Nonlinear Multi-View Representation Learning

Utsav Patel
utsav.patel10@rutgers.edu

Aniket Sanap
aniket.sanap@rutgers.edu

Sneh Desai
sd1324@scarletmail.rutgers.edu

Sanchit Thakur
st976@scarletmail.rutgers.edu

Abstract

Most real-world problems are characterized by data simultaneously collected from several sensors. For instance, an activity can be recorded by a video camera with image and audio sensors. Web pages contain text, images, audio clips, tables, all of which describe a related concept in that document. Image collections often contain tags or even complete captions written in natural text that describe the content of those images. This project deals with the exploration of different models (both linear and non linear) for the task of cross-model representation learning. Cross-modal representation learning is an essential part of representation learning, which aims to learn latent semantic representations for modalities including texts, audio, images, videos, etc. After dealing with the CCA model in the previous step of the project, the main aim here is to build upon the linear CCA multi-view model by using: (a) nonlinear deep models and (b) the triplet loss. We know that Deep CCA uses deep neural networks (DNNs) for projecting two views into a common subspace and has achieved excellent empirical performance for tasks across several domains in the setting of unsupervised multi-view feature learning. Triplet loss is a loss function for machine learning algorithms where a reference input (called anchor) is compared to a matching input (called positive) and a nonmatching input (called negative). The distance from the anchor to the positive is minimized, and the distance from the anchor to the negative input is maximized. In this way, it can be used to help the network learn better which images are similar and different to the anchor image. Hence in this step of the project, we aim to understand what gains one can obtain from using the two models described above beyond the baseline linear CCA.

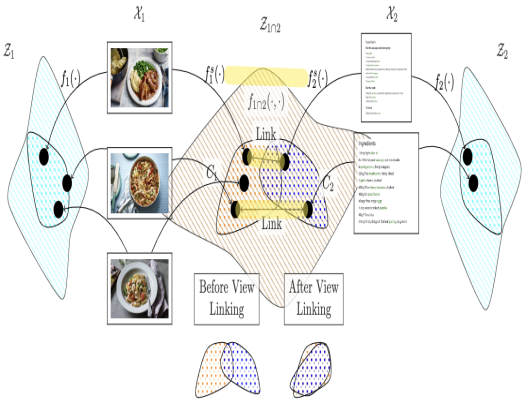


Illustration of Cross-modal Representation Learning

1. Cross Modal Representation Learning

In real world, information about a particular concept or entity is usually obtained through various means, for e.g. an encyclopedia which is known by all age groups contains information in both image and textual format. In accordance with this, most of the real world problems have information coming from multiple media which all describe a related concept. The branch related to analysis of this type of pool of common information fetched from various mediums is called as multi-view analysis. Multi-view, the word is self explanatory which tells that information is seen from multiple views(e.g. text, video and images). More colloquially, cross-model representation learning refers to learning of information coming from multiple modalities or views. Learning such representations form the backbone of various specialized tasks such as: (1) Cross-model Retrieval, (2) Cross-model Translation, and (3) Cross-model Alignment.

(1) Cross-model Retrieval: It refers to predicting infor-

mation in one view while using the input information in another view. It aims at producing flexible retrieval across various modalities. So if we are given some query in one view, the task is to generate representations in other view which comes down to analysing the similarity of information enclosed in various types of data.

(2) Cross model Translation: It refers to generating representation or information in one view while using the input information provided in another view. It uses the fact that information enclosed in various views contain similarities or similar features as they are in-fact describing a common entity. So the task is to find those common subset of features in one view that are in-fact related to another view.

(3) Cross-model Alignment: It refers to retrieve common subset of features in various views. This is in-fact a usage of conceptual similarity between the various views. So as the name suggests, we actually align the various views and come up with a space of representation that has common subset of features or representation in all those views under inspection.

This project focuses on developing models in 3 steps all addressing to the common problem of "Cross-model Representation Learning." The models are made in steps meaning that each successive models is built on top of the predecessor model and this successive model delivers greater performance on the preceding one. In the first step of the project, we explored the classical multi-view representation learning, and used linear models in the process, like the CCA. Now, we build upon the linear CCA model using non-linear deep models and the triplet loss, with the goal of understanding what gains one can obtain from using the models beyond the baseline linear CCA.

2. Model Used

The baseline model being used here is the linear model named Canonical Correlation Analysis(CCA) to performs the task of classical multi-view representation learning. CCA is actually like PCA which deals with extracting Principle Components from a single dataset while keeping the amount of variation at maximum but CCA differs from PCA by the fact that it involves extracting correlation between multiple datasets. CCA uses Canonical Variables at its core to analyse the similarity between the datasets. Canonical Variables are basically linear combinations of various variables in one dataset and as the task is to extract dataset similarity, CCA uses pairs of Canonical Variables wherein each pair contains Canonical Variables from both the datasets under consideration. So as a first step, we use CCA- a linear model for pairwise data to handle the task of multi-view representation learning. Here we have used the dataset provided by the professor.

Non Linear Deep Model

Deep CCA (DCCA), which uses DNNs for projecting two views into a common subspace, has achieved excellent empirical performance for tasks across several domains in the setting of unsupervised multi-view feature learning.

Triplet loss

Triplet loss is a loss function for machine learning algorithms where a reference input (called anchor) is compared to a matching input (called positive) and a non-matching input (called negative). The distance from the anchor to the positive is minimized, and the distance from the anchor to the negative input is maximized.

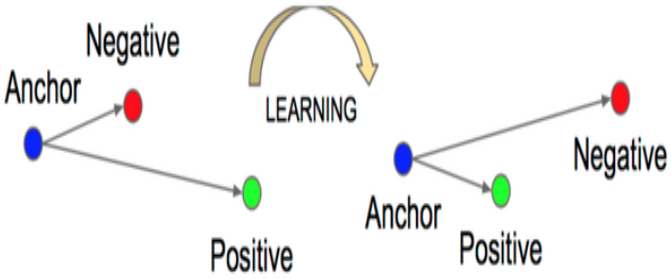


Illustration of Triplet Loss Function

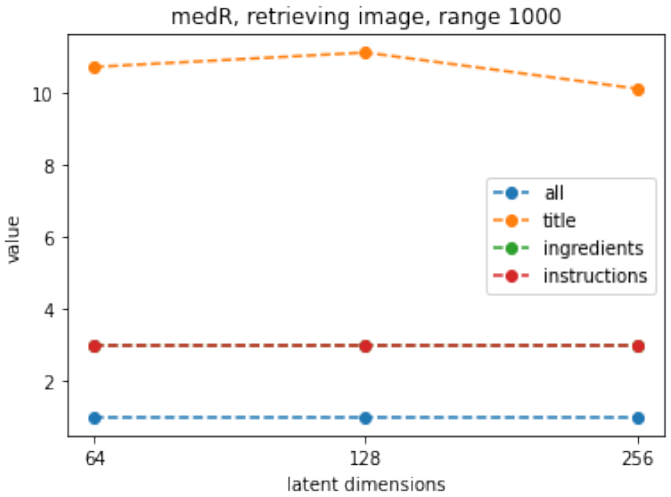
By enforcing the order of distances, triplet loss models embed in the way that a pair of samples with same labels are smaller in distance than those with different labels. Unlike t-SNE which preserves embedding orders[further explanation needed] via probability distributions, triplet loss works directly on embedded distances. Therefore, in its common implementation, it needs soft margin treatment with a slack variable α in its hinge loss-style formulation. It is often used for learning similarity for the purpose of learning embeddings, such as learning to rank, word embeddings, thought vectors, and metric learning.

3. Dataset Used

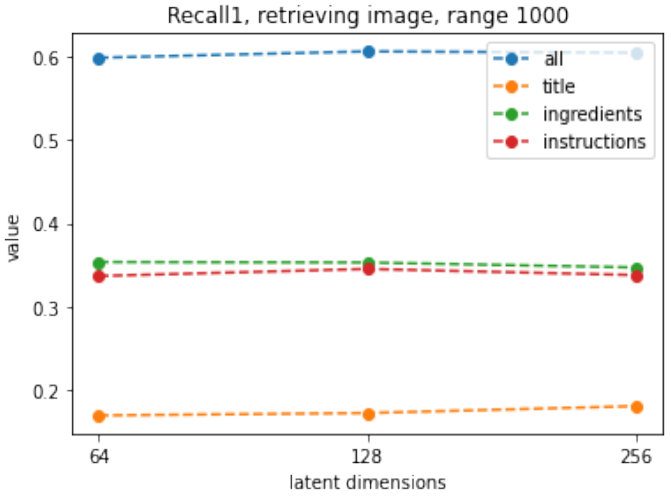
The Recipe 1 million(R1M) dataset is used for all the models developed developed throughout the scope of this project. This dataset consists of $\sim$1M text recipes that contain titles, instructions and ingredients in English. Additionally, a subset of $\sim$ 0.5M recipes contain at least one image per recipe. Data is split in 281598 train, 60422 validation and 60740 test image-recipe pairs, in accordance to the official data release provided in R1M.

4. Evaluation

For Step 2, we evaluate the CCA model on standard retrieval metrics names median rank(medR) and recall rate at top K(RK) as the model has to be evaluated on the retrieval tasks both on text-to-image and image-to-text representation retrieval. If the retrieval performance of the CCA model are of superior quality then the medR is low in value and RK is high in value i.e, RK measures the percentage of true positives being ranked within the top K returned results and inline with previous works we report values at k= 1, 2, 5, 10, 15, 20, 30, 50, 75, 100, 150, 200, 300, 500, 700. Both medR and RK are calculated based on a search pool of either 1k or 10k test samples, with the average over 10 different subset reported. We also perform ablation study wherein we check for various settings of dimensions of shared space. Also for text view, we check which elements from 3 main ones namely: title, ingredient and instruction, has high impact on the performance. We even visualize and gain further insights into the learned embeddings to explain the model's performance behaviour in a sensible manner by visualizing "Muffin" and "Salad"'s latent representation using TSNE.



As we can see, the medR value for the 1k test sample and text to image representation remains nearly constant throughout for all 3 elements and latent dimensions, however it does peak and have the largest value for the 'title' element.



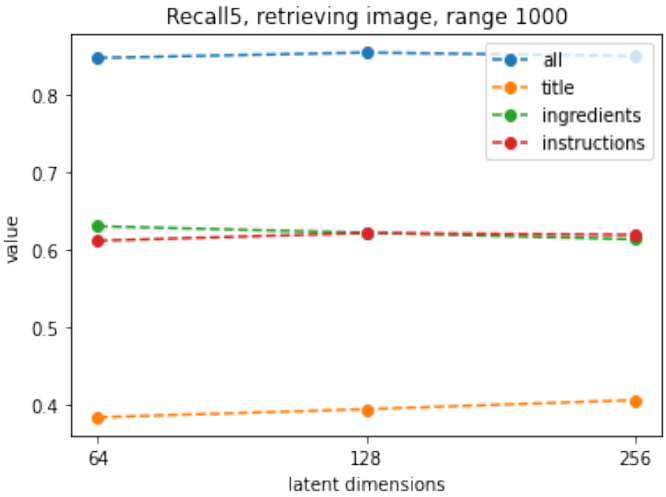
As seen in the above plot, the recall value at k=1 for the 1k test sample and text to image representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.

5. Results and Conclusions

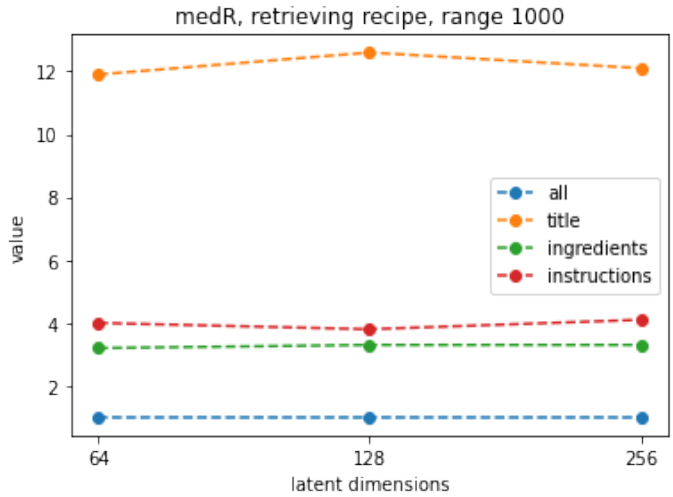
On comparing the results of using triplet loss and MSE Loss, we observe that triplet loss generally gives better results. This might be since we explicitly penalize negative samples to be further away (with the margin). This is not done in MSE loss where we just minimize the "distance" between an anchor and a positive but do not take negative samples into consideration.

Triplet Loss

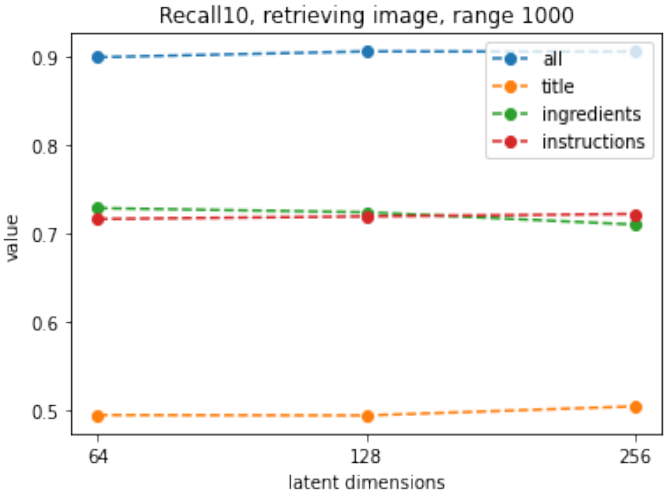
The results and graphs obtained for triplet loss are shown below are shown below:



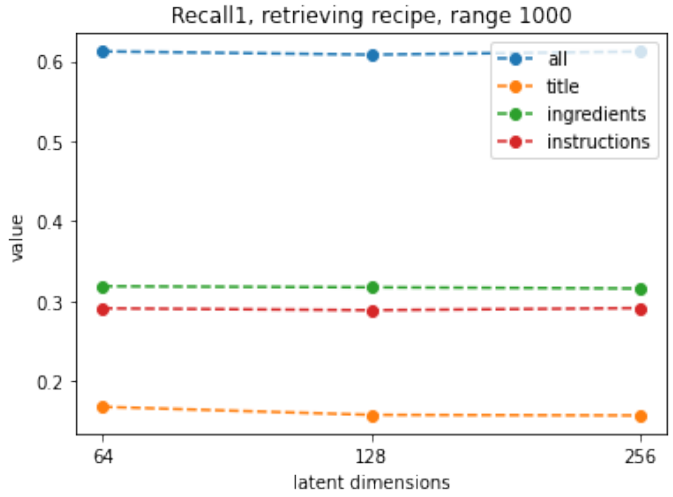
As seen in the above plot, the recall value at k=5 for the 1k test sample and text to image representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



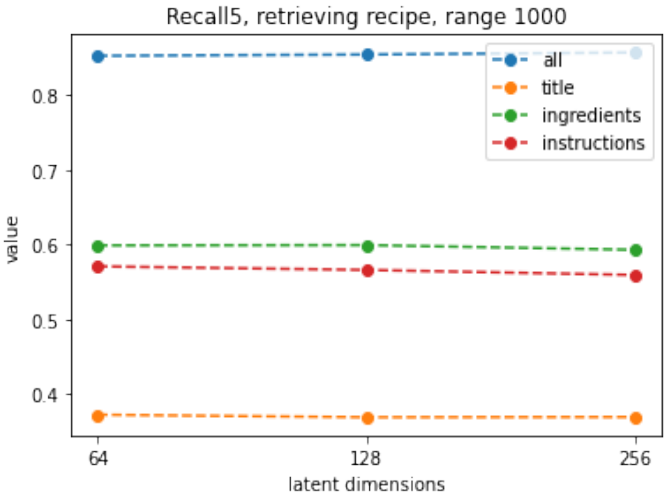
Here, for the 1k test sample and image to text representation, we compare the the medR value and see that it remains nearly constant throughout for all 3 elements and latent dimensions, however it does peak and have the largest value for the 'title' element.



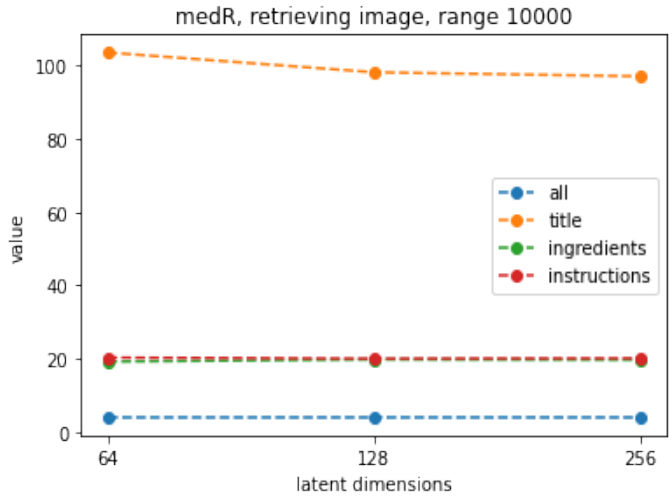
As seen in the above plot, the recall value at k=10 for the 1k test sample and text to image representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



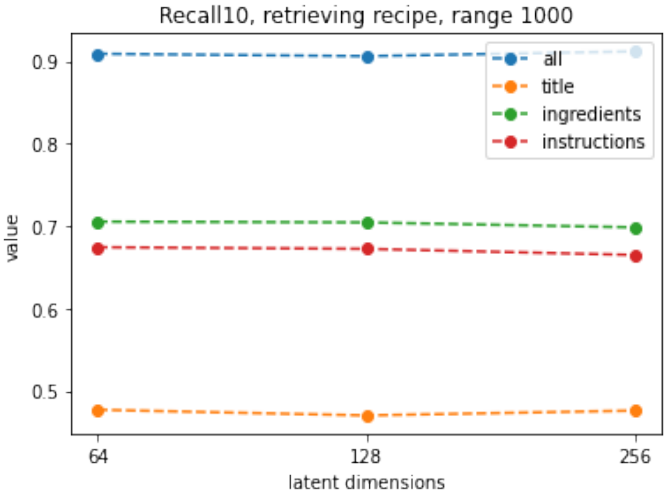
As seen in the above plot, the recall value at k=1 for the 1k test sample and image to text representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



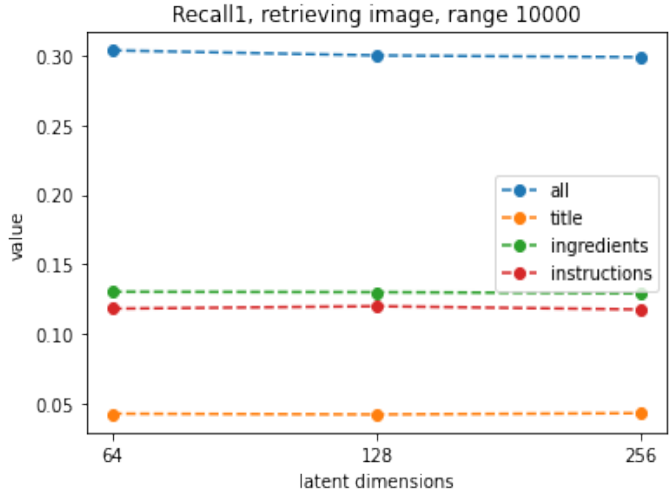
As seen in the above plot, the recall value at k=5 for the 1k test sample and image to text representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



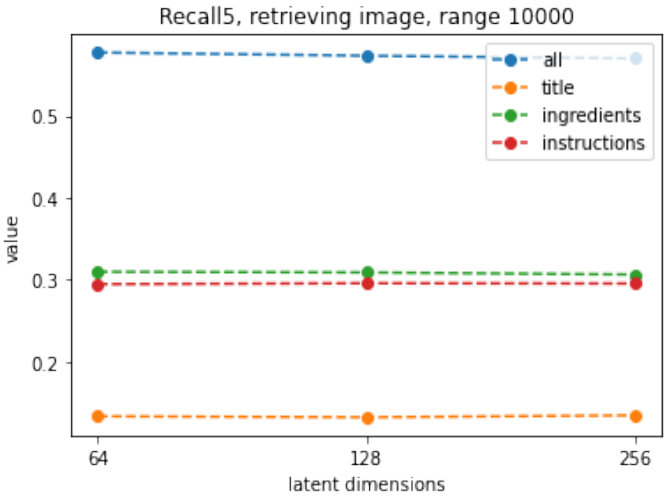
As seen in the above plot, the medR for the 10k test sample and text to image representation remains nearly constant throughout for all 3 elements, however the largest value is obtained for the 'title' element.



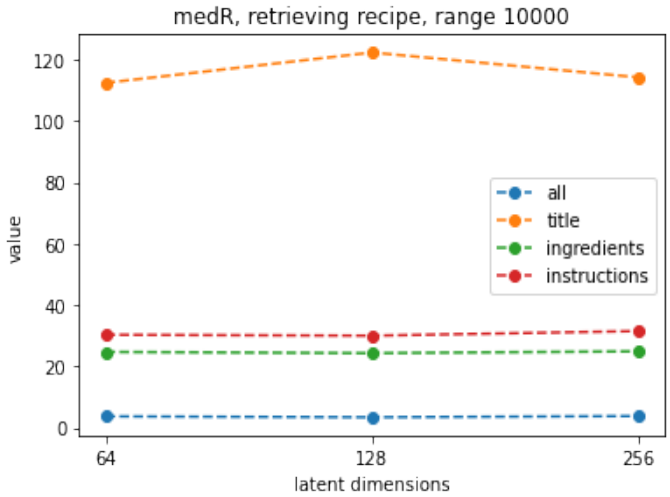
As seen in the above plot, the recall value at k=10 for the 1k test sample and image to text representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



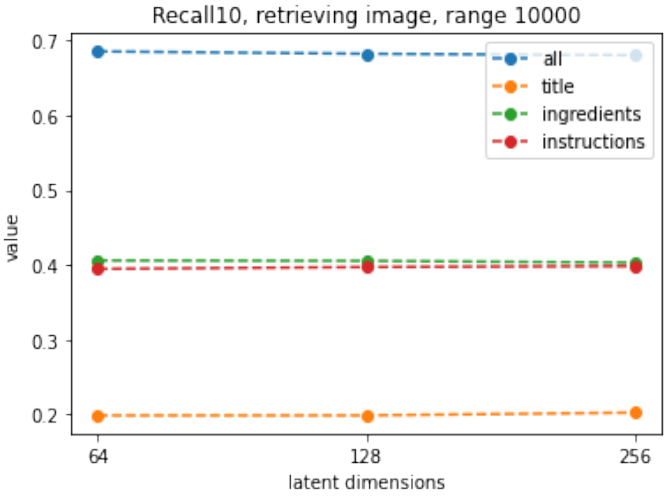
As seen in the above plot, the recall value at k=1 for the 10k test sample and text to image representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



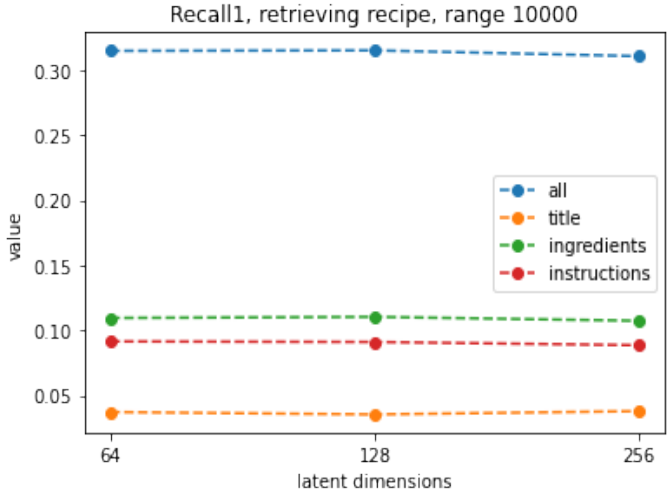
As seen in the above plot, the recall value at k=5 for the 10k test sample and text to image representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



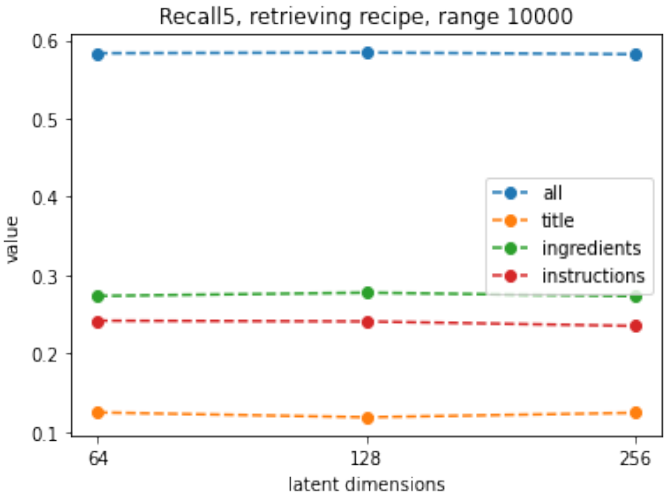
As seen in the above plot, the medR value for the 10k test sample and image to text representation remains nearly constant throughout for all 3 elements, however it does peak and have the largest value for the 'title' element.



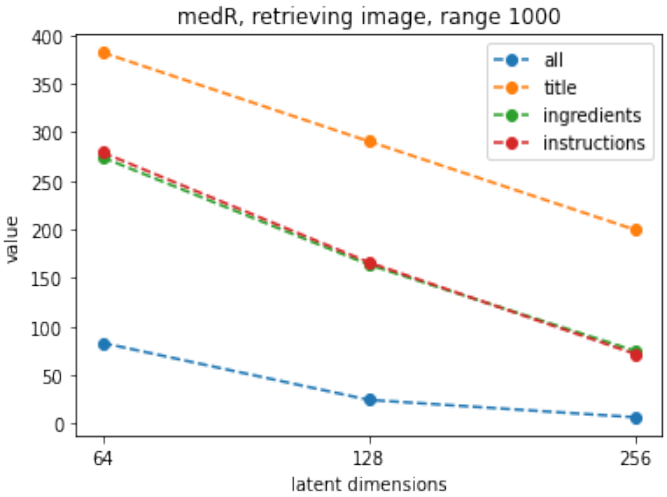
As seen in the above plot, the recall value at k=10 for the 10k test sample and text to image representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



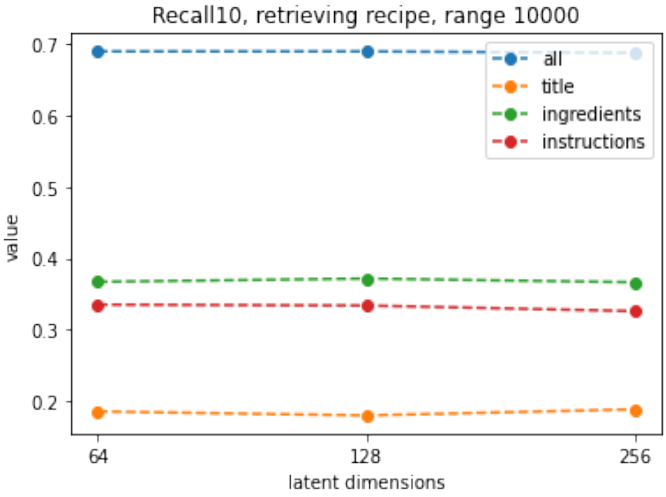
As seen in the above plot, the recall value at k=1 for the 10k test sample and image to text representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



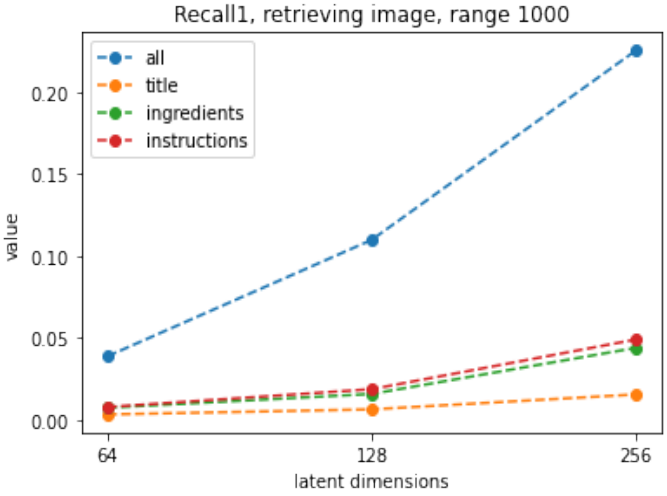
As seen in the above plot, the recall value at k=5 for the 10k test sample and image to text representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



As we can see, the medR value for the 1k test sample and text to image representation decreases with latent dimensions for all 3 elements, however it is observed that the largest value is obtained for the 'title' element.



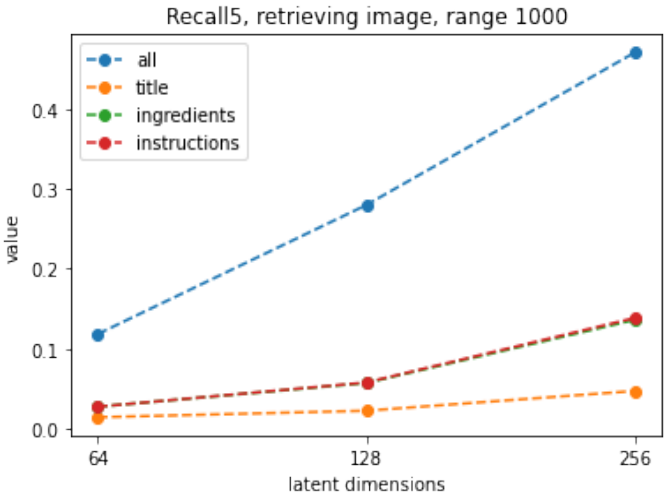
As seen in the above plot, the recall value at k=10 for the 10k test sample and image to text representation remains nearly constant throughout for all 3 elements, however the largest value is for all elements together as opposed to the elements individually.



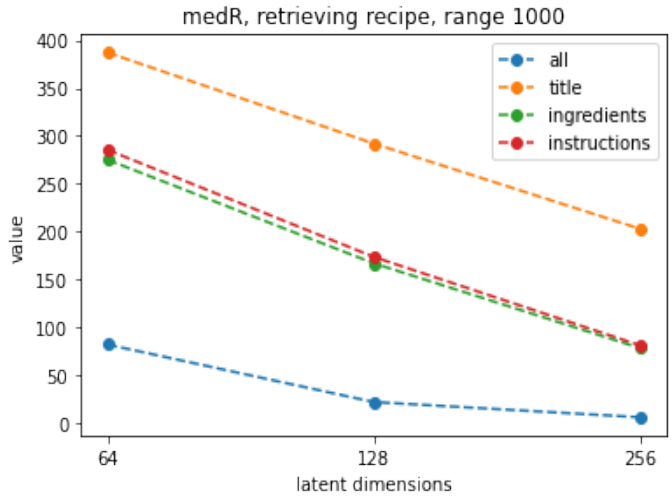
As seen in the above plot, the recall value at k=1 for the 1k test sample and text to image representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.

MSE LOSS

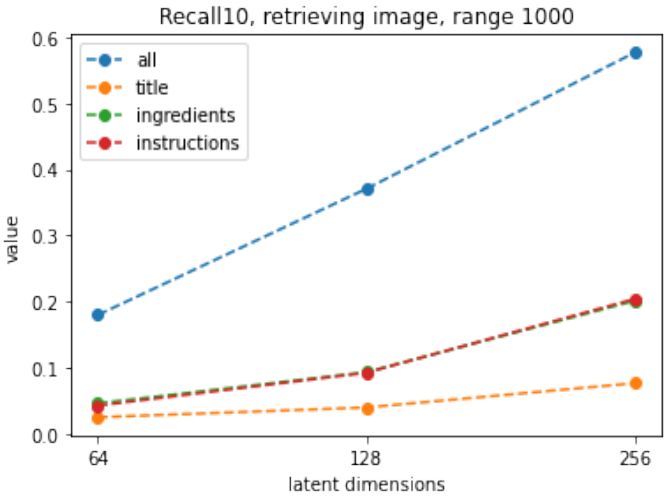
Following are the plots obtained for the MSE loss function



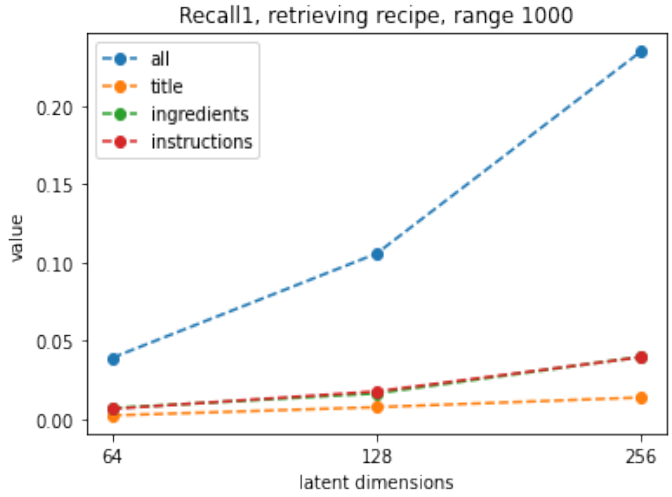
As seen in the above plot, the recall value at k=5 for the 1k test sample and text to image representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



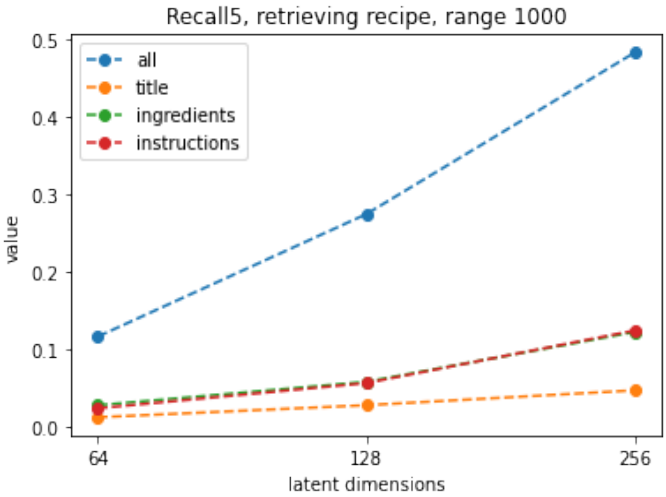
Here, for the 1k test sample and image to text representation, we compare the the medR value and see that it decreases for all 3 elements and latent dimensions, however it is observes that the largest value is obtained for the 'title' element.



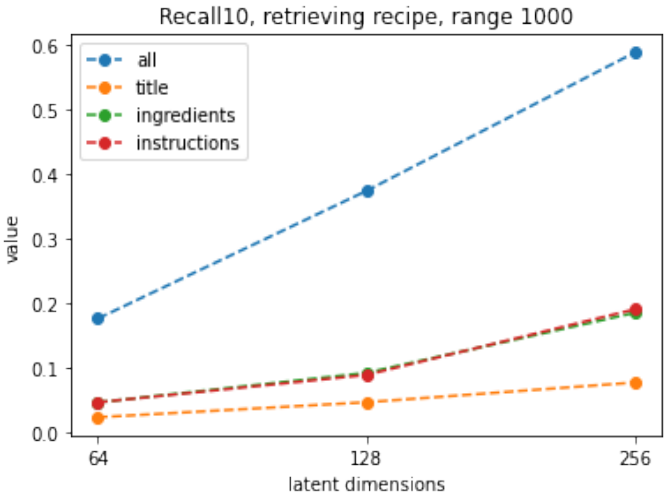
As seen in the above plot, the recall value at k=10 for the 1k test sample and text to image representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



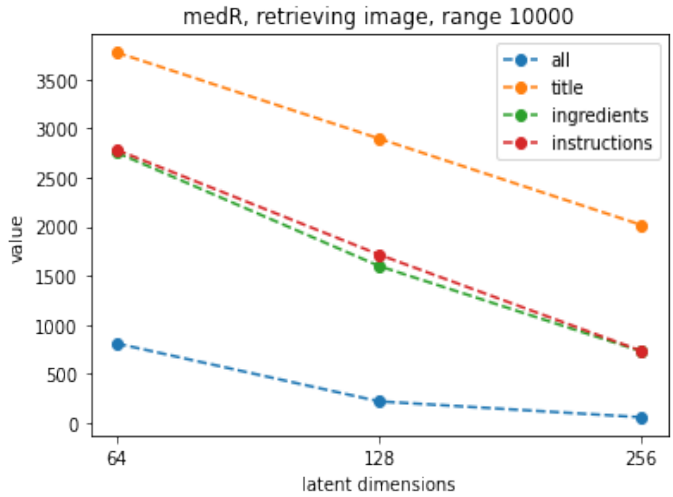
As seen in the above plot, the recall value at k=1 for the 1k test sample and image to text representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



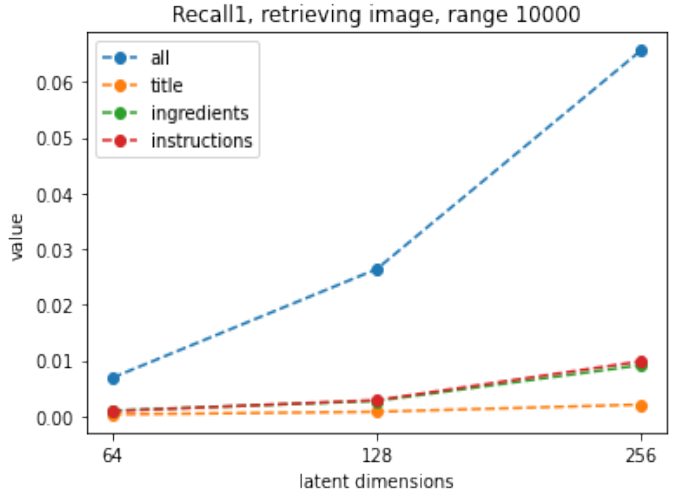
As seen in the above plot, the recall value at k=5 for the 1k test sample and image to text representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



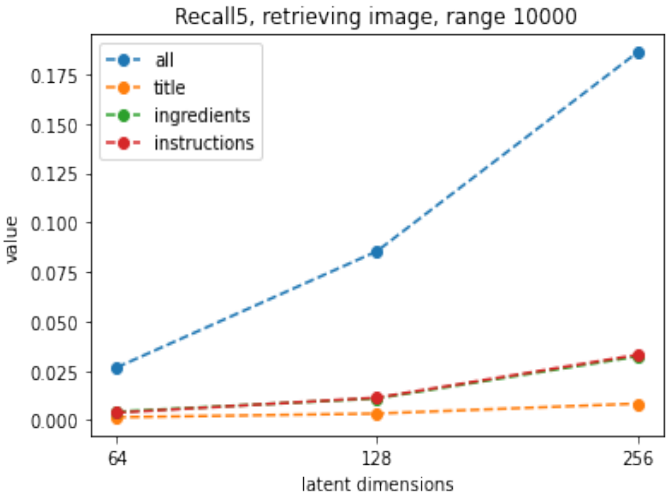
As seen in the above plot, the recall value at k=10 for the 1k test sample and image to text representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



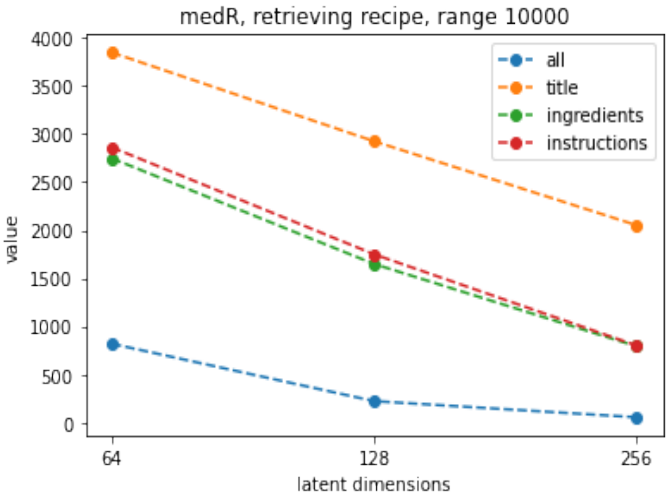
As seen in the above plot, the medR for the 10k test sample and text to image representation remains decreases for all 3 elements, however the largest value is obtained for the 'title' element.



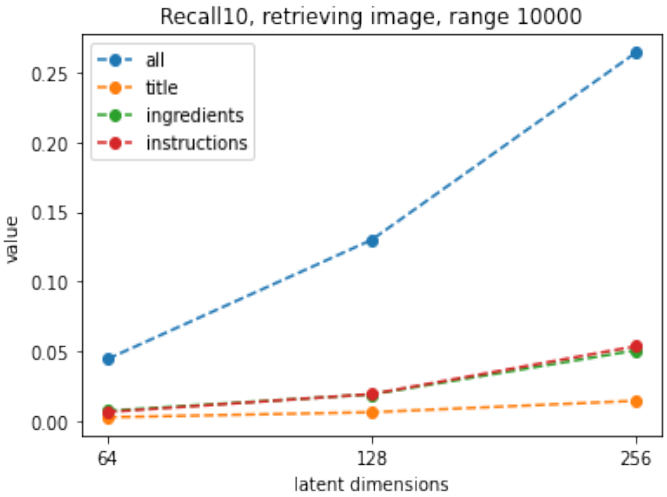
As seen in the above plot, the recall value at k=1 for the 10k test sample and text to image representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



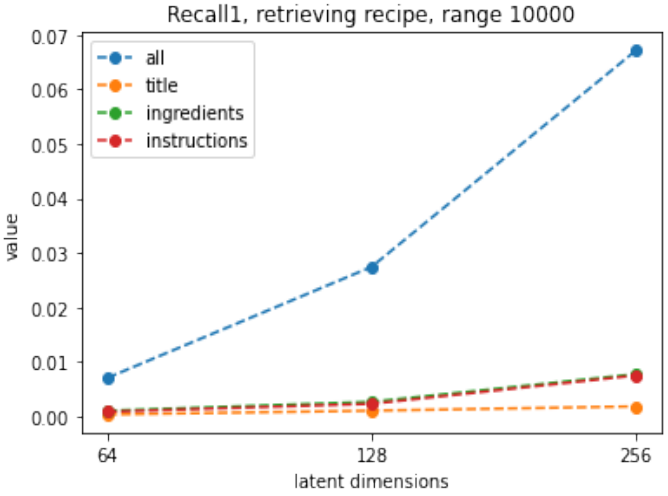
As seen in the above plot, the recall value at k=5 for the 10k test sample and text to image representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



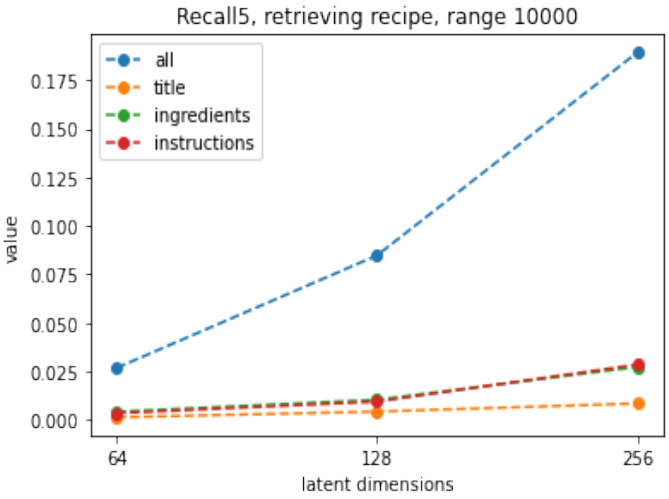
As seen in the above plot, the medR value for the 10k test sample and image to text representation decreases for all 3 elements, however it is observed that the largest value is obtained for the 'title' element.



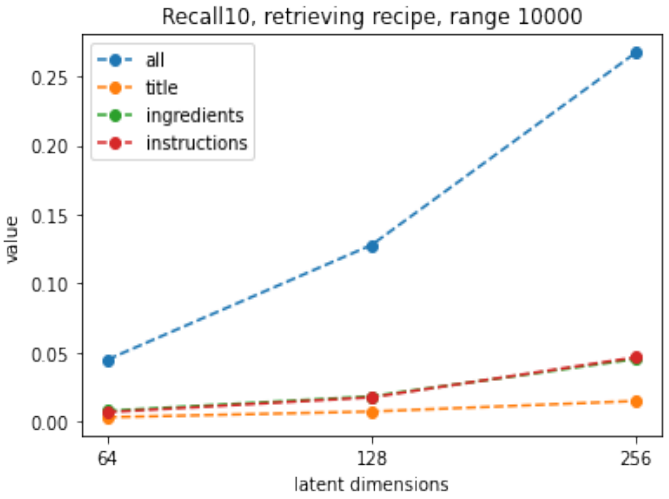
As seen in the above plot, the recall value at k=10 for the 10k test sample and text to image representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



As seen in the above plot, the recall value at k=1 for the 10k test sample and image to text representation increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



As seen in the above plot, the recall value at k=5 for the 10k test sample and image to text representation remains increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.



As seen in the above plot, the recall value at k=10 for the 10k test sample and image to text representation remains increases with latent dimensions for all 3 elements, however the largest value is for all elements together as opposed to the elements individually. The graph for the 'title' element shows a near constant graph though.

We have also visualized the latent dimension of triplet loss and MSE loss by reducing its dimension to 2 using TSNE. We extracted embeddings with "muffin" and "salad" in their title. We reduced the dimension of those latent space features generated by CCA using TSNE. Our model can clearly distinguish between "muffin" and "salad," both in image and text.