applied sciences

MDPI

Article

Text Augmentation Using BERT for Image Captioning

Viktar Atliha and Dmitrij Šešok *

Department of Information Technologies, Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania; viktar.atliha@vgtu.lt

* Correspondence: dmitrij.sesok@vgtu.lt

Received: 27 July 2020; Accepted: 25 August 2020; Published: 28 August 2020

Abstract: Image captioning is an important task for improving human-computer interaction as well as for a deeper understanding of the mechanisms underlying the image description by human. In recent years, this research field has rapidly developed and a number of impressive results have been achieved. The typical models are based on a neural networks, including convolutional ones for encoding images and recurrent ones for decoding them into text. More than that, attention mechanism and transformers are actively used for boosting performance. However, even the best models have a limit in their quality with a lack of data. In order to generate a variety of descriptions of objects in different situations you need a large training set. The current commonly used datasets although rather large in terms of number of images are quite small in terms of the number of different captions per one image. We expanded the training dataset using text augmentation methods. Methods include augmentation with synonyms as a baseline and the state-of-the-art language model called Bidirectional Encoder Representations from Transformers (BERT). As a result, models that were trained on a datasets augmented show better results than that models trained on a dataset without augmentation.

Keywords: image captioning; augmentation; BERT

1. Introduction

Image captioning is the task of automatically generating a textual description of an image [1]. The goal pursued by the researchers is to make these textual descriptions as similar as possible to how a human would describe an image. Systems of such generation capability can be used to help visually impaired people improve a human-computer interactions by introducing visual concepts to a computer and create better features for information retrieval using images [2,3].

In recent years, most approaches to solving the image captioning task include using neural networks. Most of the used neural networks architectures are of an encoder-decoder type for example [4–8]. In such models, an image is first encoded to its hidden representation and then a textual description of this image is generated (decoded) based on this hidden representation. Convolutional neural networks, such as VGG [9] and ResNet [10], are most often used as encoders, because they have proven themselves in a variety of different computer vision tasks. Recurrent neural networks, such as RNN [11] or LSTM [12], are used as decoders due to their wide applicability for natural language processing tasks.

However, the training and use of recurrent neural networks is quite challenging. To preserve the information from the previous steps RNN uses a hidden state which task is to summarize and store this information in a single vector. Nevertheless, this vector usually has a rather small dimension and, as a result, a limited ability to remember information. In addition, as the length of the sequence increases the complexity of training, the model for generating a sequence of words also increases.

Such methods as attention introduced in [13] and transformers described in [14], which has proven their efficiency in solving a large set of NLP tasks, particularly sequence to sequence tasks,

allow for dealing with mentioned problems for caption generation task. The attention mechanism helps RNN to attend areas of the image that are the most important on a next word generation step. This helps to improve the quality of generated captions and gives some interpretation of the generation process. The transformer has the self-attention and multi-head attention modules, which allow you to get rid of recurrence while not losing the quality of taking into account the context and history of previously generated words.

Year after year, models for image captioning are becoming increasingly complicated. However, high-quality training of complex models (in particular neural networks) requires a sufficiently large amount of labeled data. Although such commonly used for image captioning dataset as MSCOCO [15] and Flickr30K [16] have a relatively large number of images in the training set, they still have a lack of different descriptions per one image. Each image has in average no more than five different captions. More than that some captions within the same image are quite similar and do not show all the diversity with which a person can describe an image. Therefore, many models has poor performance from the generated descriptions diversity point of view.

Problems with an insufficient number of diverse answers that a person can give in response to a specific problem also exist for other vision-language tasks. Accordingly, for example, in the main datasets [17–21] for the visual question answering problem there are only a few correct answers for each question. However, the various answers that a person could give in reality are quite large. The similar problems can be encountered when solving the visual dialog task. The datasets [22,23] for this task contain only a limited number of replicas.

To tackle this problem, we applied text augmentation methods to image captions from a MSCOCO dataset. The dataset augmentation is widely used for boosting a performance and obtaining model stability in a many machine learning tasks. The various language model can be used for a text augmentation. One of the most recent successful language models at the moment is BERT introduced in [24]. BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. Transformer consists of two separate mechanisms—an encoder that reads the text input and a decoder that produces a prediction for the task. As opposed to directional models, which read the text input sequentially, the Transformer encoder reads the entire sequence of words at once. This characteristic allows for the model to learn the context of a word based on all of its surroundings.

Augmentation of a MSCOCO dataset using BERT has helped us to speed up several of the state-of-the-art model for image captioning training and achieve better results when compared to the same model trained on a dataset without augmentation at all.

The main contributions of this paper are as follows:

- we proposed the use of augmentation of image captions in a dataset (including augmentation using BERT) to improve a solution of the image captioning problem. To our best knowledge this is the first attempt to improve quality in image captioning through captions augmentation;
- we compared different augmentation methods—using synonyms and using BERT, as well as the various increase in a dataset size on the training three state-of-the-art image captioning models and its final quality;
- intensive experiments have shown that the proposed augmentation methods improve model performance on the commonly used image captioning metrics, such as CIDEr and SPICe; and,
- the proposed methods are not limited to an image captioning task and can be used for different vision-language tasks.

2. Related Work

2.1. Image Captioning

Image captioning is a difficult task on the intersection of computer vision (CV) and natural language processing (NLP), which involves the generation of a short sentence describing the image [1].

Encoder-decoder architectures are commonly used for this task which shows decent results. The task of generating captures is a task of a sequence generation. More than that images can be treated like a sort of "visual language" so the image captioning task can be considered as a machine translate task from a "visual language" to a human language. Regarding this, many methods that have shown themselves in the machine translation problem have also been successfully adapted to solve the image captioning problem.

One of the most used methods that significantly improved the quality of basic models is a mechanism called attention [13]. As an attention special case visual attention mechanism is widely used in current state-of-the-art image captioning models. The main idea of visual attention is to allow models to selectively concentrate on objects of interest.

Another approach is to use advantages of a transformer-based architectures [14]. They are the state-of-the-art methods in sequence modeling tasks like machine translation [25] and language modeling [24]. A number of image captioning models use transformer architectures in order to improve their quality, like [26–28]. The great advantage of transformer models over other sequence generation models is that transformers use a fully connected neural network instead of a recurrent one. This simplifies model training and increases the ability of the model to take context into account. Recent studies have shown that such approaches can be also applied to the image captioning task.

2.2. Augmentation

Augmentation is a method of creating additional training data from existing dataset. There are many different augmentation techniques for data of a different nature. Augmentation showed itself well for main tasks that are associated with the analysis of structured data such as images [29] and text [30–32]. Usually augmentation methods depend on a specifics of the task however there are some general approaches. For example for image augmentation [29] horizontal flip, random crop and more are used. The various combinations of mentioned methods may also be used. Small image changes do not affect its class and its content what allows you to expand the training set with such manipulations. For the text augmentation tasks, random word deletions and insertions, synonymous replacements, and word permutations within a sentence are usually used [30].

However, augmentation is rarely used in vision-language tasks, such as image captioning, visual question answering, or visual dialog. The most of them, such as [33,34], use standard images augmentation methods described above. The others use text augmentation techniques. For example, in [35], simple augmentations, such as word permutations for captions or random words replacement, are used for better evaluation metrics design. However, there are several works about more complex augmentation methods for vision-language tasks. For example, in [36], the authors use the template based generation method that uses image annotations and a LSTM based language model for generating question-answer pairs about images.

2.3. BERT

BERT—Bidirectional Encoder Representations from Transformers [24]. The idea behind BERT learning is to teach the model to predict missed words in a sentence. In order to do this, part of the words in the sentence are replaced with a special token (MASK) and the task of the model is to predict those words by their context. More than that to increase the model quality and to teach neural network to understand the relationship between sentences at the same time it is taught to predict whether one phrase is a logical continuation of the second. As a result, BERT is pre-trained on a large set of texts in an unsupervised manner having high-quality vector representations of words and a model that can predict words by context and the presence of a connection between sentences. In addition, thanks to its architecture, BERT can be easily fine-tuned and adopted for specific tasks from the NLP field. All of these features allowed it to become an essential part of almost all state-of-the-art NLP models for today.

3. Methodology

In this section, we will describe main methods of text augmentation. Particularly, we concentrate on methods used in our research. In Section 3.1, we describe captions augmentation using a synonymous replacement that we used as a baseline method for our studies. In Section 3.1, we concentrate on an augmentation using language models, especially contextualized word embeddings and BERT.

3.1. Synonymous Augmentation

Unlike image and speech processing augmentation by making random noises into the input signal (characters) is not suitable for text augmentation. The relative order of letters and their presence in a word can significantly affect the semantic meaning of the word itself. The best method of text augmentation is to rephrase sentence as a person would do. However, this approach is very complicated due to the size of a training datasets. One of the simpler, but still qualitative method, is synonymous replacing augmentation first introduced in [30].

Let I be an image from a training set, $C = \{c_1, \ldots, c_k\}$ be a set of captions corresponding to that image where each caption is a sequence of words $c_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,l_i})$, l_i is a length of i-th caption. Additionally, fix some synonymous thesaurus T and let $T(w_{i,j}) = (s_{i,j,1}, s_{i,j,2}, \ldots, s_{i,j,m_{i,j}})$ be a list of synonyms to a word $w_{i,j}$ sorted in descending order of semantic closeness to the most frequently seen meaning of the word $w_{i,j}$, $m_{i,j}$—the number of synonyms of the word $w_{i,j}$ in the thesaurus T.

The following operation is performed in order to generate a new caption c'_i based on a c_i. Fix some probability p. For every word $w_{i,j} \in c_i$, which has synonyms in a dictionary i.e., for which $m_{i,j} > 1$ replace it with synonym with a probability p. To chose with which of the synonyms it will be replaced fix another probability q. With a probability q, replace a word with the most semantically close synonym $s_{i,j,1}$. If the word was not replaced by the first synonym replace it with the second closest one $s_{i,j,2}$ with a probability q and so on. Thus, the probability of replacing word with synonym $s_{i,j,r}$ is equal to q^r and it exponentially decreases with the semantic similarity of the synonym to the original word. This operation of replacing a word with it synonym occurs independently for each word in a sentence.

The above operation of obtaining a new caption based on an existing one is done d times, where d is called the augmentation coefficient. Accordingly, if the image has k captions after applying augmentation with a coefficient d it will have kd captions, i.e., training set is increased d times.

3.2. Contextualized Word Embeddings Augmentation

To augment with contextualized word embeddings approach similar to described in [31,32] can be used. Similarly to a synonymous replacement, let for some image I there is a set of sentences $C = \{c_1, \ldots, c_k\}$ describing that image. Each sentence is a word sequence $c_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,l_i})$. For the purpose of augmentation fix a language model LM that can predict the probability that a particular word w will occur in a certain context. More formally, consider some caption c_i and the j-th word of this caption. Let its context be the entire caption, except of the word itself, which is $c_i \setminus \{w_{i,j}\} = (w_{i,1}, w_{i,2}, \ldots, w_{i,j-1}, w_{i,j+1}, \ldots, w_{i,l_i})$. Thus, $LM(c_i, j) = P(\cdot | c_i \setminus \{w_{i,j}\})$ is a probability distribution over the words that can stand on a place j in a caption c_i taking its context into account.

The following procedure needs to be done to obtain a new caption c'_i based on an existing caption c_i using the language model. Fix the probability p that each concrete word from the caption should be replaced by another word. In order to replace the word $w_{i,j}$ with another calculate $LM(c_i, j)$. After that, generate the word $w'_{i,j} \sim LM(c_i, j)$ and take it as the next word of the new caption c'_i. Repeating the procedure for each word $w_{i,j}$ will create an augmented caption. Performing this operation of augmenting the sentence d times for each of the captions will lead to obtaining kd sentences describing the corresponding image.

Contextualized word embeddings models, such as BERT [24], DistilBERT [37], RoBERTA [38] or XLNet [39], are the most suitable for using during the procedure described above.

4. Results and Discussion

4.1. Dataset

We used MSCOCO [15], which is the largest and most used dataset for image captioning as a base dataset for performing augmentation, in order to compare the effectiveness of the augmentation methods. Its standard version consists of 82,783 training images and 40,504 validation images. There are five different captions for each of the images. For offline evaluation, we used the standard Karpathy split from [6], which is used by the most of the articles for results comparison. As a result, the final dataset consists for 113,287 images for training, 5000 images for validation, and 5000 images for testing.

After dataset augmentation, we also performed postprocessing by replacing all the words that occurred less than five times in the final dataset with a special token <UNK>. More than that, because the vast majority of captions was no more than 16 words length we truncated the words to maintain the maximal length of the caption equal to 16.

4.2. Implementation Details

We compared five different augmentation techniques for augmenting the original dataset. Augmentation was only used for a train part of the dataset. In one of the options, we did not augment the dataset at all and used this result as a baseline for comparison. In the other methods, we augmented dataset using BERT with augmentation factor d equal to 2 and 3. We also performed augmentation using synonyms with an augmentation factor equal to 2. The default value of replacement rate p was equal to 0.1 in all cases except during the studies about replacement rate influence.

The effect of described augmentations was compared while training a model [40], which is one of the state-of-the-art models with an open source code. We conducted extensive experiments to choose the best augmentation method suitable for this model. Models on all datasets were trained for 12 epochs in a regular way and than for seven epochs in a self-critical way described in [4]. For captions generation during the testing phase beam search algorithm with a beam size of 5 was used.

Additionally, we have confirmed our results on the other state-of-the-art models, such as [41,42], on the best variant of dataset chosen based on the previous experiments with [40].

For all models, we used open source code released by the authors of the corresponding papers. We used https://github.com/aimagelab/meshed-memory-transformer for [40], https://github.com/husthuaan/AoANet for [41], and https://github.com/JDAI-CV/image-captioning for [42]. For augmentation, the nlpaug library [43] was used. The models were trained and tested using the Google Cloud Platform on a cloud machine with 8 CPU cores, 30 GB operative memory and two Tesla K80 GPUs.

4.3. Results

To compare the results we used BLEU [44], METEOR [45], ROUGE-L [46], CIDEr [47], and SPICE [48], which are widely used for image captioning models comparison. BLEU is metrics which is widely used for machine translation task. It uses n-grams precision to calculate a similarity score between reference and generated sentences. METEOR is also a metric that is based on n-grams, but it uses synonym matching functions along with exact word matching. ROUGE-L is based on a longest common subsequence statistics. The longest common subsequence problem takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically. We especially focused on a CIDEr and SPICE metrics, as they are a human consensus metrics. CIDEr measures the similarity of the captions generated by the model and the captions that created by a person using n-grams and performing TF-IDF weighting each of them. SPICE is based on a semantics graph parsing is used to determine how well model has captured the attributes of objects and their relationships.

It is important to note that all of the models trained based on their open source code show slightly worse result than that reported in corresponding papers.

In Figure 1, you can see a comparison of augmentation using BERT with $d = 2$ for different values of replacement rate p. It can be seen that, with a large value of p equal to 0.5 model performs worse than with smaller value equal to 0.1. This may be due to the fact that too much information is lost regarding the original caption and therefore the sentences become less grammatically correct and less human-like. On the other hand, model with a value of 0.1 performs better than all others. Such small changes adds some variety to the dataset without a damage to meaning and grammatical structure.

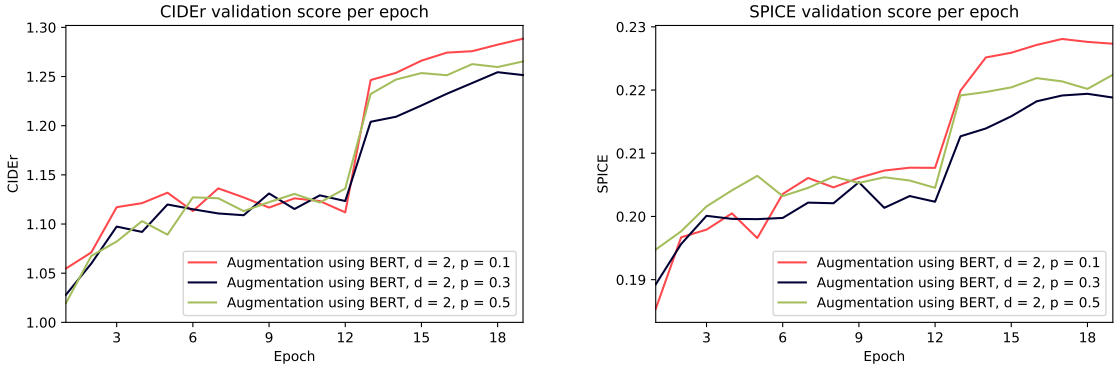

Figure 1. Validation scores per epochs during a training phase on a datasets with Bidirectional Encoder Representations from Transformers (BERT) augmentation with $d = 1$ and $p = 0.1, 0.3, 0.6$.

The comparison of models trained using various augmentation techniques—synonyms and BERT—with a model trained on an original dataset is shown in Figure 2. The model trained on a dataset with synonymous augmentation is slightly better that the original model. A model that was trained on a dataset augmented with BERT shows significantly better results than both original model and the model trained on a synonymous augmented dataset.

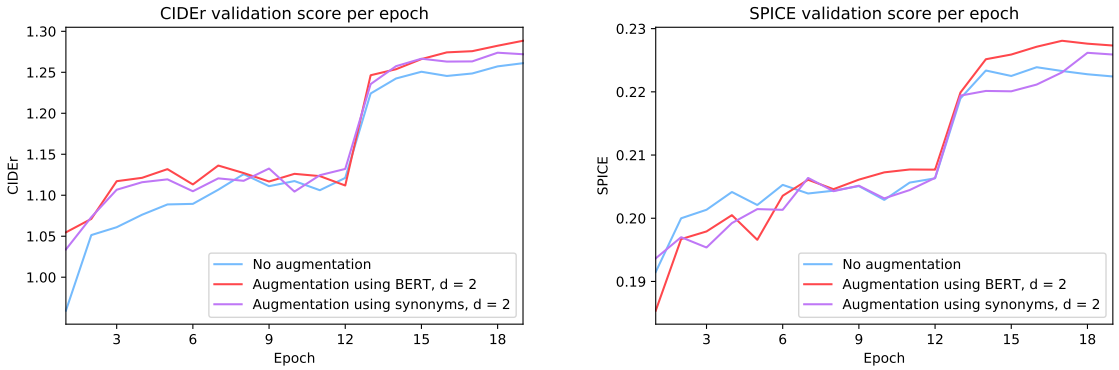

Figure 2. Validation scores per epochs during a training phase on a dataset without augmentation, on a dataset with BERT augmentation with $d = 2$ and on a dataset with synonyms augmentation with $d = 2$.

Models that are trained on a BERT augmented dataset with $d = 2$ and $d = 3$ and $p = 0.1$ (as the more promising replacement rate) are compared on a Figure 3. Comparison shows that the training on the more augmented dataset don't increases model performance. With a three-times increase in dataset size the model trains worse than with a two-times increase. This shows some boundaries of the proposed method and that the quality do not increase with the significant increase of the dataset more than two-times.

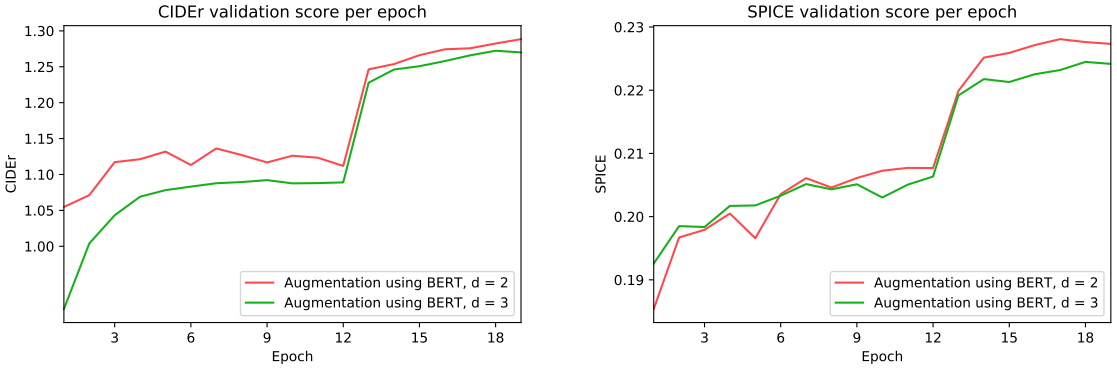

Figure 3. Validation scores per epochs during a training phase oon a dataset with BERT augmentation with $d = 2$ and $d = 5$.

Table 1 summarizes the final test scores for all of the trained [40] models. The model trained on the two-times increased dataset obtained using BERT with $p = 0.1$ augmentation shows the best results in almost all of the metrics significantly exceeding the model trained on an original dataset by 2.7 points for CIDEr and 0.2 points for SPICE. This proves the applicability of the proposed augmentation method for models designed for image captioning task quality improvement. Augmentation can be widely used to increase the quality of existing state-of-the-art approaches without any changes to that models.

Table 1. Evaluation results of the [40] model trained with our augmentation methods.

Augmentation	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
No augmentation	**38.3** [1]	28.6	58.0	126.1	22.6
BERT, $d = 2, p = 0.1$	37.8	**28.9**	**58.3**	**128.8**	**22.8**
BERT, $d = 2, p = 0.3$	37.3	28.5	57.8	125.4	21.9
BERT, $d = 2, p = 0.5$	37.3	28.8	58.0	126.5	22.3
BERT, $d = 3$	37.9	28.6	57.9	127.2	22.4
Synonyms, $d = 2$	37.7	28.7	57.8	127.4	22.2

[1] The best result for each metric is marked with bold.

Table 2 summarizes the results for all three models trained with BERT augmentation with $d = 2$ and $p = 0.1$. All of the models trained with augmentation show better results than the corresponding models trained without augmentation. This confirms our conclusions about benefits of the proposed augmentation usage for training state-of-the-art image captioning models.

Table 2. Evaluation results of all of three tested models trained with our augmentation methods.

	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
M^2 Transformers [40], no augmentation	38.3	28.6	58.0	126.1	22.6
M^2 Transformers, augmentation	37.8	28.9	58.3	128.8	22.8
AoANet [41], no augmentation	38.2	28.6	57.9	125.9	22.3
AoANet, augmentation	37.6	28.8	58.0	128.4	22.5
X-LAN [42], no augmentation	38.8	28.9	58.2	126.1	22.8
X-LAN, augmentation	37.9	29.2	58.3	128.6	22.9

4.4. Quantitative Analysis

We selected several examples of captions generated by the resulting models on a test data. They are presented in Figure 4. Here, "Ground truth" denotes a real human-generated capture from a test set used in a quality measuring. "Original" denotes a caption that is generated by a model trained on an original dataset without augmentation. We can see that augmentation helps the model trained

on augmented datasets to construct more elegant and rich sentences than the model trained on an original dataset.

Original: a fire hydrant on the side of a street
BERT, d = 2, p = 0.1: a black and silver fire hydrant on the side of a street
BERT, d = 2, p = 0.3: a fire hydrant on the side of a street with orange cones
BERT, d = 2, p = 0.5: a fire hydrant on the side of a street with two cones
Synonyms, d = 2: a fire hydrant on the side of a street with a cones
Ground truth: A bicycle is lying on the sidewalk beside a fire hydrant.

Original: a small airplane is parked on the runway
BERT, d = 2, p = 0.1: the front of a propeller airplane on a runway street
BERT, d = 2, p = 0.3: a small plane is sitting on the runway
BERT, d = 2, p = 0.5: an airplane is sitting on the runway at the airport
Synonyms, d = 2: a small plane is parked on the runway cones
Ground truth: A small airplane taking off from an airport runway.

Figure 4. Examples of captions generated by the proposed models.

Additionally, some examples of augmentation of the original captions are presented in Figure 5. Here, we can see that both augmentation methods using synonyms and BERT can diversify the captions adding a model potential to learn more complex and general ideas about textual description of the image. Additionally, since augmentation does not take into account the content of the image itself sometimes the augmented captions do not reflect the essence on the image well enough. This simulates the noise that may be present in the descriptions that are created by humans. Although augmentation is not perfect, in general captions are similar to the ground truth ones.

Original: A restaurant has modern wooden tables and chairs
BERT: Luxurious restaurant with modern wooden tables and chairs
BERT: A cafe has modern wooden walls and chairs
Synonyms: A dining room has modern wooden form and chairs
Synonyms: A restoration has modern wooden tables and directors

Original: A large bus and some people on the street
BERT: One city bus and fifteen people on the street
BERT: Another large bus carrying some people crowded the street
Synonyms: A enormous bus involved some people requests the street
Synonyms: A widespread minibus and some people on the highways

Figure 5. Examples of proposed captions augmentation.

5. Conclusions

In this work, we proposed the use of augmentation of image captions in a dataset using synonyms and contextualized word embeddings. Comparison of the results achieved by the models during training on augmented datasets based on the MSCOCO dataset showed that the proposed augmentation methods improve the quality of models for solving the image captioning problem. It has also been shown that augmentation with contextualized word embeddings helps more than with a synonymous replacement. In addition, the larger than two-times increased dataset after

augmentation do not improve the results of the models. This may indicate the limits of the proposed augmentation methods.

Despite the good results, it is worth noting that the captions generated by proposed augmentation methods cannot completely replace human ones. In addition we augment the captions only at the word level that limits the structure of augmented sentences (in particular, the number of words). Each word may not have as many words that can be used instead in a particular context. This can also make it difficult to automatically generate new captions for a training dataset.

In future works, other augmentation methods that work at the sentence level can be explored or text paraphrasing methods can be used for the same purpose. Additionally, the applicability of the proposed methods for the visual question answering task can be studied.

Author Contributions: Conceptualization, methodology, software, writing—original draft preparation, visualization, investigation, editing V.A.; writing—review, supervision, project administration, funding acquisition D.Š. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Google Cloud Platform Education Programs Grant.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Staniūtė, R.; Šešok, D. A Systematic Literature Review on Image Captioning. *Appl. Sci.* **2019**, *9*, 2024. [CrossRef]
2. Zafar, B.; Ashraf, R.; Ali, N.; Iqbal, M.K.; Sajid, M.; Dar, S.H.; Ratyal, N.I. A novel discriminating and relative global spatial image representation with applications in CBIR. *Appl. Sci.* **2018**, *8*, 2242. [CrossRef]
3. Belalia, A.; Belloulata, K.; Kpalma, K. Region-based image retrieval in the compressed domain using shape-adaptive DCT. *Multimed. Tools Appl.* **2016**, *75*, 10175–10199. [CrossRef]
4. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
5. Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; Weston, J. Engaging image captioning via personality. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 12516–12526.
6. Karpathy, A.; Li, F.-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.
7. 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.
8. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 2015 International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
10. 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.
11. Mikolov, T.; Karafiát, M.; Burget, L.; Černockỳ, J.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010.
12. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
13. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2015**, arXiv:1409.0473.

14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
15. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
16. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
17. Ren, M.; Kiros, R.; Zemel, R. Exploring models and data for image question answering. In Proceedings of the 2015 Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2953–2961.
18. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
19. Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; Xu, W. Are you talking to a machine? dataset and methods for multilingual image question. In Proceedings of the 2015 Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2296–2304.
20. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
21. Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2901–2910.
22. Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J.M.; Parikh, D.; Batra, D. Visual dialog. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 326–335.
23. Jain, U.; Lazebnik, S.; Schwing, A.G. Two can play this game: Visual dialog with discriminative question generation and answering. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5754–5763.
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019.
25. Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding back-translation at scale. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018.
26. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled Transformer for Image Captioning. In Proceedings of the 2019 IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8928–8937.
27. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [CrossRef]
28. Zhu, X.; Li, L.; Liu, J.; Peng, H.; Niu, X. Captioning transformer with stacked attention modules. *Appl. Sci.* **2018**, *8*, 739. [CrossRef]
29. Wang, J.; Perez, L. The effectiveness of data augmentation in image classification using deep learning. *arXiv* **2017**, arXiv:1712.04621.
30. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In Proceedings of the 2015 Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 649–657.
31. Fadaee, M.; Bisazza, A.; Monz, C. Data augmentation for low-resource neural machine translation. *arXiv* **2017**, arXiv:1705.00440.
32. Kobayashi, S. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv* **2018**, arXiv:1805.06201.

33. Wang, C.; Yang, H.; Bartz, C.; Meinel, C. Image captioning with deep bidirectional LSTMs. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 988–997.
34. Wang, C.; Yang, H.; Meinel, C. Image captioning with deep bidirectional LSTMs and multi-task learning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 1–20. [CrossRef]
35. Cui, Y.; Yang, G.; Veit, A.; Huang, X.; Belongie, S. Learning to evaluate image captioning. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5804–5812.
36. Kafle, K.; Yousefhussien, M.; Kanan, C. Data augmentation for visual question answering. In Proceedings of the 10th International Conference on Natural Language Generation, Santiago de Compostela, Spain, 4–7 September 2017; pp. 198–202.
37. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
38. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTA: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
39. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 5754–5764.
40. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
41. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the 2019 IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4634–4643.
42. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-Linear Attention Networks for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10971–10980.
43. Ma, E. NLP Augmentation. 2019. Available online: https://github.com/makcedward/nlpaug (accessed on 1 August 2020).
44. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
45. 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, 29 June 2005; pp. 65–72.
46. Lin, C.Y.; Och, F.J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004; p. 605.
47. 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.
48. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 382–398.