Cross Modal Representation Learning

Dhruv Metha Ramesh
Rutgers University
Department of Computer Science
dhruv.metha@rutgers.edu

Anindita P Chavan
Rutgers University
Department of Computer Science
anindita.chavan@rutgers.edu

Jash Mitesh Gaglani
Rutgers University
Department of Computer Science
jash.gaglani@rutgers.edu

Sahil Rajendrakumar Raut
Rutgers University
Department of Computer Science
sahil.raut@rutgers.edu

1. Methodology

1.1. Dataset

We use the Recipe 1 Million Dataset, which consists of approximately 1 Million text recipes with titles, instructions, and ingredients in English. We use both the **complete dataset** which contains 238999 train, 51119 validation, and 51303 test image-recipe pairs – a total of 800K (train + test + validation) images and a subset of approximately 0.5 Million recipes containing at least one image per recipe that contains about 400K (train + test + validation).

1.2. Preprocessing

We perform analysis with the **features provided** with the assignment.

1.3. Model

We train a non linear Canonical Correlation Analysis (CCA) model on the extracted features of the train text-image pairs. We train two single layer encoding networks with batch norm, dropout and the relu activation for image and text respectively. We train these models on 2 objectives and compare the results.

Mean Square Loss - Traditional CCA Loss The objective is to reduce the distance between the non-linear projection of text X and image Y, thereby increasing the correlation between the projected text and image, i.e, bringing the image projected points closer to the text projected point. We use the following objective:

$$L = \underset{\theta_1,\theta_2}{\operatorname{argmin}} ||f_{\theta_1}(X) - f_{\theta_2}(Y)||_2$$

Triplet Loss The triplet loss provides a little more signal to the model than just the MSE loss, by also telling the model to push away points that are not in the same category. This additional signal in the loss function makes better separated projected points by keeping points in the same category closer to each other and pushing away points that are not in the same category. To facilitate this in our training, we need to set up the DataLoader to sample points in the form of a triplet $<$ anchor, positive, negative $>$ - (a, p, n). The **anchor** is used as the query text feature. The **positive** is the image feature for this query text feature. We fetch the **negative** sample using the hard negative sampling strategy, where we look for an image feature that is not in the same category as the query text and has a very high cosine similarity with it (this needs to be pushed away). The objective is to minimize the following function, where d is the cosine similarity.

$$L = d(a, n) - d(a, p) + \epsilon$$

Category generation The categories for each data point is found using a bigram technique over the ingredients of each recipe. After finding the most common 2000 bigrams and processing the data, we use the title of each feature and find the most overlapping bigram and label it with that bigram. This gives us a coverage of 50% of the data, rest of which is labelled as a special class ”background”. We need these labels to generate the negative samples for triplet loss.

2. Evaluation

We evaluate on the pairs of models for both the optimization functions.

2.1. Metric

Our model described above is used to find an optimal shared latent dimension for both the image to recipe and recipe to image retrieval tasks. We evaluate these retrieval tasks of the CCA models using two benchmark metrics, namely median rank (medR) and recall rate at top K (RK).

	medR	**R@1**	**R@5**	**R@10**
Title	111.7	2.81	9.29	15.07
Ingredients	38.55	6.54	18.98	27.57
Instructions	32.00	8.24	22.2	31.41
All	**5.55**	**24.14**	**49.39**	**60.3**

Table 1. MSE Loss Results for 1K sample size

	medR	**R@1**	**R@5**	**R@10**
Title	314.25	1.05	3.21	5.42
Ingredients	15.3	12.35	32.15	43.25
Instructions	15.95	12.58	32.28	42.79
All	**2.0**	**46.65**	**73.39**	**81.31**

Table 2. Triplet Loss Results for 1K sample size

Median Rank medR: We compare the generated latent features of the images and text using "correlation" score for each with the other. The rank here is defined as the rank of the score of the ground truth image latent vector when compared with it's text counterpart or vice-versa. Computing these ranks for the entire test set and taking the median is defined as the **medR** metric.

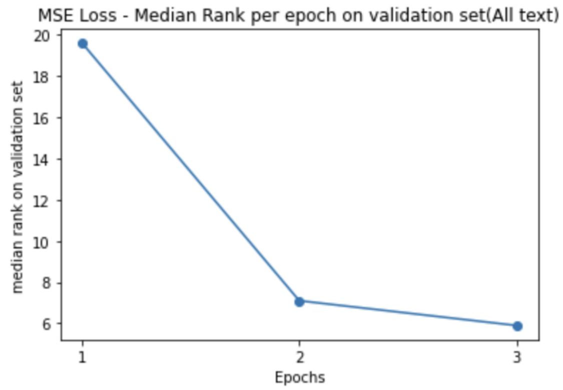

Figure 1. medR (MSE Loss) for full recipe image retrieval

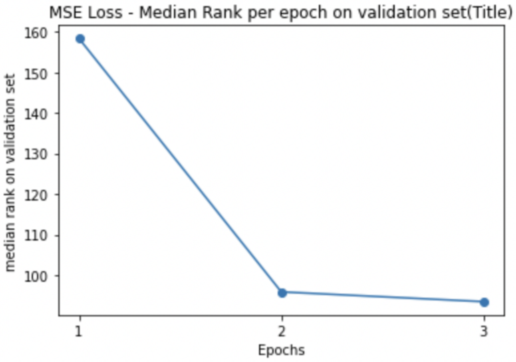

Figure 2. medR (MSE Loss) for title of recipe image retrieval

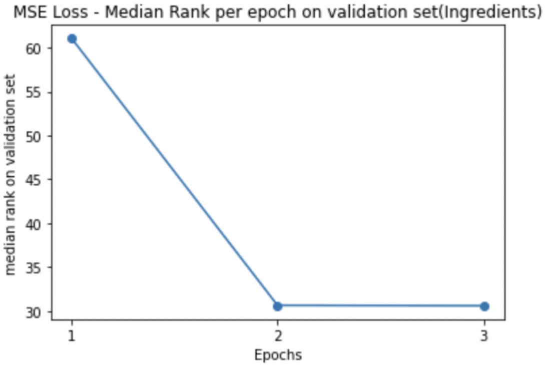

Figure 3. medR (MSE Loss) for ingredients of recipe image retrieval

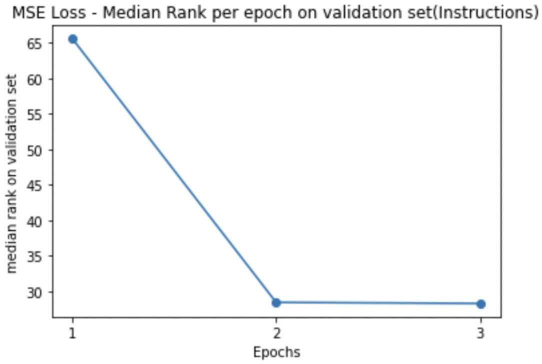

Figure 4. medR (MSE Loss) for instructions of recipe image retrieval

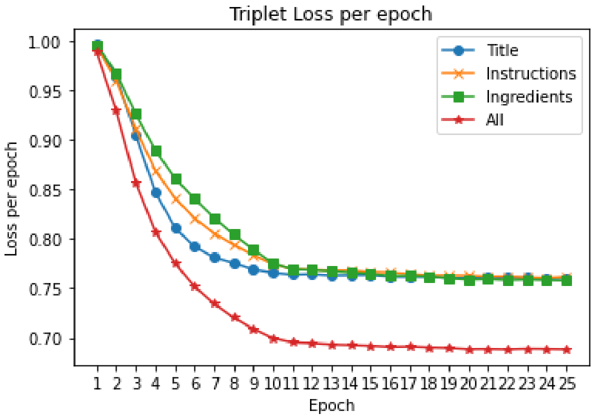

Figure 9. Training loss chart for Triplet Loss

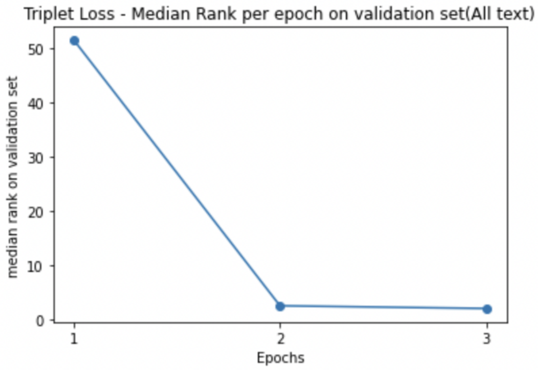

Figure 5. medR (Triplet Loss) for full recipe image retrieval

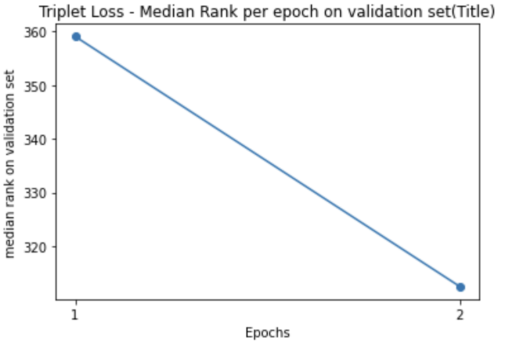

Figure 6. medR (Triplet Loss) for title of recipe image retrieval

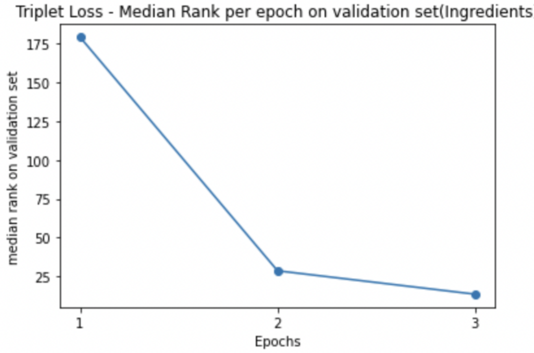

Figure 7. medR (Triplet Loss) for ingredients of recipe image retrieval

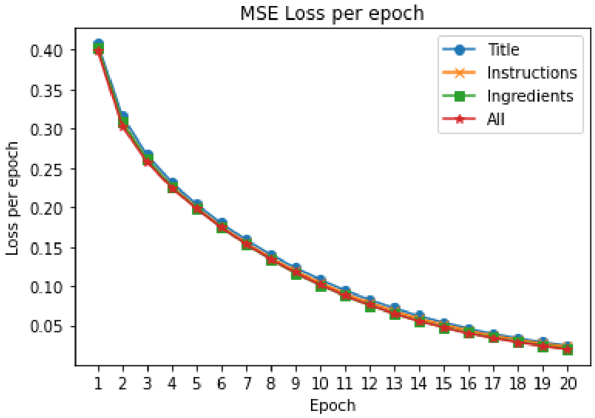

Figure 10. Training loss chart for MSE Loss

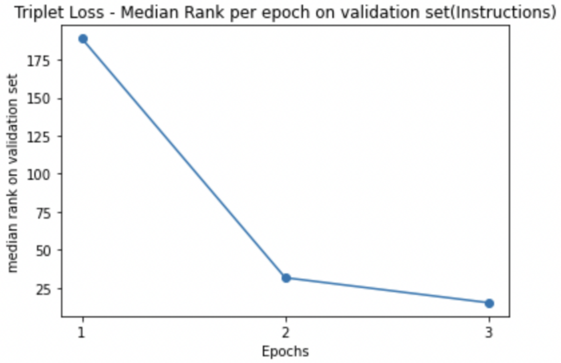

Figure 8. medR (Triplet Loss) for instructions of recipe image retrieval

Recall Rate R@K: This is a standard retrieval metric, that tells us how often is the ground truth image retrieved in the top K ranks (as defined above), where for this project K = 1, 5, and 10.

2.2. Results and Ablation

Using the evaluation metrics given in Sec. 2.1, we can see the results in Table 1, 2 and Fig 1.

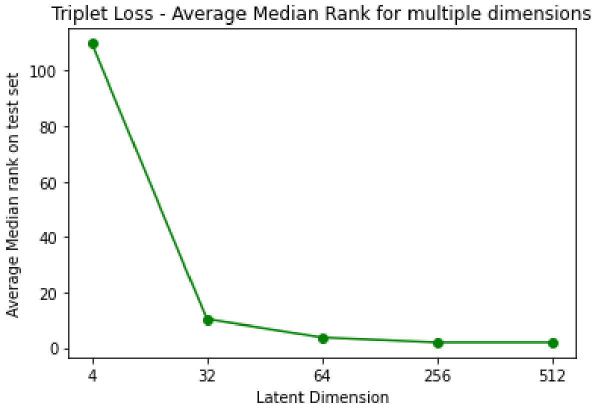

Figure 11. Triplet Loss - Dimension wise median rank

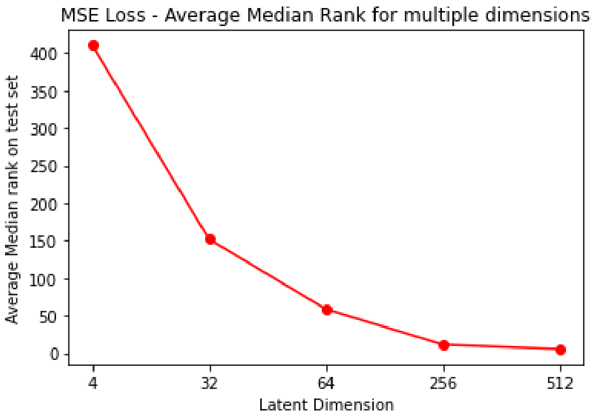

Figure 12. MSE Loss - Dimension wise median rank

2.2.1 Triplet Loss vs MSE

The triplet loss does far better in general on the **medR** and the **R@K** metric in comparison to the MSE loss because the triplet loss takes into account also the negative examples. This helps the data of a similar category be more clustered and different category away from such clusters, whereas in the MSE loss only the latter happens. This can also be seen from the T-SNE plots of the triplet lossand MSE loss.

Ablation for title We can see clearly from the results that the titles behave strangely as it gives a better medR for MSE as compared to Triplet. This is not surprising as the title does not contain enough information about the features in the image, and finding negative samples for such may lead to choosing a positive sample as a negative one, something like a ”prawn curry” and ”shrimp curry”, they may be classified differently, but they are the same. The title feature vector holds very little semantic information of the detailed recipe, and hence pushing them away causes it to perform worse.

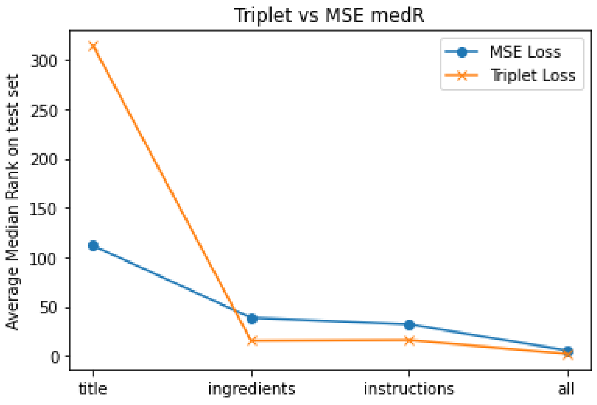

Figure 13. Triplet vs MSE medR

Ablation for the others As we can see, it performs as expected, where the triplet loss features are much better than the MSE feature.

More Ablation from earlier From the results we can see that **instructions and ingredients find a lot more correlation with their images**, owing to the fact that the ingredients and instructions contain many pseudo labels for it's corresponding images. Like for example, using onions as a garnish, onions can be detected by the Resnet features and it's corresponding text ”onion” in the recipe from the Bert features. CCA can learn this correspondence if we have multiple such recipe-image pairs in the dataset. This when compared to it's title text of ”Indian Savoury Dish”, does not contain this fine-grained information. The title talks about the more coarse information in the image.

We can conclude from this that **title** captures more **coarse** image features similar to image labels, but **ingredients and instructions** give us a more **fine** view into the image, for example the way something is cooked can define the color of the dish or the ingredients we had can influence the texture of the dish. Such representations are captured more for instruction and ingredients as compared to the title.

The **concatenated text** feature vector now contains **both the coarse and fine descriptors** (features) of the image, thus following the reasoning given above we see improved performances.

2.2.2 Visualization

We try to further make more inferences from the ablation studies using the t-SNE plots comparing the MSE loss and the triplet loss.

Ablation We see from these plots in Fig 14 Fig 15 that the clusters being formed in triplet are more prominent and have visible boundaries as compared to MSE loss which seems a little more muddled. This again shows how well a triplet loss works with that extra signal of not keeping the negative sample closer to it.

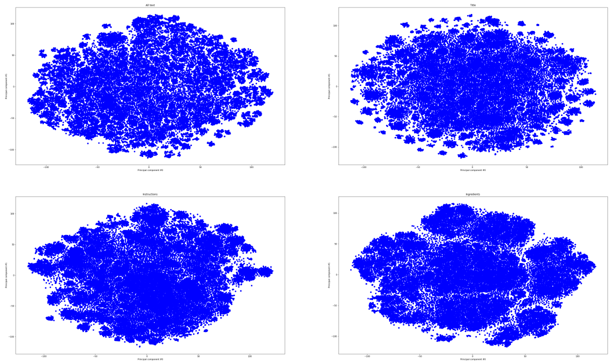

Figure 14. T-SNE (Triplet loss) Text Clusters - All text, Instructions, Ingredients, Title

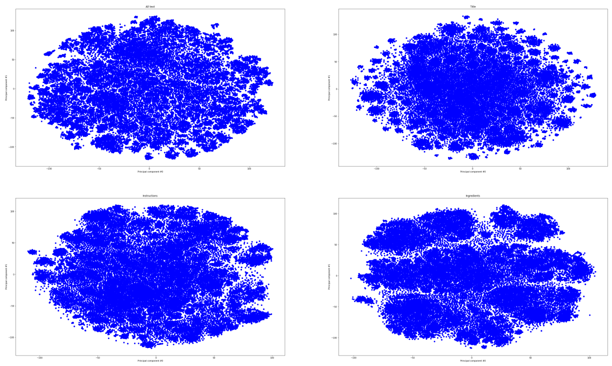

Figure 15. T-SNE (MSE) Text Clusters - All text, Instructions, Ingredients, Title

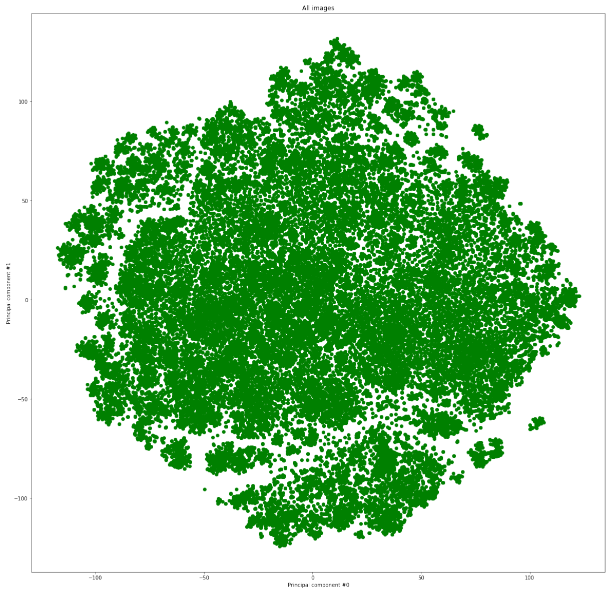

Figure 16. T-SNE (MSE loss) full recipe to image non-linear feature spread

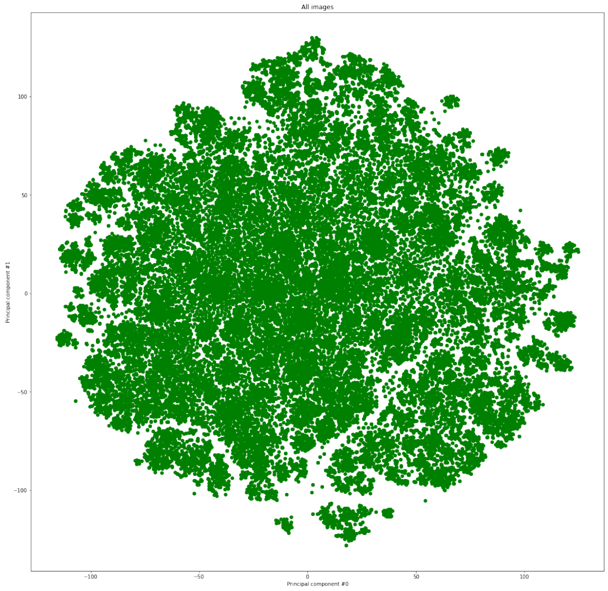

Figure 17. T-SNE (Triplet loss) full recipe to image non-linear feature spread

3. Conclusion

Using a contrastive loss like the triplet loss can substantially improve the latent space separation of the data. Creating margin boundaries between categories that help do crisp data retrieval between different modalities. The boom of contrastive losses is well justified as we have seen from the results in the presentation.