a single image, which captions should contain various expressions and describe various characteristics of the image content. When users desire to search a content clip using a description sentence, the proposed model allows more detailed searches because various representations can be applied by each user. Early models of image captioning are limited because they only generate one safe caption from each image. If the captioning system generates descriptions focused not only on the whole image, but also on the various characteristics of the image, a search is more likely to be successful. This is true even if the content that the user wants to search is in only a part of the image. In addition, diverse descriptions generated by a service for the visually impaired or an automatic system for subtitle generation in videos would help to diversify the expressions that the user encounters. A system able to generate diverse captions

is valuable because the situations faced by visually impaired users is not so varied and scenes with similar backgrounds and situations appear repeatedly in video content such as movies, dramas, CCTV, and vehicle black boxes.

In this paper, a method to generate multiple captions for a single image using the variational autoencoder (VAE) [9] structure and image attention information is proposed. Previous studies on generating captions using VAE deal with sentence information. They transformed the last hidden state of the RNN, the sentence embedding, to latent representation [10–12]. When generating captions, variation is given to the latent variable to create sentences of various styles. However, the methods only focused on the sentence generation problem and they did not use image information. In the image captioning task, information about images plays an important role. We use image related features in the VAE-based captioning model and have verified that the network generates more diverse captions than those in previous studies. Image-caption pairs in the dataset give information about the image regions to which the captions are related and this was applied to our model. There are not yet many studies about making multiple captions from a single image, so there are no popular or established methods for diversity evaluation at this time. Therefore, we propose an evaluation procedure that contains sampling methods and metrics that refer to previous papers [11,13]. The main contributions of our paper are as follows:

- A method for extracting caption attention maps (CAMs) from an image–sentence pair is proposed.
- A captioning model is proposed that generates multiple captions per image by adding randomness to the image region-related information using VAE.
- A methodology is introduced for evaluation of multiple image captioning tasks, which has not been done in previous studies, and it is used to evaluate our base and proposed model.

2. Related Works

In the field of image captioning, research has been mainly focused on improving the accuracy of the captioning system. Recently, studies have been conducted to generate diverse captions. A traditional captioning system converges to generate a description suitable for multiple similar images after training, resulting in general and safe captions. As a result, the diversity of generated captions is reduced, often resulting in the same explanations for similar but different images.

Diverse captioning was designed to generate different captions for different images and to generate novel sentences not included in the training set. To generate diverse captions, popular generative models such as VAEs [9] and generative adversarial networks (GANs) [14] were used in previous work. Dai et al. [15] constructed a conditional GAN model that adds an image as a condition. The generator produced captions and the discriminator distinguished the captions generated from the natural ones. After learning the generator and discriminator, the researchers separated the evaluator and proposed the E-GAN metric, which is a natural sentence evaluation scale. Shetty et al. [13] also proposed the GAN model, which enabled back propagation in categorical distribution by selecting words in long short-term memory (LSTM) using Gumbel softmax. The criterion score in the discriminator included not only image-to-sentence distances but also sentence-to-sentence distances for distinguishing real from fake. This scheme induced large variation between sentences.

In this paper, a system is proposed that generates multiple captions using VAE. Such VAEs have advantages, such as stable training and generating meaningful latent representations that enable us to derive the desired results. Wang et al. [10] proposed a Conditional VAE model that takes image content such as objects, as conditions for generating sentences associated with them. Two priors, Gaussian mixture model (GMM) and average Gaussian (AG), were used to determine how latent space was modeled. Experimental results showed that diversity and accuracy were both improved in relation to conventional LSTM systems and that the model with AG was more diverse and controllable than the one with GMM. In addition to captioning, there were examples of applying VAE to visual question and answering (VQA) tasks [11] and to paragraph generation [12]. Jain et al. [11] projected the hidden states of the last step of the encoder LSTMs to the latent space and succeeded in generating novel

questions that were diverse and not in the training set. Chatterjee et al. [12] proposed a model for generating paragraphs using the Stanford image paragraph data set. The proposed VAE-based model consists of a topic generation network and a sentence generator network that generates each sentence. As in earlier work [11], the hidden state vector of the last step of the sentence generator network RNN model was also transformed to latent space. In previous methods, training was used to construct a latent distribution of sentence embedding information, and variation was added to sentence style when generating the captions.

In the image captioning task, not only the sentence information, but also the image information plays an important role. However, this was not reflected in the previous works. In the work reported in this paper, we use this image information to generate multiple captions that describe various content in the image. There has been research on generating different descriptions from the training set, and methods of generating different captions in similar images. However, there are not many pieces of research on generating several different sentences for one image; therefore, there was no existing relevant evaluation standard or popular evaluation method. Hence, an evaluation method is proposed for the captioning of multiple images.

3. Multiple Image Captioning

In this paper, a VAE-based multiple image captioning system is proposed that uses CAMs. The method generates multiple captions describing diverse content in each image and includes a variety of expressions. The caption generation model VAE [9] was applied to the network, which consists of an encoder, a decoder, and a latent space. At the VAE encoder, the CAM containing the region information of an image is extracted and transformed to a latent representation. Variables related to caption attention are randomly sampled in the latent space to produce various sentences. At the VAE decoder, a description is generated by inserting the variable as well as an image feature as the condition. Before discussing the proposed method, we give a background description of VAE. Then, we represent the extraction of the CAM from the encoder model. After that, our whole model structure and sampling method are described.

3.1. Background on VAE

VAE consists of two neural network modules, encoder, and decoder, for learning the probability distributions of data. The encoder creates a latent variable z from raw data x and transforms it into latent space. The decoder plays the role of recovering x using z extracted from the latent space. Let $q(z|x)$ and $p(x|z)$ be the probability distributions of the encoder and the decoder, respectively. Through training, we obtain the parameter maximizing marginal likelihood $\log p(x)$. Expanding the equation and finding the evidence lower bound (ELBO) yields:

$$\log p(x) \geq E_{z \sim q(z|x)} [\log p(x|z)] - D_{KL}(q(z|x)||p(z)) = ELBO. \quad (1)$$

Because the objective is to maximize the ELBO on the right side, the loss function can be written as Equation (2):

$$Loss = -E_{z \sim q(z|x)} [\log p(x|z)] + D_{KL}(q(z|x)||p(z)). \quad (2)$$

" $-E_{z \sim q(z|x)} [\log p(x|z)]$ " is the reconstruction error and " $D_{KL}(q(z|x)||p(z))$ " is the KL divergence regularization term. Because posterior $q(z|x)$ is intractable, the variational inference method is applied to approximate the easier-to-handle distribution, for example, the Gaussian distribution. The variational reasoning computes the KL divergence between $q(z|x)$ and $p(z)$ and updates $q(z|x)$ in the direction in which the value of KL divergence decreases. The latent variable distribution $p(z)$ is approximated to a computable function, usually the standard normal distribution $N(0, I)$. In the neural network model, the feature is extracted from the VAE encoder and then the mean μ and variance σ^2 values are obtained through the linear layer. We draw a sample from $z \sim N(0, I)$ for training, shift and scale it using $z = \mu + \sigma\epsilon$, with the condition $\epsilon \sim N(0, I)$, and pass it to the decoder.

3.2. Caption Attention Map

The CAM is defined as a vector representing the weight value of how much the corresponding part of the image is related to a sentence about that part. The information that can be obtained from the image-caption set is an attention area in the image that the sentence describes. Intermediate results from the model in the paper [8] are used to induce a CAM. We refer to this method as KWL after the title of the paper “Knowing When to Look” [8]. The KWL derives the attention area associated with each word. We combined this information to provide attention to the entire sentence.

An example diagram of this process is represented in Figure 1.

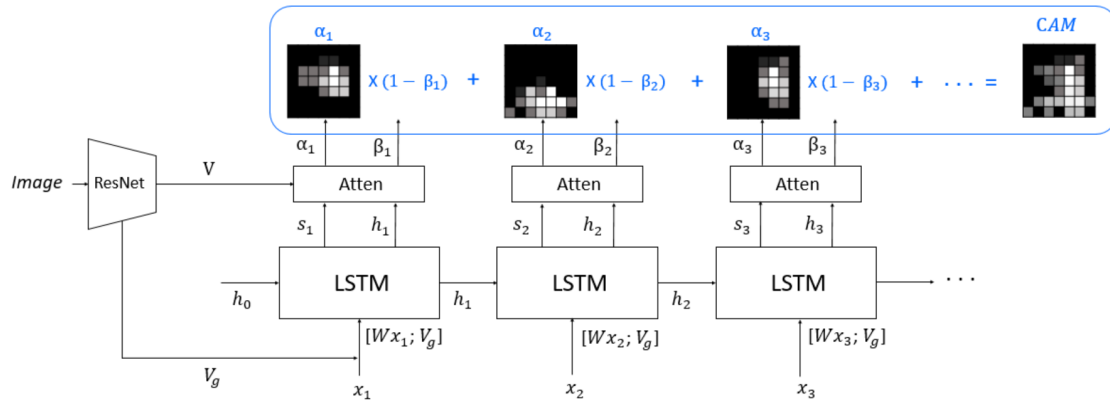


Figure 1. Illustration of the process generating a caption attention map (CAM). An attention module for each step derives attention weight α and visual grounding $(1 - \beta)$ [8]. The calculation method of the CAM is described in the blue box, which is the weighted sum of α and $(1 - \beta)$.

The attention module in the model that is attached after each step of the RNN highlights the image area and effects when predicting words. An image is divided into k areas: 7×7 , for a total of 49 areas. Following [8], $A = \{a^1, \dots, a^k\}$, $a^i \in R^{2048}$ is a spatial image features extracted from ResNet at each of the k grid locations. A global image feature can be obtained by: $a_g = \frac{1}{k} \sum_{i=1}^k a^i$. Single layer perceptron and rectifier were used to transform the image feature vectors a^i and a_g into new vectors v_i and V_g with dimension d . The transformed spatial image feature form $V = [v_1, \dots, v_k]$. The V is inserted to attention modules and the V_g is concatenated with the term $W_i x_i$ and inserted into each step of LSTM, where x_i and W_i represent an i th word of the sentence and a weight matrix, respectively. h_t denotes the hidden state of each time step t in LSTM. The LSTM memory stores both long and short term visual and linguistic information. For each time step t , a visual sentinel s_t provides a latent representation of what the decoder already knows. $\alpha \in R^k$ is the attention weight over features in V . Sentinel gate β_t represents the weight of whether to attend the image or to attend the s_t when deriving the word of the step. The β_t is derived from the s_t , the h_t , and an image feature in an attention module, and its value is between 0 and 1. The smaller this value is, the higher the weight that is placed on the image. Conversely, the larger the value of the visual grounding $(1 - \beta_t)$ probability, the more information is obtained from the image.

The α_t and β_t are extracted within an attention module for each step of LSTM. Then, the weighted sum of α_t and $(1 - \beta_t)$ for words in the caption is calculated to obtain CAM. The more a word is associated with the information in an image, the larger the weight of the region corresponding to that word in the image. The $CAM = [cam_1, \dots, cam_k]$ for one image-caption set can be expressed as:

$$cam_i = \sum_{t=1}^N \alpha_{ti} * (1 - \beta_t), \quad \text{for } i \in [1, k], \quad (3)$$

where t and N denote the step index of LSTM and the total number of steps, respectively. The attention of each area is expressed as a value between 0 and 1.

Figure 2 represents examples of a generated CAM derived from each caption for images in the training data. In each image, the part corresponding to the main content described in the sentence is highlighted. In the case of (a), the diversity of the ground truth captions is low (Div-1 = 0.321). Div-1 is the diversity evaluation metric and it is described in Section 4.2). All five captions mainly describe “A woman takes a picture in front of a mirror”, and the words appearing are similar to “woman”, “picture”, “photo”, and “mirror”. Consequently, similar parts are highlighted in the resulting CAM. In the case of (b), the diversity of the ground truth captions (Div-1 = 0.791) is higher than for (a) and different CAMs are generated according to the captions from the same image. For caption (1), the part around the “jet” is highlighted. For caption (2), the wide range highlighted in the image is affected by the word “tarmac” and “airport”. Caption (3) highlights the background and character part of the airplane, while caption (4) highlights an engine and the wing on which it is mounted. The CAM helps the proposed VAE-based model to generate diverse captions according to image region characteristics.













	Image	(1)	(2)
(a)			
		A woman taking a picture of herself in a bathroom mirror.	A woman takes a photo of herself in a rest room mirror.
	(3)	(4)	(5)
(a)			
	A woman takes a photo of herself in a rest room mirror.	A woman takes a picture of herself in a bathroom mirror.	A woman takes a picture of herself in a bathroom mirror.
	Image	(1)	(2)
(b)			
		The passenger jet is being examined by airline personnel.	An airliner sits on the tarmac at a small airport.
	(3)	(4)	(5)
(b)			
	Plane on tarmac with "Regional" written on it.	A large plane has two engines on each wing.	Workers are preparing an airplane at the gate for its next flight.

Figure 2. Examples of a generated CAM derived from each caption for an image. (a) with low diversity (Div-1 = 0.321) and (b) with relatively high diversity (Div-1 = 0.791). The highlighted parts of five CAMs are more diverse in (b) than in (a).

3.3. VAE-Based Captioning Model Using CAM

The VAE structure was used to construct a network that provides diversity in captions. While previous studies provided variation in sentence style, we proposed a method to give variation in relation to image region characteristics. The KWL [8], which is one of the attention-based caption generation networks, is applied to the VAE encoder and decoder. The overall framework is described in Figure 3.

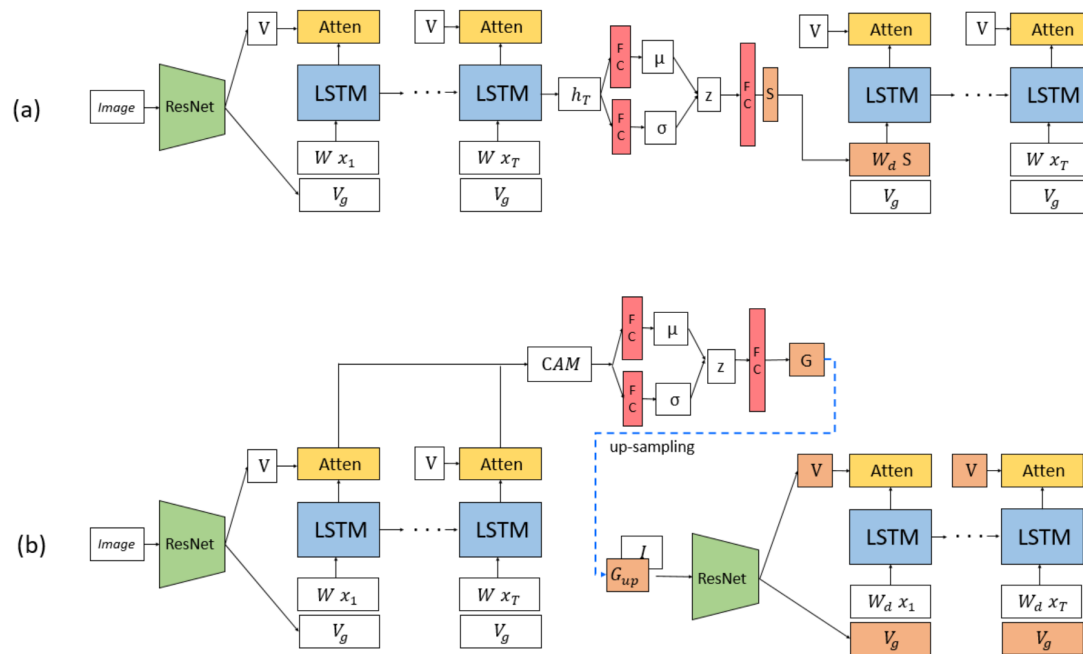


Figure 3. Illustration of variational autoencoder-based multiple image captioning models. (a) base model that projects sentence embedding to a latent distribution; (b) proposed model that projects CAM to a latent distribution.

The sentence generation model with VAE basically uses RNN as the encoder and decoder, and most simply projects the hidden state of the last step of RNN to a latent distribution [11,12,16]. The structure of the base model is represented in Figure 3a. The hidden state h_t is equivalent to the embedding of a sentence, which indicates the characteristics of the caption. h_t is transformed into a latent dimension through a linear layer to obtain mean μ and variance σ^2 , respectively. While decoding a latent variable, z (representing the sentence style) is sampled from the latent space. Different samples affect the orange shaded vectors to induce the generation of different sentences. z is projected to vector s via a fully connected layer. The size of s is the same as the word embedding size, and it is inserted into the first step of the LSTM after being multiplied by a weight. Jain et al. [11] used hidden state mapping when solving the question generation problem, by setting the latent variable dimension to 20. The dimension size of z is borrowed from [11].

The proposed model transforms CAM to a latent representation. In the image captioning task, information about images plays an important role. An image and its ground-truth caption pair gives the style of the image regions that the caption describes. The structure of the proposed model is illustrated in Figure 3b. First, the 49-size CAM is extracted from the VAE encoder using the method described in Section 3.2. Next, this vector is projected to a latent space via linear layers. If we set the size of the latent variable too large or too small, training does not work properly. The dimension of 256 was chosen for μ, σ, z , whose size is reasonable for representational power. In the decoding process, a sampled latent variable z is projected as a guide map G after going through the linear layer. The vectors affected by z are shaded in orange. This 49-size guide map G is designed to apply different weights for different areas of the image. We extracted the region-guided image feature by inserting G

into the CNN with the image. Specifically, G is scaled up to G_{up} , whose size is the image input size, and then concatenated to the fourth channel of the image. The four-dimensional tensor, including this map, is inserted as an input to the image encoder, ResNet-152 [17]. The input of the first convolution layer of ResNet is changed to four-dimensional, and the remaining layers are the same as in previous work [17]. The outputs of the ResNet, image feature vector V and global image feature vector V_g can be expected to reflect the region weighting of the image.

The latent distribution is trained to be a form of the normal distribution, as described in Section 3.1. Using inference, latent variables are sampled from the distribution. Various sampling methods can be applied according to the purpose of each model. To extract several captions with different style or expressions, we should sample points that are far away from each other in the latent space. On the other hand, to improve the accuracy of the sentence, which calculates similarity using the ground truth caption, it is necessary to sample a point near the origin. The latent variable is randomly sampled to evaluate whether our model is well constructed. Jain et al. [11] sampled from the normal distribution and uniform distribution for the evaluation. Similarly, we also extract the variables from those distributions. The uniform distribution has three ranges— $(-20, 20)$, $(-30, 30)$, and $(-40, 40)$ —and they are denoted U20, U30, and U40, respectively.

4. Experiments

4.1. Experimental Setups

The proposed model was trained and evaluated using the popular MS COCO dataset [18]. The dataset contains 82,783 images for training and 40,504 for validation. Because annotations for the test set are not publicly available, we reported the results with the widely used splits [19] which contain 5000 images for validation and test, respectively.

First, the VAE encoder used the pre-trained parameters obtained by training the KWL [8] alone. The encoder parameters are fixed while training the entire VAE model. The weights of the linear layers before and after the latent space projection and that of the decoder are updated during training. The vocabulary size was 9956, the word embedding size was 256, and the hidden size was 512. Adam optimizer [9] was used with learning rate 4×10^{-4} . Early stopping was applied based on the CIDEr score. The Adam optimizer [9] was used with the learning rate of 3×10^{-3} and stopped after 15 epochs.

4.2. Multiple Image Captioning Evaluation Metrics

In this paper, a methodology for evaluation of a multiple image captioning model is proposed that contains metrics for accuracy and diversity. The purpose of this model is to create multiple captions including a variety of expressions from a single image, with each caption correctly describing the image. Therefore, it is necessary to estimate both accuracy and diversity simultaneously.

The accuracy was measured using METEOR [20], ROUGE [21], SPICE [22], and CIDEr [23], which are metrics commonly used for image captioning. The diversity of the generated caption set, S_p , is evaluated as sentence-level diversity and corpus level diversity. The sentence level diversity metric, Div-s, includes an estimate of the diversity of the sentence structures. Even when a sentence includes the same word found in other sentences, we count the sentences with different sentence structures, that is, those in which the order of words is different. The metrics for corpus level diversity, Div-1, are referred to as was done previously [13]. This metric can be used to evaluate the diversity of words and expressions used in the sentences.

- Accuracy: the result of calculating the metrics METEOR [20], ROUGE [21], SPICE [22], and CIDEr [23] between the generated caption and the ground truth captions.
- Div-s: ratio of a number of unique sentences in S_p to the number of sentences in S_p . A higher ratio is more diverse.
- Div-1: ratio of a number of unique unigrams in S_p to the number of words in S_p . A higher ratio is more diverse.

4.3. Results

For evaluation, the captions were generated by random sampling of the latent distribution and their performance was measured. We randomly extracted 100 samples of M-dimension from the standard normal and uniform distributions. Both accuracy and diversity metrics had to be calculated to evaluate the multiple captioning systems.

The accuracy, METEOR [20], ROUGE [21], SPICE [22], and CIDEr [23] metrics were measured for the 100 descriptions generated by the models. To improve the reliability of the random extraction evaluation, the sampling and evaluation processes were repeated five times and the mean value is shown in Table 1. Table 2 compares the diversity score of the base model with that of the proposed model. Div-s was measured for S_p , which contained all 100 descriptions. In the case of Div-1, other combinations of S_p should be used to match with the scale of the evaluation score [13]. Shetty [13] used the GAN network to derive five different descriptions for each image and to evaluate the diversity of these five descriptions. With this method, when the network is trained, five captions are regarded as one set, and the parameters are learned in the direction that increases the variation among the descriptions in the set. However, with the proposed method, only one caption was generated from one sampling in the network, and the performance varied depending on where the latent variable was extracted in the latent space. Therefore, it is difficult to compare this with the results of [13]. The following method was used for numerical comparison. We generated combinations of five different captions out of 100 and they were treated as S_p s. If there were fewer than five different captions, the rest of the captions were filled with the shortest sentence. In Table 2, k represents the number of images with five or more different captions out of 5000 test sets, and this is represented in the table. The max Div-1 represents $MEAN_i[MAX_s(Div - 1)]$, where i and s denote the image index and caption combination set index, respectively. The average Div-1 represents $MEAN_i[MEAN_s(Div - 1)]$.

Table 1. Experimental results of captioning accuracy with the base model and the proposed model. Both evaluation results are comparable with N1 sampling.

Sampling	Base Model				Proposed Model			
	METEOR	ROUGE	SPICE	CIDEr	METEOR	ROUGE	SPICE	CIDEr
N1	0.249	0.532	0.183	0.976	0.248	0.533	0.182	0.980
U20	0.246	0.527	0.180	0.959	0.240	0.523	0.174	0.929
U30	0.244	0.526	0.179	0.952	0.230	0.510	0.163	0.864
U40	0.244	0.524	0.178	0.948	0.218	0.493	0.149	0.780

Table 2. Experimental results of the diversity of the base model and the proposed model. The result of the proposed method is more diverse than the base model is with N1 sampling.

Sampling	Base Model				Proposed Model			
	Div-s	k	max Div-1	average Div-1	Div-s	k	max Div-1	average Div-1
N1	0.015	38.8	0.192	0.191	0.017	133.0	0.197	0.196
U20	0.090	3859.6	0.359	0.307	0.128	3825.75	0.387	0.317
U30	0.116	4355.8	0.391	0.319	0.201	4431.75	0.457	0.347
U40	0.133	4550.2	0.410	0.326	0.274	4721.75	0.516	0.373

As represented in Table 1, the accuracy of the proposed model and the base one were comparable with N1 sampling. Moreover, the proposed method showed better performance in the diversity evaluation with N1 sampling, as shown in Table 2. Div-s was larger with the proposed model and k increased greatly from 38 to 133. This means that the model generated more different sentences in 100 samples. Both max Div-1 and average Div-1 values were also larger with the proposed model. U20, U30, and U40 cannot be compared one by one, but the table shows that the larger the sampling

range, the lower the accuracy and the higher the diversity. For reference, we computed the Div-1 of the MS COCO data set, which consisted of captions written by humans, and the value was 0.53.

Figure 4 shows an analysis of the relationship between the diversity of content the image contains and the diversity of the captions generated. Assuming the images will tend to contain a variety of content if they include a variety of captions, the Div-1 value of the ground truth captions was used to represent the richness of the content. Div-1s were obtained for the ground truth images and the images were divided into five groups according to their Div-1 values. The divisions were as follows with Div-1 values between (0.0–0.2], (0.2–0.4], (0.4–0.6], (0.6–0.8], and (0.8–1.0]. Moreover, the caption was generated by the model for an image corresponding to each section. Then, the maximum Div-1 value was obtained by constructing five caption combinations per image in the same way as the experiment in Table 2, and the average Div-1 of the images was obtained within the interval. To reduce the random sampling error, the same experiment was repeated five times and the average value is shown. The value was plotted on the vertical axis in Figure 4. The results of the base model are represented in dashed lines and the results of the proposed model are shown in solid lines. As the image content diversity increased, the Div-1 value of the proposed model increased more rapidly than with the base model. In the case of the base model, the diversity of the generated caption increased as the diversity of the ground truth caption increased, but the incremental rate was low. Therefore, a large gap developed between the Div-1s of the generated and the ground-truth captions for the rich images. Our method modeled the image region feature into a latent space. When decoding, the latent variable was calculated in conjunction with the image and gave a weighting effect to various parts of the image. Therefore, it was more affected by the image content than was the base model using only the sentence features. It is expected that, with this system, if an image contains a greater variety of content, it will generate a greater variety of captions.

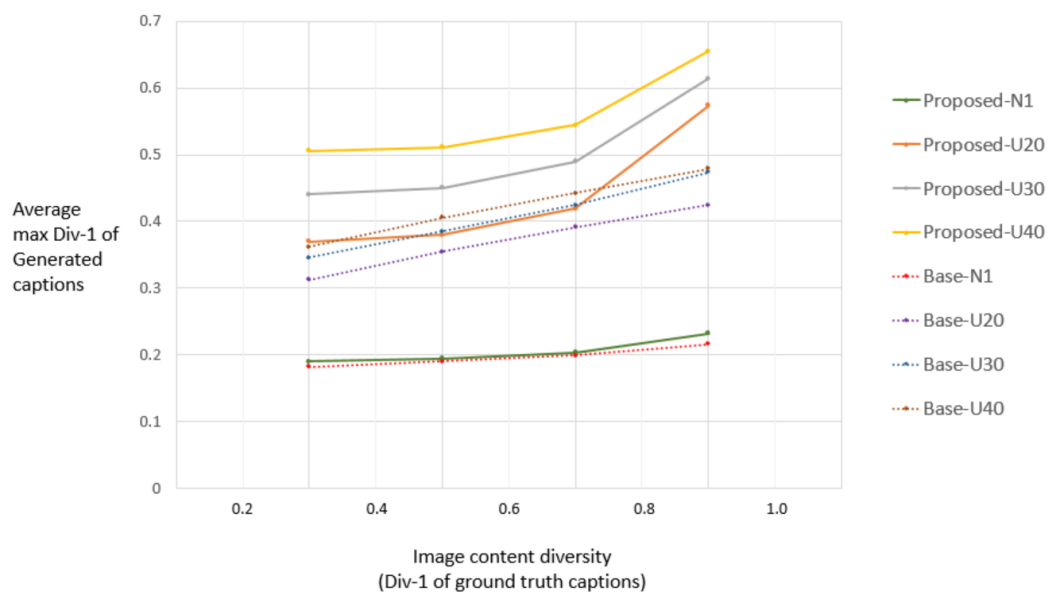


Figure 4. Representation of the relationship between the image content diversity and the diversity of captions generated. As the image-content diversity increases, the Div-1 value of the proposed model increases more rapidly than with the base model.

4.4. Discussion

Figure 5 illustrates several results from extracting various captions using the proposed network. One-hundred captions were extracted through random sampling and no duplicate sentences are represented in the figure. The second column shows the results from the U30 sampling of the base model, and the last column shows the results of the U20 sampling of the proposed model. For a

reasonable comparison, we adopted Base-Model-U30 as a comparison target of which the accuracy and the values related to Div-1 are higher than with the Proposed-Model-U20: the values evaluated with the accuracy metrics METEOR, ROUGE, SPICE, and CIDEr, and related Div-1 metrics k, max Div-1, and average Div-1 are higher. For the measure of Div-s, Proposed-Model-U30 was higher than Base-Model-U30; therefore, more different captions were generated. Captions extracted from the proposed model included more diverse expressions. An image information CAM was used in the proposed model; therefore, the generated captions were related to the different characteristics of the images. In image (1), most of the captions in (a) indicate that the “bus is parked”. Parking places vary: “in front of a bus”, “parking lot”, and “side of the road”. In (b), besides the information about the position of the bus, captions about the graffiti drawn on the bus are also generated. In image (2), the captions in (c) state that there are plates, forks, etc. on the table. In contrast, various expressions such as “slice of”, “bite taken”, “half eaten”, and “small piece” are included in the captions in (d).



	Base Model – U30	Proposed Model - U20
	(a)	(b)
(1)	 <ul style="list-style-type: none"> • a bus with a lot of people on it • a bus is parked in front of a bus • a bus is parked in a parking lot • a bus is parked in a lot of a parking lot • a bus is parked in a lot with a lot of people • a bus that is parked on the side of the road • a bus is parked on the side of the road 	<ul style="list-style-type: none"> • a bus with graffiti on it on a street • a bus with graffiti on it is on the street • a bus with graffiti on it on the side of the road • a bus with graffiti on it is parked on the side of the road • a bus with a mural of people on it • a bus with a mural of a bus on it • a colorful bus with graffiti on it on a street • a colorful bus with graffiti on it on the street • a bus with graffiti on it is on the side of the road
	(c)	(d)
(2)	 <ul style="list-style-type: none"> • a plate of food on a table • a plate of food that is on a table • a plate of food on a table with a fork • a plate of food with a fork and a knife • a close up of a plate of food on a table • a plate with a pastry on it and a fork • a plate with a pastry and a fork on it • a plate with a muffin and a muffin on it • a plate with a donut on it and a fork • a donut with a bite taken out of it • a plate with a muffin on it sitting on a plate 	<ul style="list-style-type: none"> • a plate with a doughnut on it • a plate with a slice of cake on it • a small plate with a slice of cake on it • a cupcake with a bite taken out of it • a doughnut with a bite taken out of it • a hot dog with a bite taken out of it • a half eaten doughnut on a plate with a fork • a half eaten doughnut sitting on a plate • a small doughnut sitting on top of a plate • a slice of cheesecake with a bite taken out of it • a piece of cake on a plate with a fork • a small piece of food on a plate

Figure 5. Comparison of the quality of multiple captions generated by the base model and the proposed model. Captions extracted from the proposed model contain more diverse expressions.

The observed failure cases are shown in Figure 6. When the captions in the training data are biased towards similar images and the captioning system is induced to generate diverse captions, generated captions converge on biased data. Despite the absence of “umbrella”, “kite”, or “man” in the image, these words appear in the captions. This is because the images in the training data set taken in the background of the beach mostly contained parasols and kites.

For further work, we plan to carry out experiments in the future after data reinforcement. It would also be useful to study a method by which extracting a caption is different from previous ones. To do this, a method for latent space modeling and extracting variables that are far apart from each other should be considered.


	Base Model – U30	Proposed Model - U20
	(a)	(b)
	<ul style="list-style-type: none"> • a beach with chairs and a boat on the beach • a boat is sitting on the beach near the water • a beach with a chair and a chair • a boat is sitting on the beach near the water • a beach with chairs and a chair on it • a boat is on the beach with a chair and chairs • a chair and a chair on a beach • a boat is sitting on the beach • a couple of chairs are on the beach • a beach with chairs and a boat on the beach 	<ul style="list-style-type: none"> • a beach with a beach chair and two chairs • a beach with a beach chair and chairs • a beach with a boat and a boat on it • a beach with a beach chair and a beach • a beach with a boat and a on it • a beach with chairs and a beach umbrella • a beach with chairs and umbrellas on the beach • a beach with a beach chair and a beach umbrella • a beach with a chair and a table with chairs • a beach with a pair of scissors and a beach chair • a couple of kites are sitting on the beach • a couple of chairs and a umbrella on a beach • a boat is sitting on the beach • a man sitting on a beach next to a kite • a man is standing on a beach with a kite • a person sitting on a beach with a kite • a couple of people are standing on a beach • a person is sitting on a bench with a kite • a couple of people sitting on top of a beach

Figure 6. Failure cases of the proposed model that generate image-independent captions.

5. Conclusions

In this paper, a network is proposed that generates multiple captions by changing the guide map that reflects the region-related style of the image. First, we obtain a CAM at the VAE encoder and the CAM vectors constitute a latent space. Next, the vector that reflects the image-attention region style is extracted from the latent space. At the VAE decoder, this vector is used as the CNN input, and acts as a condition to generate the caption. The proposed model outperforms the base model in terms of diversity when the accuracy is comparable. Using sample cases, we confirmed that captions generated by the proposed method have a greater variety of content and of expressions than the ones generated by the base model. The proposed method could be useful for many applications requiring multiple caption generation.

Author Contributions: Conceptualization, B.K.; methodology, B.K.; software, B.K.; validation, B.K.; formal analysis, B.K.; investigation, B.K.; resources, B.K.; data curation, B.K.; writing—original draft preparation, B.K.; writing—review and editing, S.S.; visualization, B.K.; supervision, S.S.; project administration, H.J.; funding acquisition, H.J.

Funding: This work was supported by IITP/MSIT (2017-0-00255, Autonomous digital companion framework and application).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
2. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
3. Wu, Q.; Shen, C.; Liu, L.; Dick, A.; Van Den Hengel, A. What value do explicit high level concepts have in vision to language problems? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 203–212.
4. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. *arXiv* **2015**, arXiv:1502.03044.

5. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.
6. Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; Deng, L. Semantic compositional networks for visual captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5630–5639.
7. Liu, C.; Mao, J.; Sha, F.; Yuille, A. Attention correctness in neural image captioning. In Proceedings of the Thirty-First, AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
8. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
9. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
10. Wang, L.; Schwing, A.; Lazebnik, S. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5756–5766.
11. Jain, U.; Zhang, Z.; Schwing, A.G. Creativity: Generating diverse questions using variational autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6485–6494.
12. Chatterjee, M.; Schwing, A.G. Diverse and Coherent Paragraph Generation from Images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 729–744.
13. Shetty, R.; Rohrbach, M.; Anne Hendricks, L.; Fritz, M.; Schiele, B. Speaking the same language: Matching machine to human captions by adversarial training. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4135–4144.
14. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
15. Dai, B.; Fidler, S.; Urtasun, R.; Lin, D. Towards diverse and natural image descriptions via a conditional gan. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2970–2979.
16. Bowman, S.R.; Vilnis, L.; Vinyals, O.; Dai, A.M.; Jozefowicz, R.; Bengio, S. Generating sentences from a continuous space. *arXiv* **2015**, arXiv:1511.06349.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
18. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft COCO captions: Data collection and evaluation server. *arXiv* **2015**, arXiv:1504.00325.
19. Karpathy, A.; Li, F. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
20. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 11 June 2005; pp. 65–72.
21. Lin, C.Y. Rouge: A Package for Automatic Evaluation of Summaries. *Text Summ. Branches Out* **2004**, 74–81. Available online: <https://www.aclweb.org/anthology/W04-1013> (accessed on 18 June 2019).
22. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 382–398.
23. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.

