Deep Learning for Computer Vision – HW#3

宋體淮, R09921135, Electrical Engineering

Problem 1: Image Classification with ViT

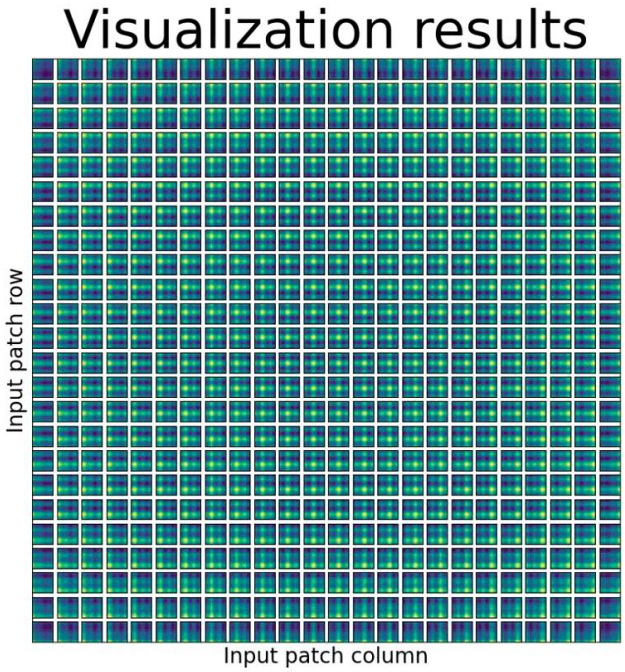
1. **Report accuracy of your model on the validation set.**

Accuracy
0.9500

- Discuss and analyze the results with different settings:

I have tried training the model from scratch or from pre-trained models. The result shows that the accuracy starts with a very low score when training from scratch, which indicates that the transformer-based model is extremely hard to train. On the other hand, if training from pre-trained models, it reaches a high performance within just a few epochs.

2. **Visualize position embeddings.**

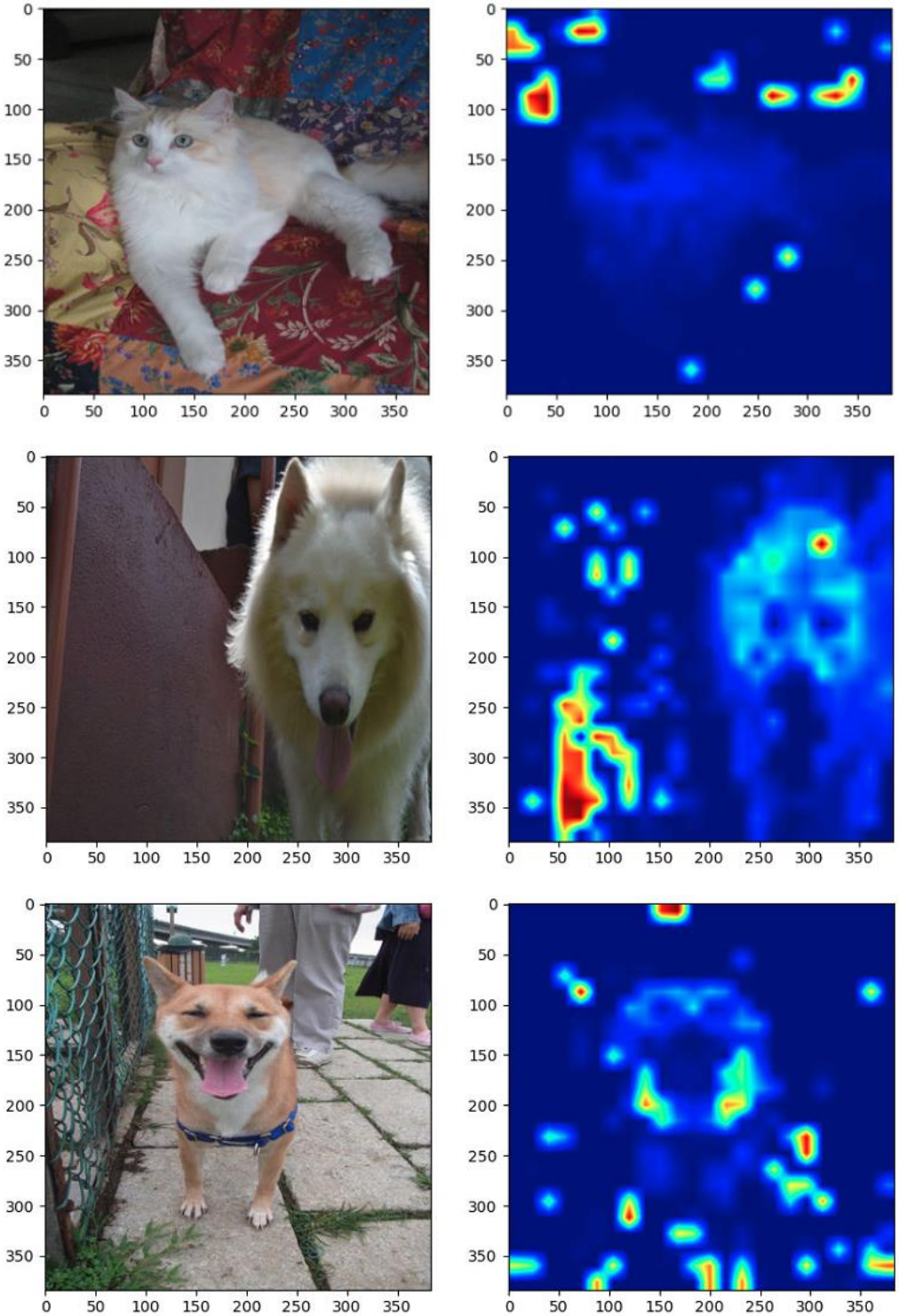


- Discuss or analyze the visualization results:

We can conclude that the model learns to encode distance information in the position embeddings from the following two observations: closer patches tend to have more similar position embeddings; patches in the same row/column have

similar embeddings.

3. Visualize attention map of 3 images.



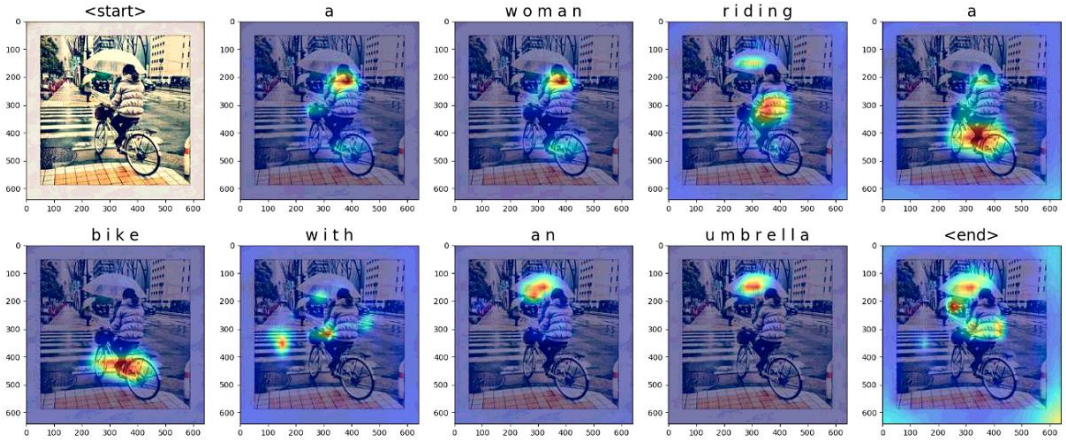
- Discuss or analyze the visualization results:

From the above results, we can see that there are some extreme values on the

outer regions, that is because when the attention maps are passed into the s*oftmax* layer, the larger values will be magnified more. Even so, we can still observe that the object regions are being attended effectively.

Problem 2: Visualization in Image Captioning

1. **Choose one test image and show its visualization result in your report.**



- Analyze the predicted caption and the attention maps for each word:

For the attention maps corresponding to the word *woman, bike, umbrella*, we can clearly see that the attended regions are well reflected to the word. And for the <*end*>, its attention map focus more on the whole image, not just a specific region.

- Discuss what you have learned in this problem:

I have gained better understanding about the cross-attention maps, like which dimension represents query, which dimension represents key or the fact that output of encoder is the right shift of input.