EECS 598-008 DL-CV Project Report: Language Supervised Vision Pre-training for Fine-grained Food Classification

Kemmannu Vineet Rao, Kshama Nitin Shah
{kemmannu,kshama}@umich.edu

1 Problem Statement

We worked on the problem of fine-grained classification of food by leveraging a vision and language pre-training model, which would be trained on culturally diverse food images that are sub-sampled from the corresponding sub-reddits present in the Reddit using data downloader used in RedCaps dataset[3]. Further, this fine-grained food classification task can be used in applications such as better recommendations on social media and self reported dietary tracking applications.

2 Related Work

Desai *et al.*[3] introduces RedCaps, a large scale dataset of 12 M image-text pairs collected from Reddit. The VirTex-v2 Model uses image captioning on Reddit data as a pre-training task to learn representations between the visual features of the image and textual information describing it. Using these representations, the authors [3] discard the transformer textual head and use the visual backbone ResNet-50 to finetune on several downstream tasks such as image classification. Vision and Language pre-training involves learning of joint multi-modal representations of image content and natural language that can be transferred to various downstream tasks such as Visual Question Answering, Image Classification, Object Detection etc. Since the inception of BERT [4], many recent methods make use of transformers to jointly learn image-text pairs. These methods using vision and language pre-trained models for tasks has been proven to produce fruitful results for downstream tasks when transferred[13, 9, 7, 12, 6, 2, 15, 8].

Li *et al.*[5] trains visual n-gram models on the YFC-100M dataset[14] to predict arbitrary phrases given the content of an image which has proven to be successful in tasks such as zero-shot transfer.

Sariyildiz *et al.*[11] used image-conditioned masked language modeling to learn visual representations from scratch which are successful in various downstream tasks. Sariyildiz *et al.*.[11] also successfully leveraged image captions to introduce both global and localized semantic information into visual representations.

3 Datasets

For training we used a subset of the redcaps dataset [3]. We wrote a shell script to download images and annotations from Reddit for the years 2017-2021 using the dataloader used to download redcaps [3]. The subreddits we used to download our training dataset are as follows: r/breakfast, r/breakfastfood, r/budgetfood, r/chinesefood, r/dessert, r/dessertporn, r/dixiefood. r/food, r/foodporn, r/healthyfood, r/knightsofpineapple, r/mexicanfood, r/tastyfood, r/todayiate, r/todaysbreakfast, r/tonightsdinner, r/veganfoodporn which contained 867,656 image-text pairs. Since it was a huge dataset, to manage the computatinal and time constraint which did not suit our purpose, we were forced to subsample the dataset - one with 700k images containing the subreddits r/food, r/foodporn and the other subsample containing 50k images with the subreddits -r/budgetfood, r/breakfast, r/veganfoodporn.

For evaluation on the downstream task, we trained and evaluated our fine grained classification model (zero-shot classifier) on the Food-101 which has 750 train images and 250 test images per category[1] benchmark dataset.

It is to be noted that we did not evaluate on the Food-X 251 dataset as we fell short of computational resources and time.

4 Computational Resources Used

We both have subscriptions to Google Colab Pro which we used to overfit the VirTex model. Once we could overfit the model, we transferred it to GreatLakes cluster available at the University of Michigan and trained the model on 4 GPUs with 4 CPU's in each GPU.

5 Proposed Method

Our approach was similar to how the VirTex-v2 model was used for downstream tasks but we first sub-sampled the RedCaps dataset to generate a dataset that is specific to our downstream task that we are targeting. Then, we scaled down VirTex-v2 model accordingly to reduce the memory and computational footprint by replacing the ResNet-50 with RegNetX-800MF, we also experimented with different number of transformer layers. Next, we performed joint optimization from scratch on the sub-sampled dataset as a pre-training task using the scaled down architecture. Finally, we used the following methods to tackle the challenge of fine-grained food classification task:

1) Zero shot classification for our downstream task of fine-grained classification using Prompt Engineering.

2) We also performed fine tuning on the visual backbone to tackle the challenge of fine grained classification.

6 Experimental Results and Analysis

Training Details

Initially, the codebase was set up to accept COCO image-text pairs.So,we modified the pre-training VirTex code in **readers.py** and **captioning.py** to load the image-text pairs from the Reddit data that we scraped. Then, we created a tokenizer with food images that was generated from all the captions we had from the subsampled dataset we had scraped. This file is named food_10k.model. This tokenizer file contains 10k words, it includes lowercase words but ignores accented words. After this, we changed the model from a ResNet-50 to a smaller version of a RegNet-X-800MF in the visual_backbones.py file. A ResNet-50 outputs a feature vector of size 2048 while a RegNet-X-800MF outputs an image feature vector of size 672. In this way, we significantly scaled down the architecture of the visual backbone. We experimented with varying the number of transformer layers with L $\in$\{1,2,3\}. We also experimented with two different subsets of the dataset.

The 2 subsamples we created are:

1. Using 50k images with image-text pairs from the following sub-reddits : -r/budgetfood, r/breakfast, r/veganfoodporn

2. Using 700k image-text pairs from the subreddits - r/food, r/foodporn from the years 2017-2021

The following are the details of the pre-training experiments conducted and their respective results:

1) **Model 1**: **Visual Backbone:** RegNet-X-800MF, **Textual Backbone, Number of Transformer Layers:** 1, No. of iterations: 250,000, Subset of the dataset: 50k images See Figure 1.

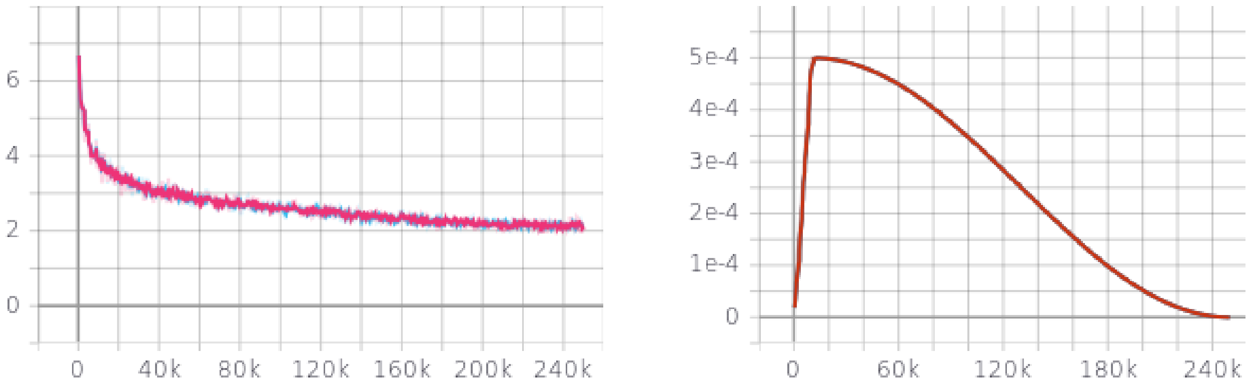


Figure 1: Training Loss Curve for Model 1 (left) and Learning Rate Curve for Model 1 (right)

2) **Model 2**: **Visual Backbone:** RegNet-X-800MF, **Textual Backbone, Number of Transformer Layers:** 2, No. of iterations: 160,000, Subset of the dataset: 700k images . See Figure 2.

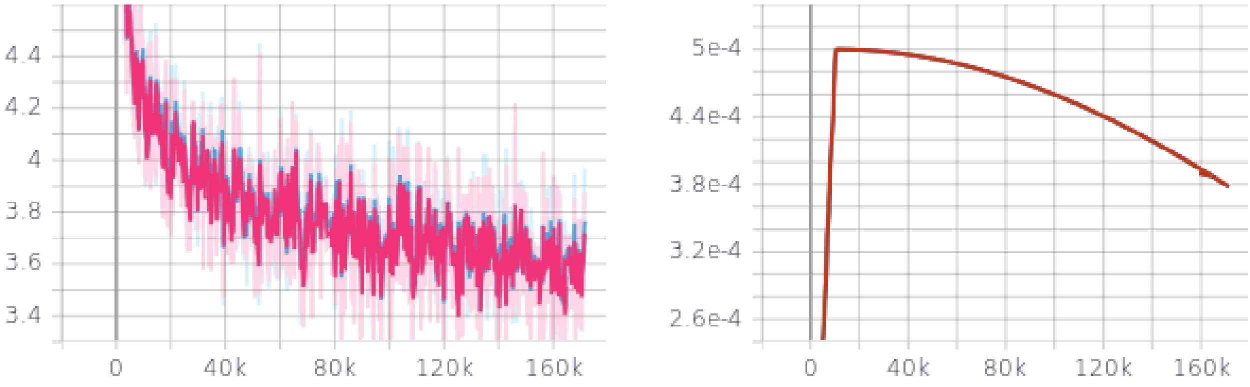


Figure 2: Training Loss Curve for Model 2 (left) and Learning Rate Curve for Model 2 (right)

3) **Model 3**: **Visual Backbone:** RegNet-X-800MF, **Textual Backbone, Number of Transformer Layers:** 3, No. of iterations: 160,000, Subset of the dataset: 50k images. See Figure 3.

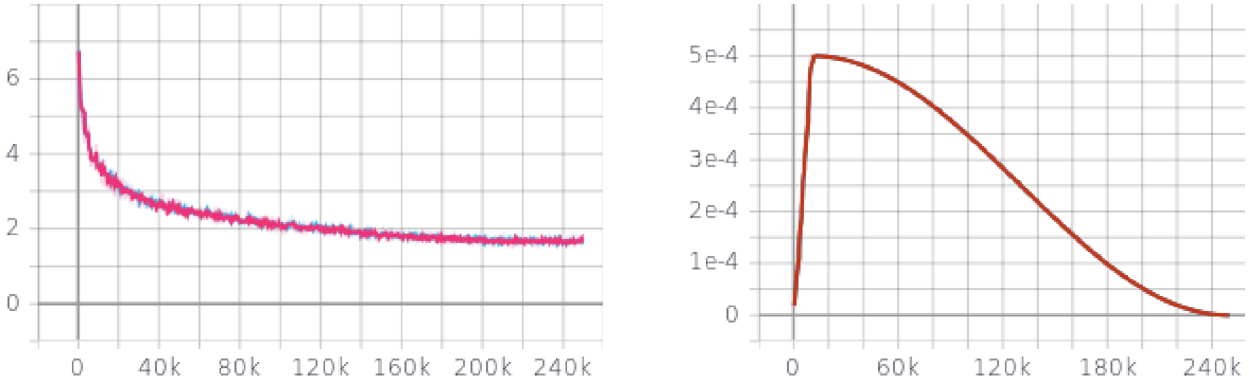
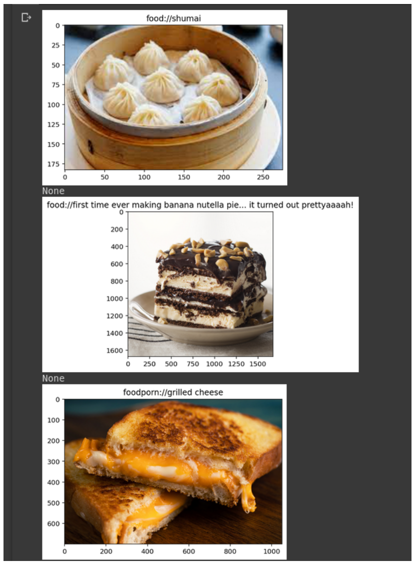


Figure 3: Training Loss Curve for Model 3 (left) and Learning Rate Curve for Model 3 (right)

From this, we can see that increasing the number of transformer layers along with scaling up the size of the dataset can help the model better learn representations.

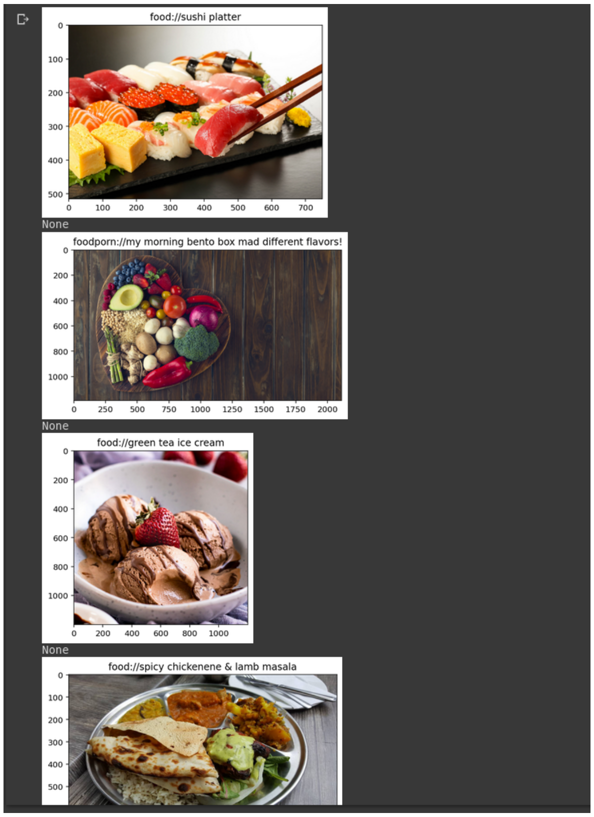
Inference Results

We also ran inference on our best performing model to evaluate the captions generated by the model on images picked randomly from Google Images. Below are the results of the above experiment:



```
food://shumai
None
food://first time ever making banana nutella pie... it turned out prettyaaaah!
None
foodporn://grilled cheese
```

Figure 4: Results of Inference on our best performing model



```
food://sushi platter
None
foodporn://my morning bento box mad different flavors!
None
food://green tea ice cream
None
food://spicy chickenene & lamb masala
```

Figure 5: Results of Inference on our best performing model

Zeroshot Evaluations:

All our configurations are done with sub-prompt:"food"

The next experiments were performed on the downstream task of zero-shot classification: We experimented with different prompts picked based on the most commonly occurring tag on the subreddits and a CLIP like prompt. It is to be noted that these experiments were performed on a subset of the **test dataset(10 %) 2500 images randomly sampled** due to computational constraints. The results obtained for these experiments are as follows:

Model	Prompt	Top 1 % Acc	Top 5 % Acc
Model 1	Homemade <cls name>	1.32 %	5.2%
Model 1	I ate <cls name>	1.6 %	5.28%
Model 1	OC <cls name>	1.24 %	5.4%
Model 1	a photo of <cls name>, a type of food	0.8 %	6.04%
Model 1	homemade <cls name>	1.08 %	5.04%
Model 2	Homemade <cls name>	2.0 %	18.96 %
Model 2	I ate <cls name>	3.28 %	17.6 %
Model 2	OC <cls name>	7.24 %	19.92 %
Model 2	**a photo of <cls name>, a type of food**	**7.16 %**	**23.76%**
Model 2	homemade <cls name>	3.24 %	17.6 %
Model 3	Homemade <cls name>	0.92 %	5.08 %
Model 3	I ate <cls name>	1.84 %	6.72 %
Model 3	OC <cls name>	1.8 %	8.36 %
Model 3	a photo of <cls name>, a type of food	0.84 %	5.52 %
Model 3	homemade <cls name>	0.92 %	5.8 %

Table 1: Zeroshot accuracy results for different prompts

From these experiments, we can conclude that the prompt used in CLIP ("a photo of <cls name>, a type of food) and the model that was trained on 700k images with general food images rather than subsampled from specific subreddits yielded better results when compared to the other configurations and prompts.

We also run our best model through entire test dataset on greatlakes and here are the following results.

Model	Prompt	Top 1 % Acc	Top 5 % Acc
Model 2	Homemade <cls name>	2.15 %	17.7544 %
Model 2	I ate <cls name>	3.21 %	16.8 %
Model 2	OC <cls name>	6.804 %	19.89 %
Model 2	**a photo of <cls name>, a type of food**	**6.85 %**	**22.47%**
Model 2	homemade <cls name>	3.31 %	18.26 %

Table 2: Zeroshot accuracy of our best model on entire Test Set

Our best model could achieve 6.85 % top 1 accuracy and 22.47 % top 5 accuracy.

There are some interesting results and extrapolations which we can infer from the following results-

1). The top 5 and top 1 accuracy gap is significant, we believe this is due to the fact that we did not pretrain the VirTex model for long enough, if we did so the gap would be much lower.

2). The test accuracy on the entire dataset is almost similar to the test accuracy on the subsampled dataset, which means that the model is scalable to higher classes and dataset sizes.

3) Food subreddit that we used as our sub-pormpt had majority of the caption starting with Homemade, OC but still the prompt that was used in CLIP performed much better when compared to the other prompts.

Prompt Ensembling:

We also run prompt ensembling using the best two performing prompts on the the models using **10%** of the test set which are randomly sampled to make it uniform across all the class.
Best two performing prompts: ["OC <cls name>", ""a photo of <cls name>, a type of food "]

Model	Top 1 % Acc	Top 5 % Acc
Model 1	1.054 %	5.84 %
Model 2	3.359 %	9.492 %
Model 3	2.08 %	6.01 %

Table 3: Prompt Ensembling results using best two prompts

Fine-Tuning the model:

We tried to freeze the parameters and finetune the model but since the model was too small, the visual backbone did not learn the visual features well enough. So we fine-tuned the entire visual backbone by adding the linear layers. WE trained this entire model, but due to time and computational constraints we were only able to train the model on 1,200 images which led to the model being extremely underfit. It was unable to learn the features well. We will try to submit a better fine-tuned model trained on the entire training data or at least a larger subset of the training data by tomorrow night even if it's not possible to consider it for grading.

7 Conclusion

We found that increasing the number of transformer layers, made the model train faster (in terms of convergence) i.e., the loss decreased at a much higher rate. We can also conclude that the model trained on 700k images(model 2) learnt more general representations of food which helped it perform better in downstream tasks such as zero-shot classification.

Since the model was very small, we can conclude that it could learn only low level features, so during the process of fine tuning the model , we didn't add linear layers and freeze the visual backbone as done by the authors in the paper, we trained the entire visual backbone along with the added linear layers. If we scale our pre-training VirTex model to add more number of transformer layers and replacing the RegNet-X-800MF with a RegNet-X-1.4GF, the visual backbone would learn more high level features along with low level features which would transfer better to downstream tasks.

For Zeroshot transfer, it was observed that the results varied slightly but not significanlty to the prompt used. The prompt which was used in CLIP [10] yeilded the best results. The prompt ensembling yielded followed the similar conjuction. We believe if we pretrain the model for a longer time then we would be getting more fruitful and definitive results.

One of the technical difficulty which we faced was in deciding which checkpoint to use. A low captioning loss did not directly correspond to a low classification loss. While having a fruitful discussion with the GSI, he did mention that after a certain checkpoint they made the visual backbone solve a small classification task and considered that as a checkpoint loss, we did not have the time and resources to implement this. We wonder if we did that the results would be more fruitful.

8 Future Work

With more computational resources, we can scale the architecture by scaling up the visual backbone and increasing the number of transformer layers, and pre-train a larger model. Once we can do that, we can transfer this model in a better way to downstream tasks. We can increase the number of linear layers during fine-tuning of the visual model.

Alternately, we can also train the model to fine tune on more number of images which would make it perform better on that particular downstream task.

References

[1] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing.

[2] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. UNITER: learning universal image-text representations. *CoRR*, abs/1909.11740, 2019.

[3] K. Desai, G. Kaul, Z. Aysola, and J. Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[5] A. Li, A. Jabri, A. Joulin, and L. van der Maaten. Learning visual n-grams from web data. *CoRR*, abs/1612.09161, 2016.

[6] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *CoRR*, abs/1908.06066, 2019.

[7] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019.

[8] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *CoRR*, abs/2004.06165, 2020.

[9] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[11] M. B. Sariyildiz, J. Perez, and D. Larlus. Learning visual representations with caption annotations. *CoRR*, abs/2008.01392, 2020.

[12] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. VL-BERT: pre-training of generic visual-linguistic representations. *CoRR*, abs/1908.08530, 2019.

[13] H. Tan and M. Bansal. LXMERT: learning cross-modality encoder representations from transformers. *CoRR*, abs/1908.07490, 2019.

[14] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015.

[15] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao. Unified vision-language pre-training for image captioning and VQA. *CoRR*, abs/1909.11059, 2019.