DLCV HW3 Report

學號：B08502141　　姓名：石旻翰　　系級：電機三

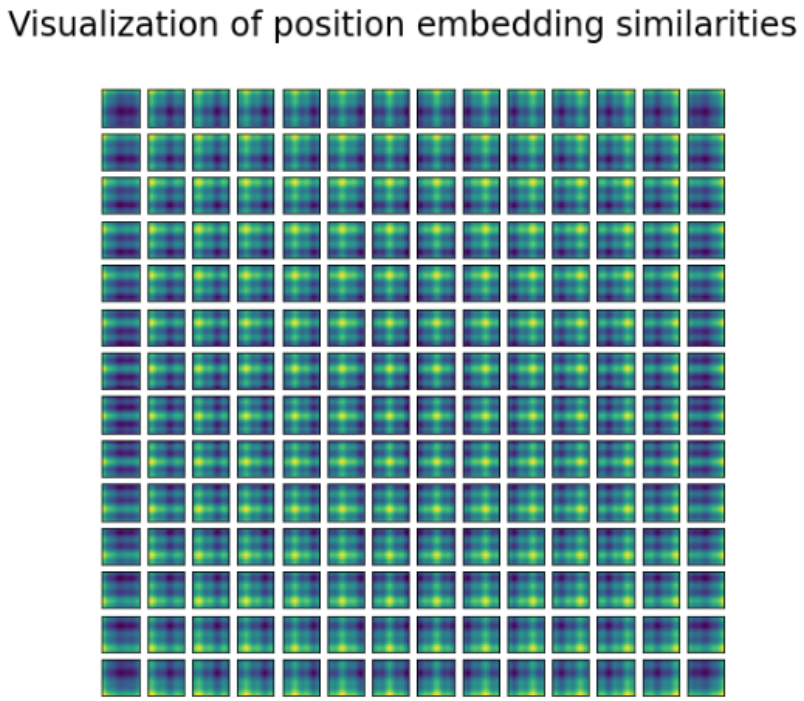
1 Problem1

1.1 Accuracy on valid set

1. Model Architecture:
See the model.txt in folder p1, for it is too long.
2. Hyperparameter:
optimizer=Adam with learning rate = 3e-5, batch size = 4, ViT pretrained = True
3. Accuracy:
Accuracy on valid set: 94.4%

1.2 Visualize position embedding

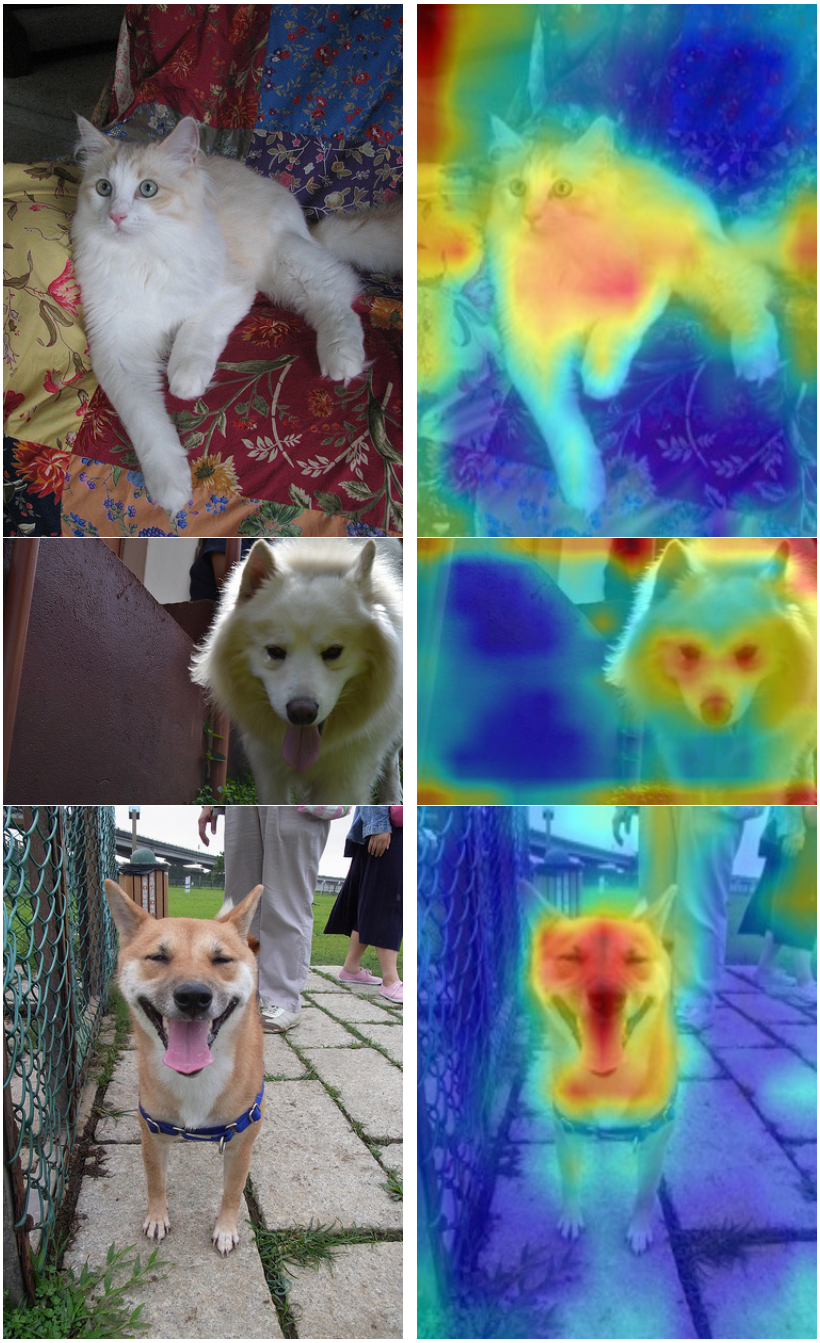
1. Visualization:



2. Analysis:
For each patch in the graph of visualizing position embeddings, there is a lightest row and column. Spot which is at the intersection of the row and the column is also the lightest. Suppose that $P[i][j]$ denotes the patch on the i^{th} row and j^{th} column, the lightest spot is on the (i, j) position in the patch. Through visualizing the cosine similarity between position embeddings, it shows that they indicate the information of the position.

1.3 Visualize attention map

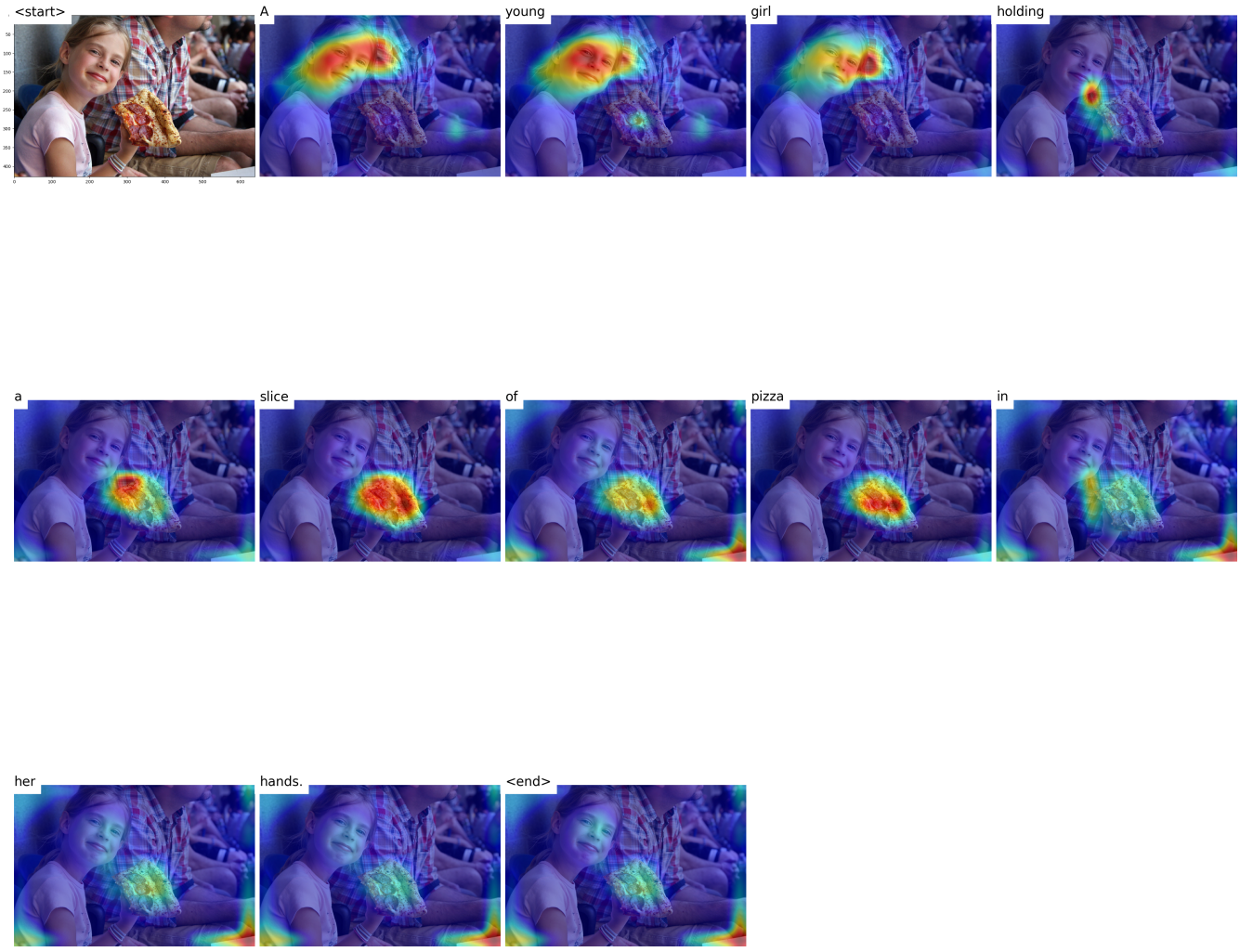
1. Attention map :



2. Analysis:
It is clear that the red, orange, and yellow areas in three images are around the animals, and the blue ones are on the background.For the first image, the cat's main part is covered by the red, orange, and yellow. However, in the second picture, the background is mainly in blue, but some areas, like the four edges in the images, turn out to be red, it is a little strange. In the first image, we can find that there are also some red areas in the edges, but comparing to the second image, it is not such strange. Yet, there are some red spots on the dog's eyes. Finally, for the third figure, the background is covered by blue and no strange area like the first and second images. Perfectly, the dog's face is covered by the yellow and red.

2 Image Caption

2.1 Visualize on test image



1. Analysis:
From above images, we can see that it does focus on the right objects, such as "girl", "pizza".
2. Difficulty:
It takes me such a long time to realize what the codes in Reference[4] are doing, specially thanks to my friend, 曾揚哲, and I modify they to meet the required formats.

References

[1] ViT: https://github.com/lukemelas/PyTorch-Pretrained-ViT

[2] Visualize Position Embedding: https://github.com/hirotomusiker/schwert_colab_data_storage

[3] Visualize Attention Map: https://github.com/jacobgil/vit-explain

[4] Image Captioning: https://github.com/saahiluppal/catr

[5] Colaborators: b07502071 陳志臻、b08902134 曾揚哲