tags: **dlcv**

# HW3

## 1.

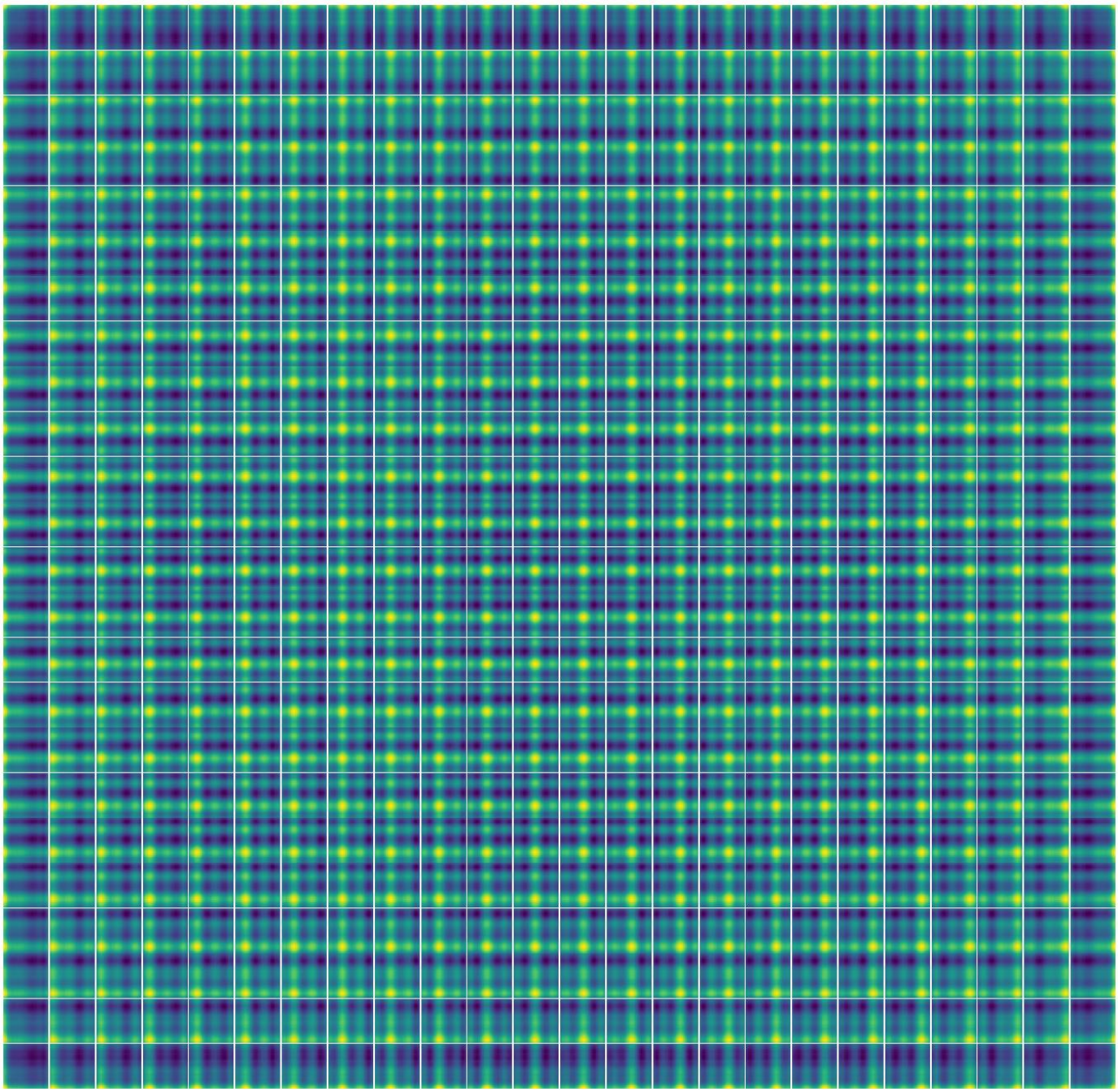
### 1.

a.
use pretrained `google/vit-base-patch16-384` on huggingface, used transforms such as autoaugment and RandomResizedcrop, small learning rate due to fine-tune.

b.
acc: 0.948

### 2.

a.



b.

Patches that are close to each other have similar embeddings, it is reasonable because their positions on the image are close.

3

a.

26_5064.jpg


29_4718.jpg


31_4838.jpg


b.

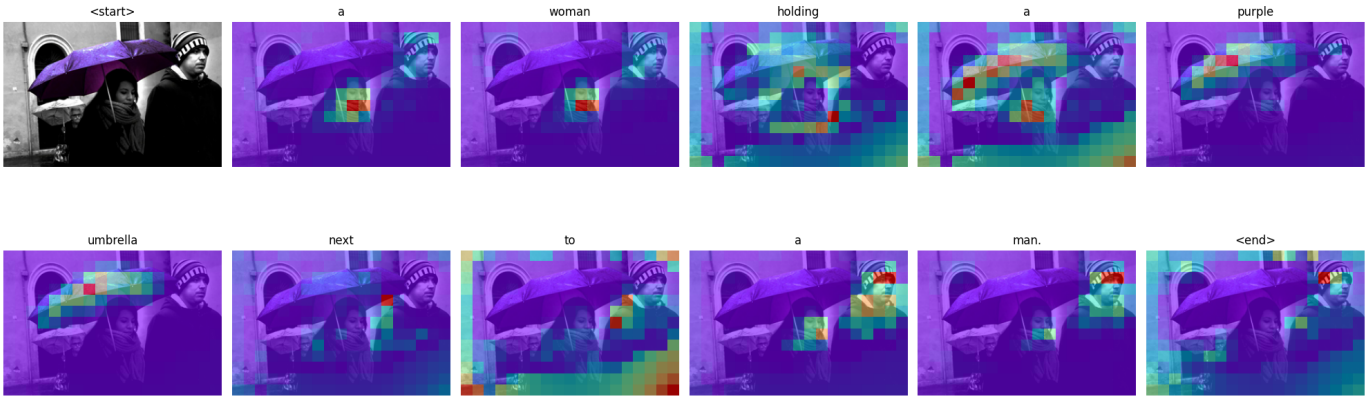
On all images, we can observe that the emphasized area covers most of the area of the class object in the picture. However, on the second image, the heatmap puts some emphasis on the left part of the image, which is quite strange.

## 2.

### 1.

2.

a.




When the word is `a` followed by a noun, the model prefers to have a similar heatmap that both emphasizes the object refered by the noun, which is reasonable because the two words are strongly connected. When the word is a non-object phrase such as `next to`, the heatmap tends not to focus on a single object, but focus on relative position or relation between objects. For the word `holding`, the heatmap focus on the hand, body and the umbrella, which is reasonable because the model need to see the gesture to translate it into action.

b.

From the homework, I learned how to visualize the results of the transformer. Besides, I found that transformer is a complicated model becuase it has many embeddings and

tokens, but it is also very powerful because it can observe the whole sequence during training. I encountered some difficulties when tracing the architecture of the model.