공학석사 학위논문

Smart Random Erasing for Image Captioning

이미지 캡셔닝을 위한 스마트 랜덤이레이징 데이터 증강 기법

2020 년 12 월

서울대학교 대학원

컴퓨터공학부

김 연 우

Smart Random Erasing for Image Captioning

이미지 캡셔닝을 위한 스마트 랜덤이레이징 데이터 증강 기법

지도 교수 이 상 구

이 논문을 공학석사 학위논문으로 제출함
2020 년 12 월

서울대학교 대학원
컴퓨터공학부
김 연 우

김연우의 공학석사 학위논문을 인준함
2020 년 12 월

위 원 장 김 형 주 (인)

부위원장 이 상 구 (인)

위 원 문 봉 기 (인)

Abstract

Image captioning is a task in machine learning that aims to automatically generate a natural language description of a given image. It is considered a crucial task because of its broad applications and the fact that it is a bridge between computer vision and natural language processing.

However, image-caption paired dataset is restricted in both quantity and diversity, which is essential when training a supervised model. Various approaches have been made including semi-supervised and unsupervised learning, but the result is still far from that of supervised approach. While data augmentation can be the solution for data deficiency in the field, existing data augmentation techniques are often designed for image classification tasks and are not suitable for image captioning tasks.

Thus, in this paper, we introduce a new data augmentation technique designed for image captioning. The proposed Smart Random Erasing (SRE) is inspired from the Random Erasing augmentation technique, and it complements the drawbacks of Random Erasing to achieve the best performance boost when applied

to image captioning. We also derive idea from AutoAugment to automatically search optimal hyperparameters via reinforcement learning. This study shows better results than the traditional augmentation techniques and the state-of-the-art augmentation technique RandAugment when applied to image captioning tasks.

Keywords: Image captioning, Data augmentation, Random erasing, Cutout, Reinforcement learning

Student Number: 2019-27584

Contents

Table Contents

Figure Contents

Chapter 1. Introduction

Image captioning is a task in machine learning that aims to automatically generate a natural language description of a given image. It has long been a topic of interest and has been considered significant because of its broad applications. Applications of image captioning include audio descriptions for visually impaired, improvement in search engine, and development of general cognitive architecture (artificial intelligence). Image captioning is also intriguing because it connects the two major fields in machine learning: computer vision and natural language processing.

Yet, a lot of research papers address lack of image captioning data and difficulty of gathering high-quality dataset[1, 2, 3]. Various approaches have been made to overcome this shortage of data, including recent studies in semi-supervised and unsupervised image captioning[3 - 6]. While the subject itself is worth exploring, it has not yet been able to outperform supervised models, and is still in need of further research.

Another possible approach to alleviate data shortage is data augmentation. Besides traditional image data augmentation

techniques including random crop and random horizontal flip[7 -10], advanced techniques such as RandAugment[11] have been developed to effectively augment image data. However, majority of these techniques are designed for image classification tasks and are not suited for image captioning. Later in Chapter 4, we experimentally show that existing image data augmentation techniques have little effect when applied to the field of image captioning.

This is because different data augmentation technique is needed for different data domain. Cubuk et al.[12] notes that data augmentation is used to teach a model about invariance in the data domain. In addition, Jackson et al.[13] states in his paper that style augmentation worsens accuracy on ImageNet as texture correlates strongly with the class label in ImageNet, and style augmentation removes this correlation. This implies that data augmentation techniques cannot be applied universally to different tasks and data domains.

Thus, in this paper, we develop a new image data augmentation technique, designed for the problem of image captioning, called Smart Random Erasing. We suggest several improvements upon Random Erasing[14], in order to develop a novel

data augmentation technique that is suitable for image captioning. With the intuition that objects are the most critical information in a given image with regards to image captioning, SRE detects objects inside the image to apply object-aware augmentation. Moreover, Smart Random Erasing (SRE) automatically finds the proper hyperparameters for different datasets via reinforcement learning. We evaluate our algorithm on two public benchmarks: Microsoft COCO[24] and Flickr30k[25]. Smart Random Erasing technique shows higher performance boost compared to the current state-of-the-art data augmentation technique RandAugment[11], as well as traditional augmentation techniques such as flip and crop.

Chapter 2. Related Work

2.1 Image Captioning Models

The most commonly used framework in image captioning is the encoder-decoder model[15]. Encoder-decoder model consists of a Convolutional Neural Network (CNN) image encoder followed by a Recurrent Neural Network (RNN) language decoder. The CNN encoder is usually pretrained with a large amount of image data such as ImageNet[16], and used to extract a feature vector from a given image. The extracted vector is then passed to the RNN decoder to

train a language model that generates sentences. For the encoder, VGGNets[9] or ResNets[10] are widely used, and for the decoder, we typically use LSTMs or GRUs[17].

Figure1. Example images and captions of [18]

There have been various improvements on the basic CNN-RNN architecture. One of the biggest progress is attention-based captioning model first introduced by Xu et al[18]. Attention technique is originally used in the field of machine translation to focus on specific part of an input sequence to generate an output sequence. Similarly, when applied to image captioning, attention is used to concentrate on parts of input images to generate captions. Figure1 shows how the model focuses on specific parts of an image when generating a sentence.

In this paper, we use the basic encoder-decoder model[15] and the attention-based model[18] as baseline models to show that SRE has effect on different kinds of model architectures.

2.2 Image Data Augmentation techniques

Traditional image data augmentation techniques include random horizontal flip, random crop, color space transformation[19, 20], kernel filter[21], and so on. Among these augmentation techniques, random crop and random horizontal flip are often used in image captioning.

Figure2. Examples of Random Erasing

Random erasing[14] is another kind of basic image data augmentation that uses simple transformations. Random erasing randomly selects an n x m patch of an image and masks it with other (0s, 255s, mean pixel values, etc.) values. Figure2 shows examples of Random Erasing. It effectively prevents overfitting by forcing the

model to focus on the entire image rather than parts of it. It is similar to dropout in that it performs a kind of regularization by random masking, but it is conducted on image data level rather than on network level. Random erasing shows one of the highest accuracies on CIFAR-10 dataset – it reduced the error rate from 5.17% to 4.31%.

However, a noteworthy disadvantage of Random Erasing is that it is not always a label-preserving transformation. In many fine-grained image classification tasks, labels can change with partial occlusion. For example, in handwritten digit recognition, “8” can be transformed into “3” when the left part of the image is cut out. Thus, in order to properly apply random erasing, some modification is necessary depending on the tasks and datasets. In this paper, we modify random erasing so that it does not occlude salient regions (object regions) of an image. This way, random erasing can effectively regularize the input without harming the perceptive nature within it. We also experimentally show that while the original Random Erasing shows little effect in image captioning, the proposed SRE (Smart Random Erasing) has noticeable effect on various datasets and baseline models.

Application and magnitude (hyperparameters) of image data augmentation techniques used to be manually designed by the programmer depending on tasks and datasets. Recently, Cubuk et al. introduced AutoAugment[12], which automatically searches for the optimal augmentation policy using reinforcement learning. Subsequently, other augmentation techniques that propose automatic augmentation policy search such as Population Based Augmentation (PBA) [22], Fast AutoAugment[23], and RandAugment[11] have also been proposed. The method we propose, namely Smart Random Erasing, adopts the method of reinforcement learning analogous to AutoAugment in order to find the optimal hyperparameters. Concrete method is described in Chapter 3.

3. Smart Random Erasing

In this section, we introduce Smart Random Erasing, a novel data augmentation technique for image captioning. We made several improvements upon the original Random Erasing, including object recognition, partial occlusion, and automatic hyperparameter search.

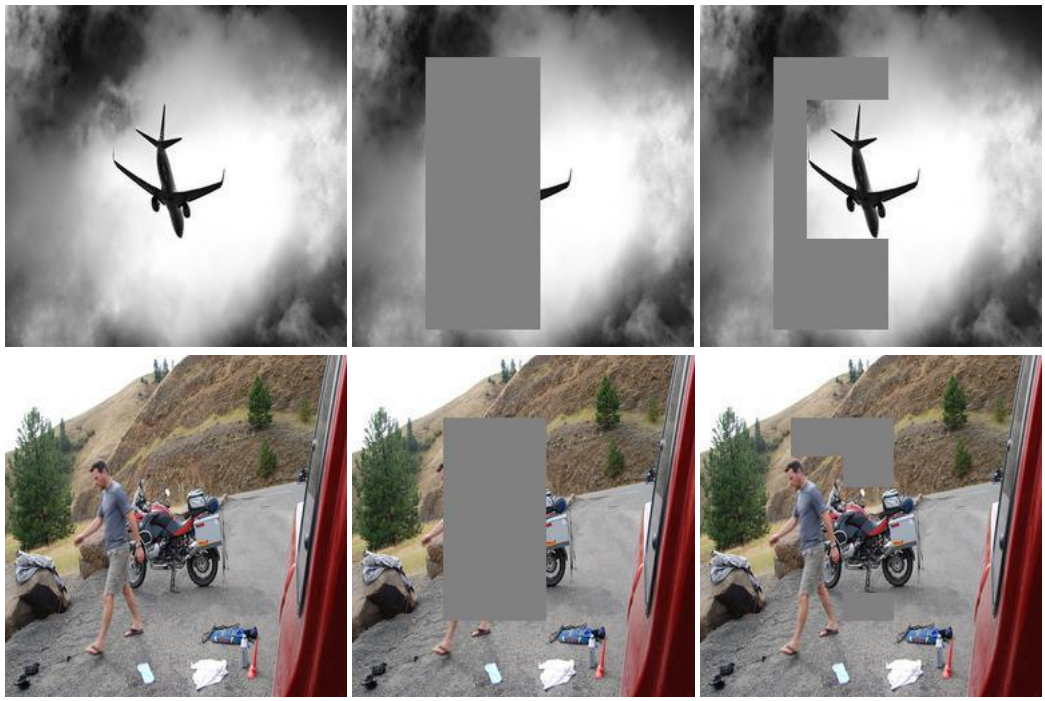


Figure3. Examples of object recovery in SRE

3.1. Object Recognition

Object can be seen as the perceptual core of an image. If an image contains a “boy” and a “ball” we describe the image as the boy doing something with the ball. One of the key drawbacks of Random Erasing when applied to image captioning is that it can also cut out the key point that decides the meaning of the entire image. Thus, in Smart Random Erasing, we randomly erase an n x m patch of the given image like the original random erasing, but recover the part that includes objects detected by the object detector. This way, we can effectively augment images without hurting its label (the

corresponding caption) and teach the model the invariant of image captioning task. Figure3 shows how object recovery is done in Smart Random Erasing and how it is different from original Random Erasing technique. We can see that while Random Erasing can hide salient objects inside a given image, our object recognition and recovery helps safely cutout random regions without the risk of erasing critical regions.

While we can try removing a rectangular patch that does not include any object, this takes considerable time searching for empty space with the right size. Experiments show that recovering objects inside the patch region shows similar or even better performance compared to cutting out a region without any object.

3.2. Object Occlusion

Again, objects are critical when trying to generate a caption of a given image. However, adequate amount of occlusion or variation can make our model more robust. Traditional image data augmentation techniques such as crop, shear, and rotation are also inspired from this intuition. As images are not always ideally given, image recognition needs to be robust to various types of images such as partially occluded images, tilted images, shadowed images, and etc.

Smart Random Erasing applies partial occlusion to detected objects to make the model more robust to various images.

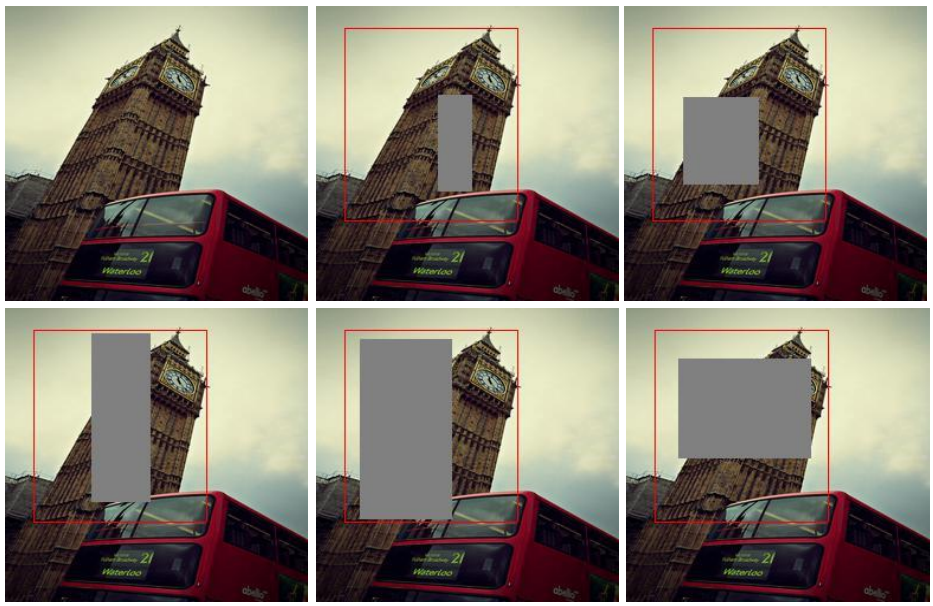


Figure4. Object occlusion depending on the occlusion rate

Specifically, partial occlusion is applied as follows. First, object detector detects several object regions. Then, objects with area bigger than threshold are selected and occluded by generating additional rectangular patch within the object area. The size of the rectangular patch to be generated is object size multiplied by occlusion rate *or*. Figure4 shows how occlusion is applied to an image with different occlusion rate. Experiments show that SRE performs best when occlusion rate is 0.2 (top-right in Figure4). Figure5 shows how an image is augmented by SRE, combining the object recovery and object occlusion in Figure3 and Figure4. We can

see that using SRE, the shape of the masked region is much more diverse, and objects are masked adequately to increase variation but still keep the perceptual core – we are able to recognize the objects.

Figure5. examples of SRE

3.3. Automatic Hyperparameter Search

With the improvements explained in Section 3.1 and Section 3.2, Smart Random Erasing operates with several hyperparameters including box size and occlusion rate. Specifically, there are 4 hyperparameters, *p, s, r, or. p* is the probability of applying SRE, *s* and *r* are hyperparameters that decide the size and shape of the rectangular patch, and *or* is the occlusion rate. While it is possible to manually search for the best hyperparameters given a dataset and a model, it is a very time-consuming and tedious job. Thus, in this paper, we develop a reinforcement learning model that automatically searches for the optimal hyperparameters for Smart Random Erasing.

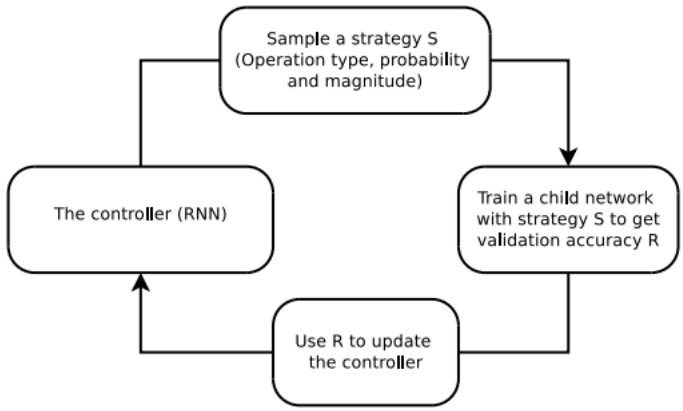


Figure6. Architecture of AutoAugment

Basic architecture of the model is greatly inspired by AutoAugment[12], which also uses reinforcement learning to find the best augmentation policy. The architecture of AutoAugment is described in Figure6. First, we discretize the range of hyperparameters as in AutoAugment (p 0.1 ~ 0.9, s 0.1 ~ 0.9, r 0.1 ~ 0.9, or 0.1 ~ 0.9). Then we train the controller shown in Figure6. The controller is a one-layer LSTM with 100 hidden units. The output of the LSTM at each step is fed into a fully connected layer and then passed to a softmax function to obtain a policy decision. This process is repeated several times to obtain decisions for different hyperparameters. The decision is then fed into the child network to obtain validation accuracy and update the controller.

4. Experiments and Results

In this section, we evaluate the performance of SRE, using two different baseline models and two benchmark datasets – MS COCO[24] and Flickr30k[25].

4.1. Experimental settings

We use two baseline models to show that SRE works on models with different architectures.

First is the implementation of Vinyals et al.'s paper[15], which is a basic encoder-decoder model. Second is the implementation of Xu et al.'s paper[18], which is an encoder-decoder model with attention mechanism. For both models, we use resnet-152 pretrained on ImageNet for the encoder and a 1-layer LSTM with 512 hidden units for the decoder. As for the optimizer, we use Adam optimizer with 0.0001 learning rate.

Note that the reason why we use these two rather old models as baselines is that majority of recent image captioning models adopt these two models as the basic architecture. As we are examining the performance of data augmentation technique, we decided that it is

more crucial to experiment on basic and widely used architectures rather than complex ones.

The controller for the reinforcement learning adopts PPO (Proximal Policy Optimization) with learning rate 0.00001 as in AutoAugment.

4.2. Evaluation metric

We use 5 evaluation metrics – BLEU[26], METEOR[27], ROUGE-L[28], CIDEr[29], SPICE[30] – that are commonly used in image captioning tasks.

BLEU calculates the overlap of n-grams (1~4) between the reference caption and the generated caption to compute the similarity. It is the most basic and widely used evaluation metric in image captioning. However, BLEU can be inexact because it does not extensively consider syntax or synonyms, and tends to score better when a sentence is short.

METEOR is the harmonic mean of weighted precision and recall. It also applies stemming and synonymy matching, resulting in more accurate evaluation. ROUGE is a recall score from reference

captions and generated captions. As for ROUGE-L, recall for the Longest Common Subsequence (LCS) is computed. It is noteworthy that ROUGE only takes recall into account, while METEOR takes both precision and recall into consideration.

CIDEr exploits the idea of TF-IDF to evaluate performance of image captioning. As in TF-IDF, it naturally outputs high scores for sentences with many words in common but penalizes words that are frequently used (e.g. a, the, man).

SPICE does not compare raw sentences but rather compares the scene graphs of two sentences to compute the similarity. Scene graph captures the essence of a sentence and better expresses the relationships between each concept. Figure7 shows an example scene graph of a reference caption. By comparing scene graphs, SPICE does not merely rely on overlap of words but tries to capture the structural similarity between the reference and generated captions.

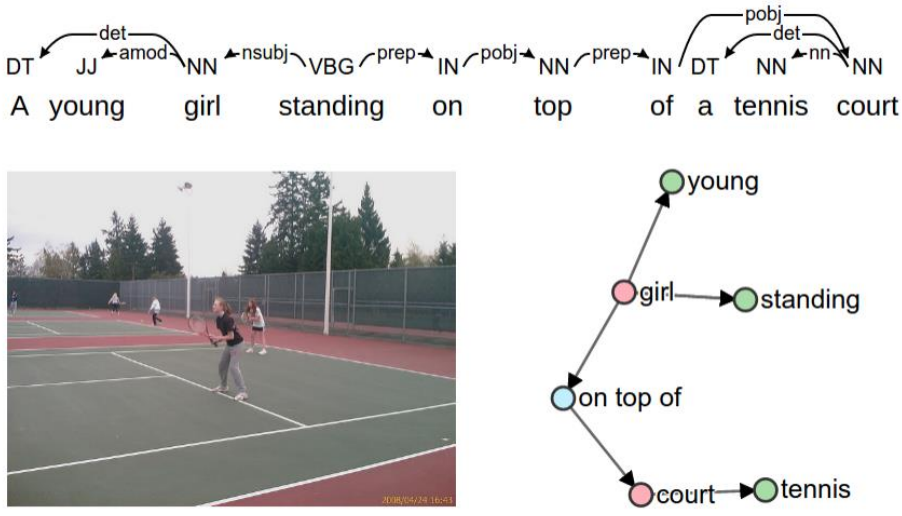


Figure7. Scene graph of an example caption

4.3. Experiment results and analysis

In this section, we quantitatively and qualitatively analyze the performance of SRE. First, we compare the performance of SRE with traditional image data augmentation techniques that are frequently used in image captioning (i.e. Random Crop and Random Horizontal Flip) and RandAugment. Second, we also perform ablation study by comparing the original Random Erasing with SRE on image captioning datasets.

4.3.1. Comparison with other DA techniques

We evaluate the performance of SRE in comparison with other data augmentation techniques, namely Random Crop, Random Horizontal Flip, and RandAugment. We show experiment results for all combinations of the three baseline augmentation techniques as shown in Chart1~4. F, C, R corresponds to Flip, Crop and RandAugment.

	base	F	C	R	F+C	C+R	F+R	F+C+R	SRE
BLEU-4	0.189	0.195	0.196	0.187	0.193	0.174	0.187	0.183	**0.198**
METEOR	0.190	0.192	0.192	0.190	0.192	0.181	0.188	0.186	**0.195**
ROUGE	0.446	0.449	0.451	0.446	0.450	0.433	0.446	0.441	**0.455**
CIDEr	0.532	0.545	0.555	0.531	0.548	0.483	0.530	0.520	**0.566**
SPICE	0.115	0.117	0.118	0.116	0.118	0.108	0.115	0.114	**0.122**

Chart1. Experiment result on COCO dataset using [15]

	base	F	C	R	F+C	C+R	F+R	F+C+R	SRE
BLEU-4	0.115	0.124	0.126	0.094	0.117	0.125	0.102	0.117	**0.129**
METEOR	0.141	0.141	0.139	0.130	0.143	0.131	0.128	0.130	**0.143**
ROUGE	0.338	0.350	0.350	0.334	0.344	0.364	0.354	0.357	**0.367**
CIDEr	0.129	0.146	0.139	0.074	0.129	0.098	0.085	0.094	**0.152**
SPICE	0.073	0.077	0.079	0.067	0.075	0.066	0.063	0.064	**0.088**

Chart2. Experiment result on Flickr30k dataset using [15]

	base	F	C	R	F+C	C+R	F+R	F+C+R	SRE
BLEU-4	0.278	0.274	0.277	0.268	0.278	0.265	0.277	0.274	**0.28**
METEOR	0.240	0.241	0.240	0.237	0.241	0.235	0.238	0.237	**0.252**
ROUGE	0.513	0.515	0.514	0.509	0.517	0.505	0.515	0.512	**0.519**
CIDEr	0.901	0.902	0.891	0.875	0.899	0.861	0.889	0.874	**0.911**
SPICE	0.173	0.175	0.171	0.170	0.173	0.167	0.169	0.168	**0.181**

Chart3. Experiment result on COCO dataset using [18]

	base	F	C	R	F+C	C+R	F+R	F+C+R	SRE
BLEU-4	0.207	0.199	0.202	0.195	0.202	0.202	0.206	0.193	**0.211**
METEOR	0.188	0.191	0.187	0.189	0.187	0.188	0.191	0.186	**0.201**
ROUGE	0.443	0.438	0.439	0.436	0.439	0.437	0.443	0.431	**0.456**
CIDEr	0.437	0.421	0.409	0.399	0.419	0.408	0.428	0.387	**0.44**
SPICE	0.127	0.134	0.129	0.132	0.127	0.126	0.133	0.125	**0.142**

Chart4. Experiment result on Flickr30k dataset using [18]

Chart1 and 2 show the result for SRE and other augmentation techniques on [15]. As for the hyperparameters, our reinforcement learning technique resulted in *p=0.5, s=0.2, r=0.3, or=0.2* for COCO dataset, and *p=0.4, s=0.1, r=0.2, or=0.2* for Flickr30k dataset. Notice that while Random Horizontal Flip and Random Crop shows a bit of performance boost for both COCO and Flickr30k, RandAugment shows no improvement in performance. This empirically shows that RandAugment does not have noticeable effect on image captioning.

The result conforms to our overall hypothesis that different data augmentation is needed for different task and data domain. The suggested SRE achieves the highest score for all 5 evaluation metrics for both COCO and Flickr30k. Considering the performance boost ratio of the traditional augmentation techniques (Random Crop and Random Horizontal Flip), we can conclude that the performance improvement of SRE shown in the chart is meaningful.

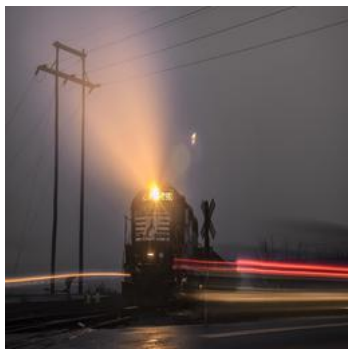
Chart3 and 4 show the result for the augmentation techniques on [18]. Hyperparameters are *p=0.4, s=0.2, r=0.2, or=0.2* for COCO and *p=0.4, s=0.1, r=0.2, or=0.1* for Flickr30k dataset. Similarly, SRE also shows the highest score for all 5 evaluation metrics.

It is noteworthy that not only RandAugment, but also Random Crop and Random Horizontal Flip do not show any significant effect when applied to attention-based model. Furthermore, the combination of augmentation technique that works best is different for all four of the experimental settings (Chart 1 ~ 4). This suggests that the traditional data augmentation techniques that are frequently used when developing image captioning model are in fact ineffective and model-dependent. The suggested SRE, however, can adjust to

different models and datasets by searching for the appropriate hyperparameters.

baseline	A small dog is sitting on a boat in the water.
ours	A man is riding a surfboard.

baseline	A man sitting on a chair next to a laptop.
ours	A man sitting on a bench with a hat on.



baseline	A large white clock tower with a large clock on top.
ours	A red train is on the tracks.

baseline	A man riding a motorcycle down a street next to a car.
ours	A man walking down a street holding an umbrella.

Figure8. Qualitative analysis of SRE

We also qualitatively check the performance of SRE. SRE algorithm greatly focuses on “objects” inside a given image, only masking regions that would not hurt the identity of the objects. By looking at 300 example images and generated captions, we realized that model’s ability to identify an object correctly noticeably

improved. Figure8 shows the example images and captions where the baseline did not identify the objects inside the images correctly, but SRE did.

4.3.2. Comparison with original Random Erasing

We perform ablation study in order to compare the performance of SRE with the original Random Erasing and analyze the effect of each suggested improvement (i.e. object recognition and occlusion).

Chart 5 shows comparison of the baseline model, Random Erasing, and SRE on COCO dataset. We used [15] for the baseline architecture. Here we can see that the original Random Erasing harms the performance of baseline model by erasing critical parts of given images. However, SRE complements this drawback and enhances the performance by object recognition and occlusion.

	base	RE	SRE(OR)	SRE(OR + OO)
BLEU-4	0.189	0.183	0.195	0.198
METEOR	0.190	0.187	0.194	0.195
ROUGE-L	0.446	0.441	0.452	0.455
CIDEr	0.532	0.509	0.557	0.566
SPICE	0.115	0.113	0.117	0.122

Chart5. Ablation study of SRE

Chapter 5. Conclusion and Future Work

In this paper, we suggest a new data augmentation technique for image captioning named SRE. SRE complements the drawbacks of original Random Erasing and effectively augments given image data without harming its perceptual nature with object recognition and object occlusion. We also adopt the reinforcement learning paradigm to automatically search for the optimal hyperparameters for different datasets. By selecting different hyperparameters for different datasets, SRE can adapt to different datasets and models without additional computational effort. As a result, SRE achieves greater performance improvement on various datasets and models when compared to traditional widely-used data augmentation techniques such as Random Crop and Random Horizontal Flip, as well as RandAugment.

We also experimentally show that image data augmentation techniques are highly dependent on tasks, datasets, and even model architectures. However, recent researches for image data augmentation technique often only focus on image classification tasks. Active research is necessary for various augmentation techniques devised for other various problems.

Lastly, while SRE shows good performance on the suggested baselines, it is not completely model agnostic. Some recent researches on image captioning models do not use CNNs to extract features of the complete image. Rather, they use object detector to find regions of interest (ROI) and give the region features as input to the model. For such models, SRE naturally does not work, and other augmentation technique is necessary.

References

[1] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon, “Image Captioning with Very Scarce Supervised Data: Adversarial Semi-Supervised Learning Approach.” ACL, 2019.

[2] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqiang Lu. “MSCap: Multi-Style Image Captioning with Unpaired Stylized Text.” CVPR, 2019.

[3] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. “Unpaired Image Captioning by Language Pivoting.” ECCV, 2018.

[4] Feng, Yang, et al. “Unsupervised image captioning.” CVPR, 2019.

[5] Laina, Iro, Christian Rupprecht, and Nassir Navab. “Towards Unsupervised Image Captioning with Shared Multimodal Embeddings.” ICCV, 2019.

[6] Jiuxiang Gu, Shafiq Joty, et al. “Unpaired Image Captioning via Scene Graph Alignments.” ICCV, 2019.

[7] A. Krizhevsky, “Learning Multiple Layers of Features from Tiny Images,” Technical report, University of Toronto, pp. 1–60, 2009.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions.” CVPR, 2015.

[9] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition.” ICLR, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition.” CVPR, 2016.

[11] Ekin. D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. “RandAugment: Practical automated data augmentation with a reduced search space.” CVPR, 2019.

[12] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. “AutoAugment: Learning Augmentation Strategies from Data.” CVPR, 2019.

[13] Philip. T. Jackson et al. “Style Augmentation: Data Augmentation via Style Randomization.” CVPR, 2019.
[14] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. “Random Erasing Data Augmentation.” AAAI, 2017.
[15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. “Show and Tell: A Neural Image Caption Generator.” CVPR, 2015.
[16] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, 2012.
[17] Kyunghyun Cho et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.” EMNLP, 2014.
[18] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning, 2015.
[19] Agnieszka M, Michal G. “Data augmentation for improving deep learning in image classification problem.” IEEE international interdisciplinary Ph.D. Workshop, 2018.
[20] Ren W, Shengen Y, Yi S, Qingqing D, Gang S. “Deep image: scaling up image recognition.” CoRR, abs/1501.02876, 2015.
[21] Guoliang K, Xuanyi D, Liang Z, Yi Y. “PatchShuffle regularization.” arXiv preprint. 2017.
[22] Daniel Ho, Eric Liang, Ion Stoica, Pieter Abbeel, Xi Chen. “Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules.” ICML, 2019.
[23] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, Sungwoong Kim. “Fast AutoAugment.” NeuralIPS 2019.
[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick. “Microsoft COCO: Common objects in context.” European conference on computer vision, pages 740–755. Springer, 2014.

[25] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.” ICCV, 2015.
[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: a method for automatic evaluation of machine translation.” ACL, 2002.
[27] Michael Denkowski and Alon Lavie. “Meteor universal: Language specific translation evaluation for any target language.” Proceedings of the ninth workshop on statistical machine translation, 2014.
[28] Chin-Yew Lin. “Rouge: A package for automatic evaluation of summaries.” Text Summarization Branches Out, 2004.
[29] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. “Cider: Consensus-based image description evaluation.” CVPR, 2015.
[30] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. “Spice: Semantic propositional image caption evaluation.” ECCV, 2016.

초록

Image Captioning 을 위한

Smart Random Erasing 기법

김연우

서울대학교 대학원

컴퓨터공학부

이미지 캡셔닝이란 입력이 이미지로 주어졌을 때, 이미지에 대한 자연어 묘사를 생성하는 머신러닝의 한 과제이다. 이미지 캡셔닝은 시각장애인을 위한 보조자막 생성, 캡션 생성을 통한 검색엔진 성능 향상 등 방대한 어플리케이션을 가질 뿐만 아니라 자연어 처리와 컴퓨터 비전 분야를 연결하는 과제로서 중요성을 지니고 있다.

하지만, 이미지 캡셔닝 모델을 학습하는데 필요한 이미지-캡션의 쌍으로된 데이터셋은 매우 한정되어 있고, 현존하는 데이터셋들 또한 생성되는 문장들의 다양성이 부족하며 이미지 분야도 매우 제한적이다. 이를

해결하기 위해 최근엔 비지도 학습 모델의 연구도 진행되었으나,
현재로서는 지도 학습 모델의 성능을 따라가기엔 아직 한참 부족하다.

데이터 부족 문제를 완화하기 위한 또 다른 방법으로는 데이터 증강 기법이 있다. 최근 이미지 데이터 증강 기법은 AutoAugment, RandAugment 등 활발하게 연구가 진행되고 있으나, 대부분의 연구들이 이미지 분류 문제를 위한 기법들이고, 이를 그대로 이미지 캡셔닝 문제에 적용하기엔 어려움이 있다.

따라서 본 연구에서는 실험을 통해 기존의 데이터 증강 기법이 문제, 모델, 데이터셋에 따라 성능이 매우 달라진다는 것을 확인한다. 그리고 기존의 데이터 증강 기법을 발전시켜 이미지 캡셔닝 문제에 적합한 새로운 기법을 개발하고, 해당 기법의 성능을 실험적으로 검증한다.