Proposal - Cross Modal 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

This project pertains to exploration and creation of different models (both linear and non linear) for the task of cross-model representation learning. The main aim is to develop a model that incorporates the techniques used to extract Cross-Model Representations on text-image pairs which is now gaining booming popularity in vision-language tasks. The project starts with studying and exploring the task of classical multi-view representation learning using Canonical Correlation Analysis- a linear model for paired data to extract shared representation between features in text and images. The project extends its reach in its successive step by incorporating non-linearity while building upon the same linear CCA multi-view model used in step 1. This step 2 aims at deriving superior gains as compared to those obtained in step 1 using the touchstone linear CCA model. This is done by leveraging non-linear deep models and also investing the usage of triplet loss.

In the concluding step, we explore and compare 2 new models [1] and [5]. The first model uses a kind of zero-shot learning method and improve on the robustness of the network and make our model more representative while dealing with task of fetching or learning visual concept from natural language supervision. The second model increases the reach of the popular BERT architecture to process 2 streams of representations(textual and visual representations) individually which later work in cohesion through co-attentional transformer layers.

So we in-turn create a model for learning task-agnostic shared representations of visual concepts and natural language. Thus through this model, we try to emphasis the importance of considering the visual grounding as pretrainable and transferable mechanism.

1. Introduction

In real world, information about a 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 itself tells us that the information is seen from multiple views(e.g. text, video and images). More colloquially, cross-modal 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 information 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.

2. Prior Work

2.1. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images [9]

This paper introduces **A**dversarial **C**ross **M**odal **E**mbedding (ACME). The paper identifies and provides solutions for the following:

1. High variation among images belonging to the same recipe making it difficult to converge using a naive sampling strategy. Hard sample mining is employed to increase convergence speed.
2. Being from heterogeneous modalities, the feature distributions can be very dissimilar. This approach tries to align these different distributions using an adversarial loss given by:

$$L_{MA} = \mathbb{E}_{i\sim p_{image}}[log(D_M(E_V(i)))] \\ + \mathbb{E}_{r\sim p_{recipe}}[log(1 - D_M(E_R(r)))]$$

   which is solved by a min max optimization as:

$$min_{E_V,E_R}(max_{D_M}(L_{MA}))$$

3. Since the embeddings might lose information during the embedding process, To try and remedy this, a cross modal translation consistency component where food images are generated using recipe text features, and the ingredients of a recipe are predicted using the image features.

2.2. Revamping Cross-Modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning [8]

This paper introduces a hierarchical transformer architecture. There are several additions like

1. Transformers are used for both modalities.
2. For the text encoder, the recipe text is divided into a sequence of sentences. Each sentence is first processed by a transformer network. The resulting sequence of sentence representations is then passed as an input sequence to a second transformer network with the same architecture but different parameters. This creates the ”hierarchical” structure of the model.
3. The recipe text is used for input as is, without any sort of preprocessing or pretraining. Additionally, the entire dataset is used for training as opposed to just the recipes which have images. This is done using a self supervised loss that is computed on top of the recipe representations of the model. The objective is to constrain recipe embeddings so that intermediate representations of individual recipe components are close together when they belong to the same recipe and far apart otherwise. This is done by introducing a triplet loss of the form

$$L_{rec}(a, b) = L_{bi}(e_a^{n=i}, \hat{e}_{b\to a}^{n=i}, \hat{e}_{b\to a}^{n\neq i}, e_a^{n\neq i}) \quad (1)$$

   where a and b can take values among title, ingredient and instructions. e is the embedding feature and $e_{b\to a}$ is projected onto another feature space as $\hat{e}_{b\to a}$ using a single linear layer.

2.3. Cross-modal Retrieval and Synthesis (X-MRS): Closing the modality gap in shared subspace learning [3]

This approach uses a simpler method which provides promising results. A ResNet [4] based image encoder is used along with a transformer based text encoder to create the image and text representations respectively. The retrieval model takes a text, image pair. The training objective is to minimize the distance between an anchor and recipe r^+ and a matching image v^+, while also maximizing the distance between the anchor recipe r^+ and a non-matching image v^-. It minimizes the margin triplet loss of (r^+, v^+, v^-). The complete recipe (title + ingredients + instructions) is passed as input to the text encoder without any preprocessing. The recipe text is augmented via back translation from German (en-de-en) and Russian (en-ru-en). Additionally, recipes are translated into German, Russian, French and Korean. These translations are obtained using pretrained models. A multilingual bert model is used as the text encoder giving this architecture a name of T-ML.
Along with the retrieval model, a synthesis model is introduced. This generative model synthesizes novel images of a recipe or visually realizes a recipe without any accompanying images. The model generates an image from the recipe embedding, the similarity of which can be measured using the aforementioned retrieval model.

3. Technical Details

This project focuses on development and exploration of different 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.

3.1. Step 1

The first step in the project is to build and analyse a linear model named Canonical Correlation Analysis(CCA) to perform 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 feature extractors at our disposal such as ResNet for image features and BERT [2] for textual features and then employ CCA to learn common representation. As the first step, we link the text and image views of a recipe and so we use the dataset Recipe 1 Million(R1M).

3.2. Step 2

For the second step, we introduce non-linearity on top of our CCA model in an attempt to better incorporate the features and learn even more useful correlations. As CCA is linear in nature, it is more likely that it will fail to deliver optimal results. So to improve upon that, we introduce non-linearity by using non-linear deep models, more colloquially, Deep Canonical Correlation Analysis(DCCA). Deep Canonical Correlation Analysis is a method to learn complex non-linear transformation of 2 views. The resulting representations are highly correlated in linear fashion. It is basically a non-linear augmentation of the linear CCA. Deep CCA outputs representations by passing the representations from 2 views through multiple stacked layers of nonlinear transformation. This way it is more successful in outputting representations with higher degree of correlation than those outputted by a normal linear CCA. We also use triplet loss to maximize the similarity and minimize the dissimilarity between the retrieved representation and the input representation. In simpler words, the triplet loss builds triplets of the form <anchor, positive, negative> where the 'anchor' can be one representation in first view, 'positive' is the similar representation in second view and 'negative' is the dissimilar representation in the second view. Thus the step 2 is actually built upon the linear CCA with the aim to produce better results.

3.3. Final Step

As stated earlier in the abstract, in this final step, we explore 2 models which can be used for the task of cross-model retrieval task. The models are explained as below:

3.3.1 CLIP

Using the techniques of zero-shot transfer, multimodal learning and natural language supervision, we build our final model that aims at bridging the gap between the visual and textual representation. In this concluding step, we build a model that performs the task of generating visual representations only by using the natural language supervision. We also formulate the following proxy training task for our neural network model: inputting the image, the model predicts the an output text snippet which is paired with the image from a pool of many randomly sampled text in our dataset. For this to happen precisely, it is necessary to learn a great number of visual concepts in the images and pair it with a text describing it. So in simpler terms, if the neural network model we build is used for classification problem on the dataset of cat vs non-cat images, the model outputs a prediction to an input image the text snippet like ”image of cat” or ”image of non-cat object” is more plausible to be paired with the image. Our model does the task of pre-training an image encoder and a text encoder to predict which images were paired with which texts in our dataset and then use this to perform the zero-shot classifier. So in essence, we are connecting both the individual space of visual and textual representations by forming a latent space and using that we define which visual concept is in close affinity with which textual concept.

3.3.2 VilBERT

We know that the strategy of using the independent visual and natural language models prepared for other specialized tasks and then as a part of task-training, learn the groundings later on provide us with poor groundings. These crippled low-quality groundings often also fail to generalize well when the image-text data is limited and biased. In the domain of computer vision and natural language processing, this method serves well because of the convenience and the power the large-scale, publicly available pretrained models have to offer. But in the domain of cross-model representation learning which deals with extracting similarity between the visual and textual representation as a subtask, working with the standard technique shows no prudence. In another words, catching only the visual representation is not enough if the downstream vision-language model fails to generate relationship between the textual and visual concepts or representations. Thus we explore this domain and prepare a model that uses the concepts of self-supervised learning that enables the model to perform proxy tasks and evidently capture the co-relation between the visual and textual representation. Our joint model aims at learning shared image-text representation by using a paired dataset consisting of both image and relating text snippets. Developing upon the famous and successful architecture of BERT we try to generate separate yet later on shared spaces for processing the visual and textual representations. The separate space is formulated to respect the individual and independent processing needs of each modality and the shared spaces(enabled using transformer layer) allows them to jointly create a common latent space to work with at differing levels of complexity. So in essence, we are connecting both the individual space of visual and textual representations by forming a latent space and using that we define which visual concept is in close affinity with which textual concept.

4. Evaluation

The dataset used and the evaluation strategy selected to derive the performance of each individual models formulated in Step 1 through Final Step is stated as below.

4.1. Dataset Used

The Recipe 1 million(R1M) dataset [6] is used for all the models developed developed throughout the scope of this project. This dataset consists of ∼1M text recipes that contain titles, instructions and ingredients in English. Additionally, a subset of ∼ 0.5M recipes contain at least one image per recipe. Data is split in 238999 train, 51119 validation and 51303 test image-recipe pairs, in accordance to the official data release provided in R1M. For the final step, we also employ the dataset r-FG-BB in [7] to solve the problem of lack of correspondence labels in R1M dataset necessary for localization evaluation. We also plan to come up with a solution to bridge the gap between the unlabeled (R1M) and the labeled (r-FG-BB) dataset that contains Japanese text incompatible with English-trained models on R1M, which is necessary for quantitative evaluation.

4.2. Evaluation strategy

For all the steps, we evaluate the models on standard retrieval metrics namely median rank(medR) and recall rate at top K(RK) as the models 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 than 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, 5, 10. 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.

4.3. Visualization

For the final step, we even visualize and gain further insights into the learned embeddings to explain the model's performance behaviour in a sensible manner. This is done by observing how the latent space (created by the Step 1, Step 2 and Final Step models) is structured. This task of visualization is done by employing TSNE- a dimensionality reduction technique. By using TSNE, we would get a clear idea of how a group of representation sharing similar attributes are cluttered together and how distant they are from the representations having contrasting attributes. By using this technique, we can get a visual analytic data of how the latent space is distributed across all those output representations under inspection.

5. Timeline

Please refer to the Gantt Chart shown in figure 1.

References

[1] M. V. Conde and K. Turgutlu. Clip-art: contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3956–3960, 2021. 1

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[3] R. Guerrero, H. X. Pham, and V. Pavlovic. Cross-modal retrieval and synthesis (x-mrs): Closing the modality gap in shared subspace. *arXiv preprint arXiv:2012.01345*, 2020. 2

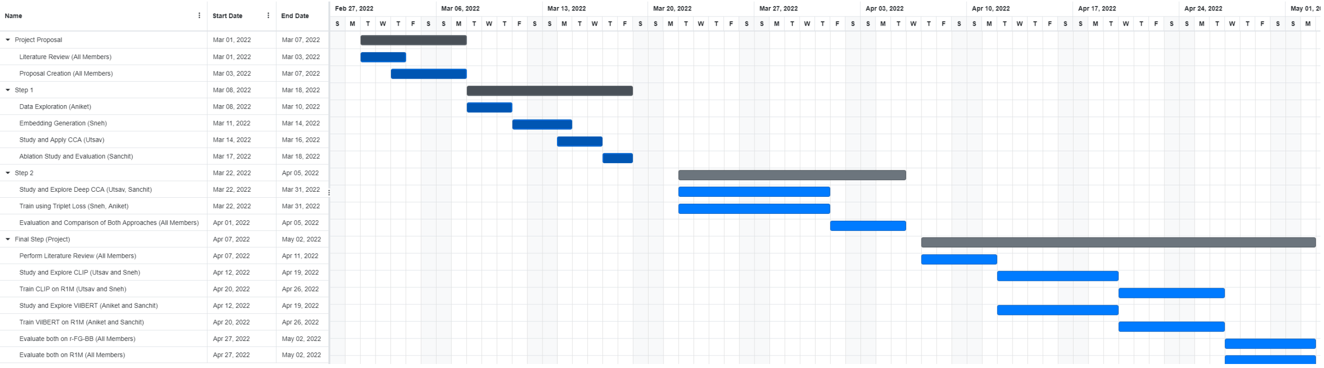
[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[6] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):187–203, 2019. 4

[7] T. Nishimura, S. Tomori, H. Hashimoto, A. Hashimoto, Y. Yamakata, J. Harashima, Y. Ushiku, and S. Mori. Visual grounding annotation of recipe flow graph. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4275–4284, Marseille, France, May 2020. European Language Resources Association. 4

[8] A. Salvador, E. Gundogdu, L. Bazzani, and M. Donoser. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15475–15484, 2021. 2



Name	Start Date	End Date
▾ Project Proposal	Mar 01, 2022	Mar 07, 2022
Literature Review (All Members)	Mar 01, 2022	Mar 03, 2022
Proposal Creation (All Members)	Mar 03, 2022	Mar 07, 2022
▾ Step 1	Mar 08, 2022	Mar 18, 2022
Data Exploration (Aniket)	Mar 08, 2022	Mar 10, 2022
Embedding Generation (Sneh)	Mar 11, 2022	Mar 14, 2022
Study and Apply CCA (Utsav)	Mar 14, 2022	Mar 16, 2022
Ablation Study and Evaluation (Sanchit)	Mar 17, 2022	Mar 18, 2022
▾ Step 2	Mar 22, 2022	Apr 05, 2022
Study and Explore Deep CCA (Utsav, Sanchit)	Mar 22, 2022	Mar 31, 2022
Train using Triplet Loss (Sneh, Aniket)	Mar 22, 2022	Mar 31, 2022
Evaluation and Comparison of Both Approaches (All Members)	Apr 01, 2022	Apr 05, 2022
▾ Final Step (Project)	Apr 07, 2022	May 02, 2022
Perform Literature Review (All Members)	Apr 07, 2022	Apr 11, 2022
Study and Explore CLIP (Utsav and Sneh)	Apr 12, 2022	Apr 19, 2022
Train CLIP on R1M (Utsav and Sneh)	Apr 20, 2022	Apr 26, 2022
Study and Explore VilBERT (Aniket and Sanchit)	Apr 12, 2022	Apr 19, 2022
Train VilBERT on R1M (Aniket and Sanchit)	Apr 20, 2022	Apr 26, 2022
Evaluate both on r-FG-BB (All Members)	Apr 27, 2022	May 02, 2022
Evaluate both on R1M (All Members)	Apr 27, 2022	May 02, 2022

Figure 1. Timeline Chart.

[9] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S. C. Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11572–11581, 2019. 2