Multi-modal Chaining

Hassan Fahmy
Department of Computer Science
Stanford University
hassanf@stanford.edu

Abstract

The idea of this paper is to train encoders to generate the same embedding for the same item if the input is from different modalities(i.e. a picture of a golden retriever and the words " a golden retriever" should generate similar embeddings. We show that it can be done and that with two encoders that are trained to match image to text we can match multi-spectral satellite data to text with the same accuracy by training the multi-spectral embeddings to be similar to the image embeddings.

1 Introduction

Contrastive learning is a self supervised learning technique that allows us to learn general features of the data without labels by teaching the model which data points are similar or different. It has shown great success in multimodal correspondence like predicting the correspondence between audio and image from a video or an image and its caption which is what Google's contrastive learning image pretraining model (CLIP) does. Which is a model consisting of an image encoder and a text encoder that are trained to output similar embeddings for an image and its caption. The idea of this project is to expand on CLIP to generate representations from different modalities in the same embedding space based using contrastive learning. If we train an audio encoder to generate audio embeddings that are similar to the image embeddings generated by CLIP, we then have audio embeddings that are similar to the text embeddings, and thus we can do audio-text tasks using those embeddings without having an audio-text dataset. For example we can do zero shot classification by comparing audio and text embeddings. That was all work I have done on this at the COCO lab. For this project I am continuing on that work with multi-spectral data, matching multi-spectral embeddings to image embeddings and then hopping from there to text. I made an encoder that takes in multi spectral data as input and returns a 512 dimension embedding just like the one returned by the CLIP model. Doing that for multi spectral data has many open ended applications in the fields of robotics and autonomous vehicles and in perception in general, because it can allow us to have one perception model that is trained on those embeddings and we can use the encoders to process the data and have the model work with many different modalities.

2 Related work

This area of work is very new. The CLIP paper by Google [1] that this work is based on came out earlier this year. Their have multiple models that are based on Resnets and Visual transformers. Since then another paper was submitted to a conference last month that did something very similar with what I did with audio dubbed WAV2CLIP[2], their approach was very similar to what I did with audio specifically, and general similar to what I am doing for this project with multi spectral data. In general the idea of matching different modalities to each other without the general idea of the common embedding space has been explored a lot. Notably in Soundnet[3] for example. Which does

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

correspondence mapping on YouTube videos.
It is however in creating common representations and embeddings for data from all different modalities that this paper has its value.

3 Dataset and Features

I use the Eurosat dataset which contains multi-spectral data in 12 spectral bands with varying resolutions from the Sentinel-2 satellite. The spectral bands include visual light, and each data point has a land use label. This in essense gives us 3+ modalites at each datapoint namely image, label text, and multi spectral data which can be treated as more than one modality if we consider data from either end of the spectrum.
The multi-spectral data can be extracted as 64 by 64 images with 13 channels.
I use the images and the multi spectral data to do the contrastive training between the CLIP embeddings for the RGB bands and embeddings I generate for other spectral data. The dataset has 27000 samples I use 2600 samples for training and 1000 samples for validation, with the target being matching multispectral data to the corresponding RGB image. Then for testing, I use the CLIP embeddings for the text labels to match that to the embeddings for the multi spectral data and compare the difference in accuracy when matching RGB data to text and matching multispectral data to text.

4 Methods

In order to build the multispectral data encoder I start by modifiying a pretrained version of Resnet-50. Resnet-50 is pretrained to classify RGB images into one of 1000 labels. It has multiple convolutional layers then a pooling layer and a linear layer to do the classification. Removing the last layer then flattening the output will give us a 2048 dimensional embedding of the photo. I then add a linear layer that maps the 2048 dimenstional embedding to 512 dimensions because that is the size of the embedding generated by clip. I left the input as three channels and feed it data from 3 different spectra with similar resolution.
The loss function I use is a cross entropy loss on the dot product of a batch of the the image embeddings and the multispectral embeddings. The idea is for the embeddings for the same item should be as similar as possible and embeddings for different items should be different. So the labels for the cross entropy function are represented in a diagonal matrix which means that the dot product between the coresponding embeddings should be 1 and the dot product for different items should be zero.
In order to prevent over-fitting I do random crops on both the spectral data and RGB data. With this approach there is a high risk of over-fitting, both in terms of regular over fitting and over-fitting by making features that represent the similarity between the two modalities more salient in the embedding rather than features that represent the content of the data. This type of over-fitting will lead to better accuracy when matching multispectral data to RGB data but worse accuracy when matching multispectral data to text.

5 Experiments/Results/Discussion

The metric used for the final evaluation of the model's performance is the matching accuracy between the satellite data and the text data. Meaning that we dot product the normalized satellite embedding with the normalized text embeddings for all the labels, and the highest dot product is selected. The metric we use to observe how well the model is training is the matching accuracy between the satellite data and the images, which is obtained the same way as the satellite text accuracy, and is a harder problem when the batch size is larger because there are more similar images that it has to differentiate.

The results from this run was done with infrared spectral bands, the aforementioned data augmentation, using a batch size of 100 in training and 512 in validation and Adam optimizer with a learning rate of 0.0001.
These settings were able to get to the highest performance on the satellite text accuracy the fastest, although the large batch size in training has also caused it to drastically over fit and optimize for the small differences between images rather than the bigger differences that constitute differences in land use.
It is observed that as the model over-fits to the accuracy between RGB and multi-spectral data

the accuracy between text and multispectral data goes down. This is slowed down by the data augmentation.

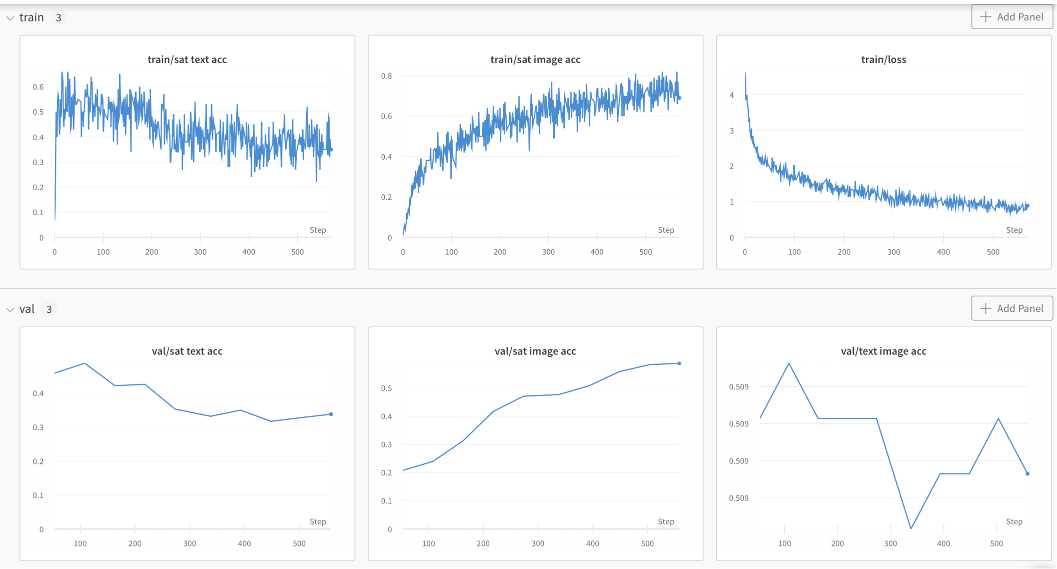
In the validation we can see that the model preforms very close in matching spectral data to text to matching RGB images to text. This shows that it is possible to do the bridging between different modalities.

In this case we are hitting the ceiling on text accuracy early on before overfitting and optimizing for the image accuracy which it is training for.

I experimented with different learning rates between 0.0001 and 0.000001. The larger learning rates converged faster and did not seem to cause any issues in training.

I experimented with small batches like 20 for training and larger batch sizes like 100. The larger batch size had better results in training. It was able to converge faster and perform better as it saw more false or dissimilar examples from different classes in the contrastive training.

Doing qualitative analysis here was interesting because not all what is visible in the spectral data is visible in images and vice versa. It was almost impossible for me as a human to tell the differences or similarities between different data points let alone what category of image it was. Factors like weather, cloud coverage and temperature of water and land had large effects on the way the spectral data looked. That was more the cases for natural cases like images including water and crops. Those were the images that really threw the model off. While images with residential or industrial pieces of land were much more consistent across the different spectra.



6 Conclusion/Future Work

In summary I have shown that we can use contrastive learning to build models that generate corresponding embeddings from multiple modalities, and that can be used to preform zero shot classification tasks to a high accuracy as high as the accuracy between the modality it was trained with and the other modalities.

For future work I think the strongest directions to go into would be to use those encoders as preprocessing that can take in any modality and pass it on to an arbitrary model that would in turn be ambivalent about the modality of the input data and expand the number of encoders for different modalities and start testing the zero shot performance on modalities that are several hops away from each other.

7 Contributions

This work was all based on/part of work that I have done under the supervision of Alex Tamkin a PhD student in the COCO lab. I worked under the supervision of Alex as well with Chris Waites a fellow CS masters student on the audio-image training that was discussed in the introduction.

References

[1] AlecRadford,JongWookKim,ChrisHallacy,AdityaRamesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, et al., “Learning transferable visual models from natural language supervision,” ICML, 2021.

[2] Ho-Hsiang Wu1, Prem Seetharaman, Kundan Kumar, Juan Pablo Bello, "WAV2CLIP: LEARNING ROBUST AUDIO REPRESENTATIONS FROM CLIP", Submitted to ICASSP 2022

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, “Sound- net: Learning sound representations from unlabeled video,” NeurIPS, vol. 29, pp. 892–900, 2016. [4] Paszke, Adam and Gross, Sam and Massa, Francisco and Lerer, Adam and Bradbury, James and Chanan, Gregory and Killeen, Trevor and Lin, Zeming and Gimelshein, Natalia and Antiga, Luca and Desmaison, Alban and Kopf, Andreas and Yang, Edward and DeVito, Zachary and Raison, Martin and Tejani, Alykhan and Chilamkurthy, Sasank and Steiner, Benoit and Fang, Lu and Bai, Junjie and Chintala, Soumith, "PyTorch: An Imperative Style, High-Performance Deep Learning Library"