### LAB 3: VAE and DCGAN

Tai-Hsiang Huang (mtk\_06584)

MediaTek Inc.

1. **Introduction**

Automatically generating captions for an image is a task close to the heart of scene understanding — one of the primary goals of computer vision. Not only must caption generation models be able to solve the computer vision challenges of determining what objects are in an image, but they must also be powerful enough to capture and express their relationships in natural language. For this reason, caption generation has long been seen as a difficult problem. It amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language and is thus an important challenge or machine learning and AI research.

In this assignment, three caption generators [1]-[3] are implemented as an exercise to learn how to combine a convolutional neural network (CNN) and a recurrent neural network (RNN). These language generators adopt a similar flow whose block diagram is shown in Fig. 1. Given an input image, the CNN based attention model would computes several 7x7 feature maps that determines the saliency region of an image. After that, a RNN with attention over the image would be applied to generate the output caption word by word. The details of each caption generator would be discussed in Section 2.

The rest of this report would be arranged as follows. The experiment setup would be described in Section 2. The experimental results would be shown and discussed in Section 3, and finally, the findings would be shown in Section 4.

1. **Experiment Setup**

This section describes the environment of the experiment conducted in this work. The dataset is described in section 2.1, the architecture of the neural network is described in section 2.2, and the training hyper-parameters are described in section 2.3.

## **Dataset**

The COCO dataset is used as the training and test dataset in this work. It is a large-scale object detection, segmentation, and captioning dataset. It consists of 330K images with 5 captions per image. Fig. 2 shows an example of an image and the 5 captions corresponding to it in COCO dataset.

## **Attention Models**



**Fig. 1.** Block diagram of the caption generator.



**Fig. 2.** Example of an image and its corresponding captions in COCO dataset.

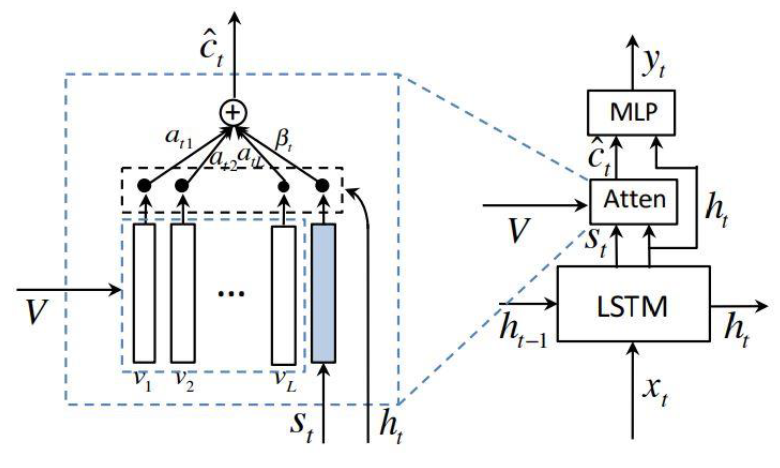
The attention model emphasizes the salient objects of the input image so that the caption generator would have higher probability to report the text corresponding to the object. Fig. 3 shows an example that the trees in the input image is emphasized and hence the word “tree” is reported in the output caption.



**Fig. 3.** The attention model emphasizes the saliency objects in caption generation.



1. (b)



(c)

**Fig. 4.** The architectures of the attention models implemented in this work: (a) SAT [1], (b) BUTPA [2], and (c) KWL [3].

Fig. 4 illustrates the three different network architectures implemented in this work: Show, Attend, and Tell (SAT) architecture [1], Bottom-Up and Top-Down Attention (BUTPA) architecture [2], and Knowing When to Look (KWL) architecture [3]. The SAT architecture concatenates the output of the attention model **ct** with the encoded word **xt** and feeds the coupled feature map into a Long-Short-Term-Memory (LSTM) neural network with hidden layers **h**. The BUTPA architecture introduces the top-down attention to the neural network by adding an extra LSTM neural network to cope with the attention. The output of this LSTM neural network would be feed into both the Bottom-up attention model and the language LSTM. The KWL architecture introduces an adaptive attention encoder-decoder framework which can automatically decide when to rely on visual signals and when to just rely on the language model.

## **Training Hyper-Parameters**

The high level training parameters are listed in Table I, and the training hyper-parameters are listed in Table II.

TABLE I

Training Hyper-Parameters

|  |  |
| --- | --- |
|  | SAT/BUTPA/KWL |
| Batch size | 10 |
| Input encoding size | 512 |
| RNN size | 512 |
| Atten hidden size | 512 |
| Fc feat size | 2048 |
| Att feat size | 2048 |
| RNN type | LSTM |
| Loss function | Cross-entropy |

TABLE II

Training Hyper-Parameters

|  |  |
| --- | --- |
|  | SAT/BUTPA/KWL |
| Method | SGD |
| Momentum | 0.9 |
| Mini-batch size | 128 |
| Total epochs | 164 |
| Learning rate | 0.1, epoch < 81,  0.01, 81<= epoch < 122,  0.001, 122<= epoch |
| Weight decay | 0.0001 |
| Weight initialization | Conv2D.weight: KaiMing\_Normal  Conv2D.bias: 0  BatchNorm2D.weight: 1  BatchNorm2D.bias: 0  Linear.weight.std: 0.001  Linear.bias: 0 |
| Loss function | Cross-entropy |

1. **experimental results**

The minimum training loss of SAT, BUTPA, and KWL are shown in Table III. For each type of networks, the results of 20-layer, 56-layer, and 110-layer are reported.

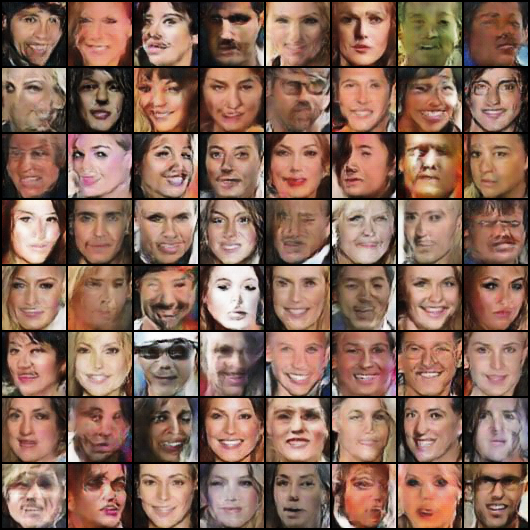
TABLE III

Minimum Training Loss

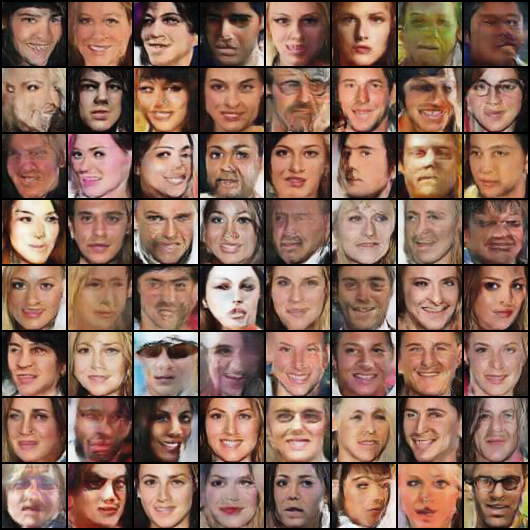
|  |  |  |  |
| --- | --- | --- | --- |
|  | SAT | BUTPA | KWL |
| Minimum training loss | 1.861 | 1.861 | 1.861 |



(a)

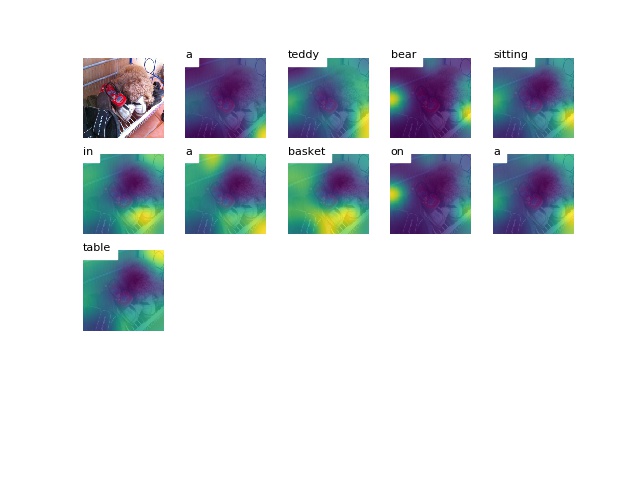


(b)

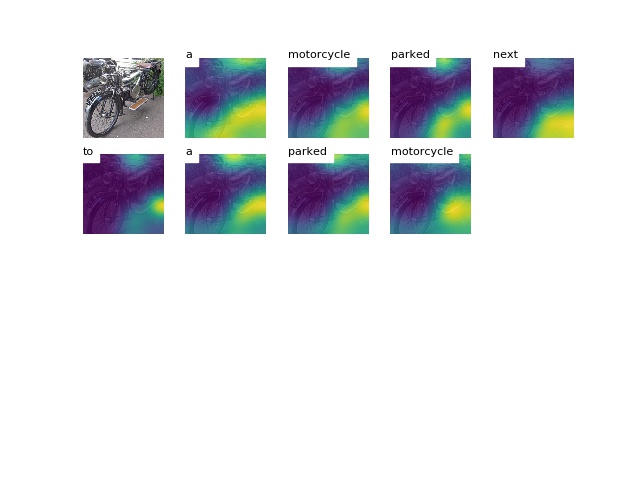


(c)

**Fig. 3.** DCGAN results: (a) EPOCH 0, (b) EPOCH 10, and (c) EPOCH 25



(a)



(b)

**Fig. 4.** Test error curves: (a) 20 layers, (b) 56 layers, and (c) 110 layers.

1. **Discussion**

In this paper, we have described a method based on the

* Importance of initial parameter

1. **References**
2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning (pp. 2048-2057).
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2017). Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998.
4. Lu, J., Xiong, C., Parikh, D., & Socher, R. (2016). Knowing when to look: Adaptive attention via A visual sentinel for image captioning. arXiv preprint arXiv:1612.01887.