Deep Learning Lab Course

Semantic segmentation, Image captioning and Image-text retrieval

Sejal Mutakekar - 5775691

May 05, 2024

Contents

1 Abstract

Semantic segmentation, Image captioning and Image-text retrieval is implemented using self-supervised learning and Image-text captioning. The code has been executed in the Linux OS, by creating the virtual environment in conda.

2 Self-supervised Learning

2.1 Results and Inferences

2.1.1 Results

Method	Crop Error (Mean)	Crop Error (Median)
Nearest Neighbor Features with Negative Feature Space Distance as Score	51.53	24.42
Nearest Neighbor Features with Negative Cycle Distance as Score	34.75	13.37

Table 1 : **Results : Self-supervised Learning ??**

Graphs showing the Crop Errors and Error values **Figure.** 1

The x-axis labelled 'rank' refers to difficulty level in training data, rising high from left to right. The y-axis labelled 'Crop error' represents the performance measures of the model. For four different cropping scenarios, four different colored line are taken. If the line matches one components from source image to the correct component in the reference cropped image, the crop error is low, if it mismatches, then the crop error is more. From the graph we can say that the model performs better on the easy tasks (lower rank) and poor on the difficult tasks (high rank).

2.1.2 Performance of Cycle Distance against Feature Distance score

From the crop error values, it is evident that the method with cycle distance indicate better performance in comparison with feature distance method. The mean crop error is slightly lower for cycle distance method, which indicates improvement in error reduction.
Hence, the results suggest that cycle distance score improves the accuracy of feature matching using Nearest Neighbour Matching, by reducing the mean and median values.

2.1.3 Advantages of using PCA

Principal Component Analysis (PCA) is applied to an image to reduce the dimensionality of image features. This implies setting pca=10 reduces the dimensionality of features to 10 components. This reduces the noise in the data and potentially improving the efficiency. As the dimensionality is reducing, the computational cost also lowers down and the search for Nearest Neighbours is sturdy .

2.1.4 Advantages of using negative cycle distance

Negative cycle distance has helped in consistent matching and accurate results. There may be random similarity estimation due to the dataset size. It penalizes the inconsistent matches by filtering out inconsistent images, thus reducing the impact of noise on the results.

3 Image-Text

3.1 Image Captioning

Output screenshots can be seen in **Figure.** 3

3.1.1 Complete caption generation with sampling

Results improve with lower temperature.
During caption generation, most probable work is selected by the Greedy method. Using 'topk' count, the word is selected from the set of K most probable words, encouraging exploration which leads to diverse results. High temperature gives less likely words a higher chance of being selected, because it explores so much, that it produces garbage captions giving less bleu score.
With lower temperature, the process is inclined towards most probable words from the topk words and explores more. It gives priority to the words from the topk with higher probability, generating more interpretable and grammatically correct sentences, with high bleu score.

Hence, high temperature leads to too much of exploration leading to meaningless captions and low bleu score. Whereas, low temperature concentrates on most probable words, giving meaningful captions and good bleu score.

3.1.2 Prompt Engineering

	topk	Temperature	Prompt	Bleu score
1.	50	0.7	a picture of	12.66%
2.	50	0.7	an image of	13.38%
3.	50	0.7	a depiction of	6.83%
4.	50	0.7	a figure of	9.24%

Table 2 : **Prompt Engineering**

The score obtained is not steadily increasing nor decreasing, but is inconsistent. For some prompts it increases, for some it decreases. Hence, there is no steady increase observed in the bleu score on changing the prompts. Also, we cannot strongly say that 'prompts' are hyperparameters and hence this may be one of the reasons for not getting promising improvements in the performance.

3.1.3 Student Hyperparameter Search

	topk	Temperature	Prompt	Bleu score
1.	50	0.7	a picture of	12.66%
2.	100	0.5	an image of	17.21%
3.	25	0.3	a depiction of	13.11%
4.	10	1.0	a figure of	6.81%
5.	75	1.0	a figure of	4.74%
6.	100	0.3	an image of	19.91%

Table 3 : **Student Hyperparameter Search**

From the above table it is evident that the blue score can be improved. According to the hyperparameter settings taken into consideration above, on reducing the Temperature and increasing the topk value, the blue score improves. However, increasing the temperature results in poor bleu score. Approximately 20% accuracy is obtained by reducing the temperature to 0.3 and increasing the topk value to 100. On the contrary, increasing the temperature to 1.0 impacts the bleu score poorly.

3.2 Image-Text Retrieval

Output screenshots can be seen in **Figure.** 4

3.2.1 Fine tune instead of training from scratch

	Finetune	Learning rate	Weight decay	Epochs	Temperature	Scores
1.	-	1e-3	1e-3	5	0.1	42.51
2.	True	1e-5	1e-3	3	0.1	59.55
3.	True	1e-5	0	3	0.1	59.69

Table 3 : **Fine tuning**

Fine tuning gives better results than training from scratch. Training from scratch gives the result which is around 43%, while by performing fine tuning, we get the result around 59%. Fine tuning builds upon the already trained model and hence tends to produce better results. Moreover, since the model is not trained from scratch, training time reduces.

3.2.2 Student Hyperparameter search

	Learning rate	Weight decay	Epochs	Temperature	Scores
1.	1e-5	0	5	0.1	1.24%
2.	5e-1	1e-3	5	0.05	16.49%
3.	5e-3	1e-4	8	0.1	43.20%
4.	1e-3	0	8	0.05	41.33%

Table 3 : **Student Hyperparameter Search - train_retrieval**

By changing the hyperparameter settings, the score could also be improved. It depends on the hyperparameters being used. From the table, it is evident that increasing the learning rate and weight decay and setting temperature to 0.1 gave a slight improvement in the result in comparison with the baseline score. For other hyperparameter settings, deterioration in the result could be observed.

4 Appendix

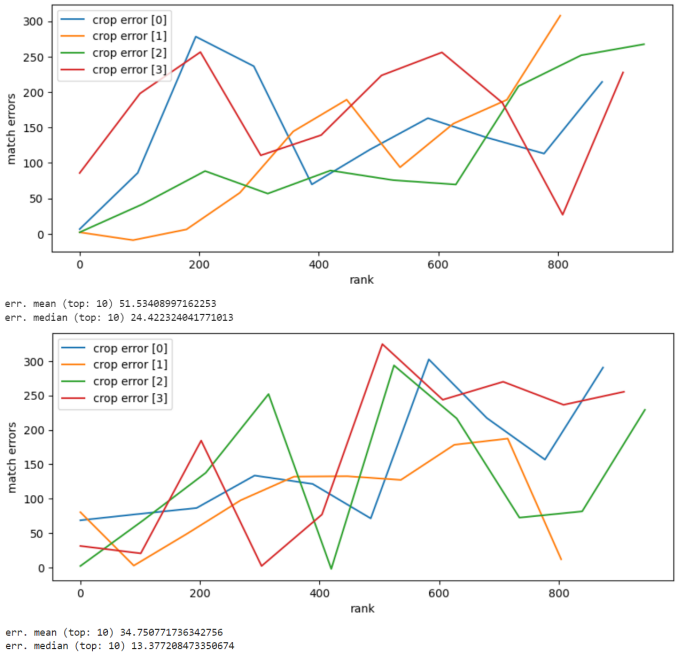
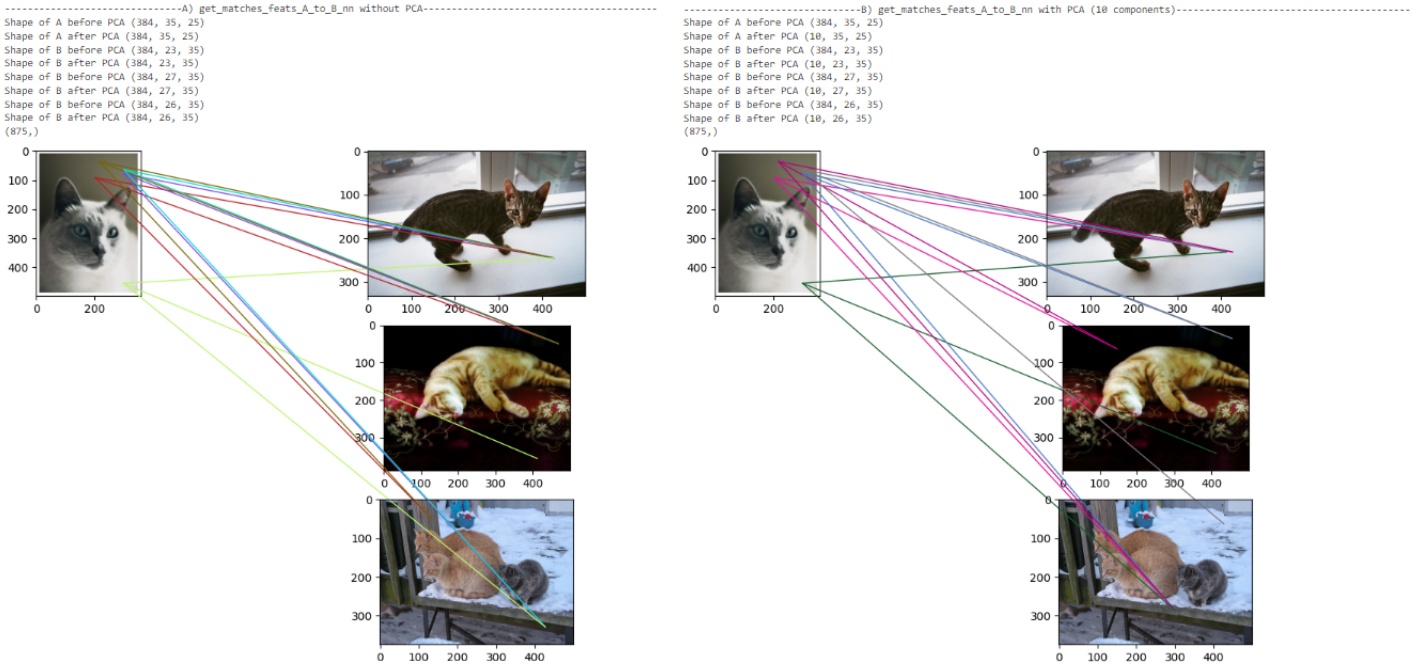


Figure 1: Graphs showing the Crop Errors and Error values
(A. Nearest Neighbor Features with Negative Feature Space Distance as Score, B. Nearest Neighbor Features with Negative Cycle Distance as Score)



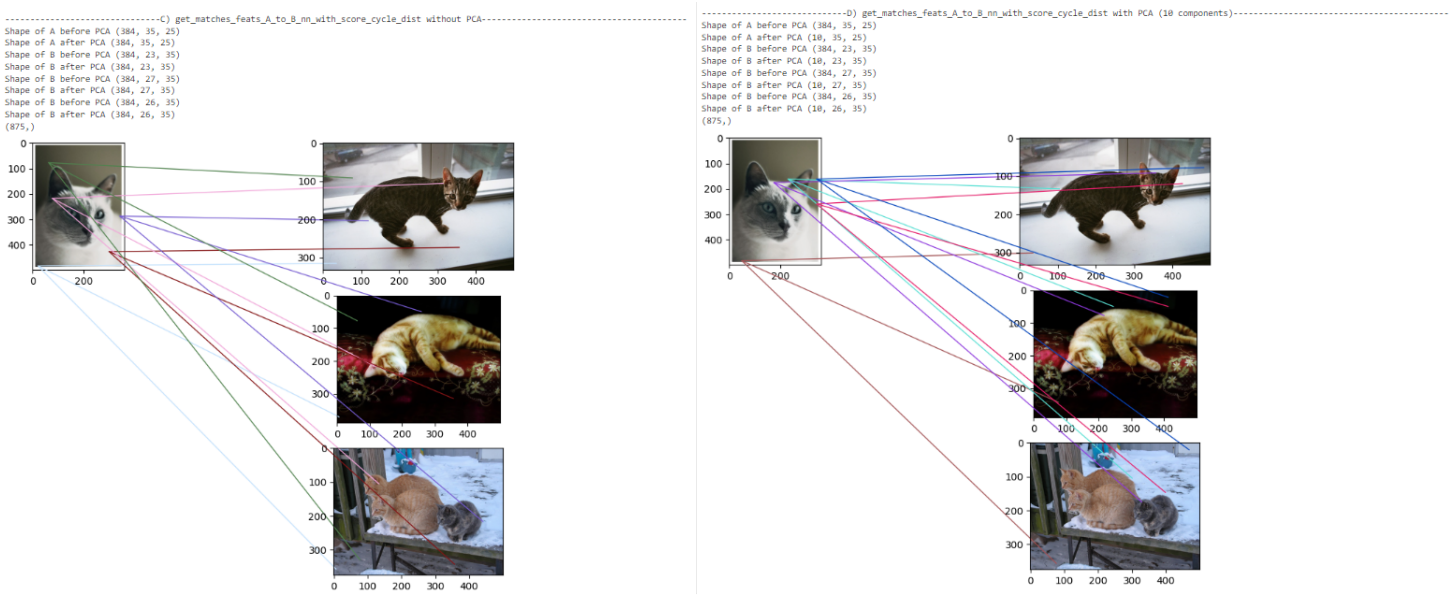


Figure 2: Visualization for Nearest Neighbour using all four Methods

Pred caption: two trains parked on the tracks

Reference caption: two white and red trains parked next to each other

Pred caption: a horse that is standing in the grass

Reference caption: a horse standing next to a green slide

Figure 3: Results for Caption Generation with Greedy Search

Sim. Score: 0.2688276767730713

Caption: a white and red fighter jet on top of a fence

Sim. Score: 0.2748887538909912

Caption: a white and blue airplane parked on a tarmac

Figure 4: Results for Search Query