MiniProject 4: Reproducibility in ML

Group 44
Max Zhang et al.

Reproducibility Summary

1 Introduction

We selected 'Beyond Categorical Label Representations for Image Classification' by Boyuan Chen, Yu Li, Sunand Raghupathi, and Hod Lipson as the paper of interest due to its novel approach and high-level concepts towards a topic covered in the course. As seen in lectures and the kaggle competition, image classification is a well-traversed topic in Machine Learning, and since the first instances of well-performing convolution neural network architectures such as AlexNet, deep CNNs have gone on to yield super-human results. Our paper of interest identifies a few concerns in the field of image classification, namely enormous data sets needed to train state-of-the-art CNN's and weakness to adversarial attacks for standard categorical models. Then, the original authors propose a novel approach to solve both pain spots: audio label representations.

In class, we saw when labeling images in datasets for m classes using e.g. ResNet, the standard label consists of a one-hot vector of length m where each value reflects the confidence the model has from its training phase. These researchers introduced and conducted experiments that used high-dimensional and high-entropy audio labels instead of categorical labels to train models. In doing so, the researchers were able to obtain similar or better test results when using vast amount of data. With a limited dataset, audio label representation was able to surpass categorical label representation in performance easily. Furthermore, the claim is verified by using low/high dimensional, low/high entropy combinations of label representations. The robustness of a model with respect to adversarial defence is investigated as well. When an image is adjusted slightly with adversarial perturbations so that it is indistinguishable to the human eye, it is common for a less robust image classifier to fail. The paper shows using audio labels provides the classifier with more ground on which to fend off such attacks.

'Beyond Categorical Label Representations for Image Classification' doesn't necessarily advocate for its results being high-performing or SOTA, rather as in it's namesake, the paper places the experiments as grounds for expanding upon this novel approach to ML research. Not much research has been done on labels themselves, rather, optimization methods, weight initialization, architecture design, and data pre-processing. The researchers use CIFAR-10 and CIFAR-100 as datasets to conduct their experiments. An image encoder is used for the various label representations to generate the input to 3 chosen Neural Nets, ResNet-32, ResNet-110, and VGG19. For high-dimensional label representations, an addition decoder is needed to once again convert the high-dimensional prediction into a category for the purposes of evaluation.

As such, the results of the research indicates that using speech as a signal for supervised learning is worth investigating further due to its data-efficient and attack-adverse properties. The paper hypothesises that the high-dimensionality and high-entropy nature of speech labels gives a strong error signal which allows the models to train efficiently.

2 Scope of reproducibility

The original paper aims to address Data Efficiency and Robustness benefits of supervised training with speech labels. Given the speech label's high-dimensional and high-entropy nature, experiments must be run to isolate both properties so to prove that both are needed to produce the performance seen in audio labels. As such, other label representations are presented: gaussian composition, shuffled-speech, uniformly distributed matrix, BERT, GloVe.

On robustness, the researchers run adversial attacks on the classifer by testing it on a modified image set, perturbed by FGSM and the Iterative Method. With efficiency, the dataset is constricted from a full set to 20, 10, 8, 4, then 1 percent to measure the effects on various label representations.

The claims made in the paper that we will be reproducing and experimenting on:

- Supervised training on CIFER-10 dataset using speech labels yields similar or better results to categorical labels, in contrast to other high-dimensional label representations.

- Supervised training on a limited CIFER-10 dataset using speech labels achieves similar or better performance than using other modalities.

3 Methodology

Using the code repository provided by the researchers, we aimed to re-implement the original method.

`https://www.github.com/BoyuanChen/label_representations#training`

3.1 Model descriptions

ResNet-110 was chosen to train on our dataset. The CNN architecture is an industry standard for image classification due to its skip-connection approach to solving the vanishing gradient problem. Tweaks were made in the dimensions and depth for category or speech labels - these were handled in the provided codebase. Results were taken in training loss, validation loss, and validation accuracy, in which the training, validation, testing split was 45,000/5,000/10,000.

3.2 Datasets

We used the CIFAR-10 dataset to reproduce results from the original experiment. (`www.cs.toronto.edu/~kriz/cifar.html`).

3.3 Experimental setup and code

Using the generated labels, train the model on 100% data, then limit the data to 4% and 1% to investigate the efficiency of other label representations. Using Google Colab, the experiment was conducted using GPU and virtual python environment. A limit of 30 was set on the epochs to train our model in favour of time.

3.4 Computational requirements

When using 100% of data for training, validation, and testing, 30 epochs took just over an hour connecting to the colab GPU. With 4% and 2%, about 10 minutes. In total, using 3 different representations for each percentage, the computing time took around 5 hours.

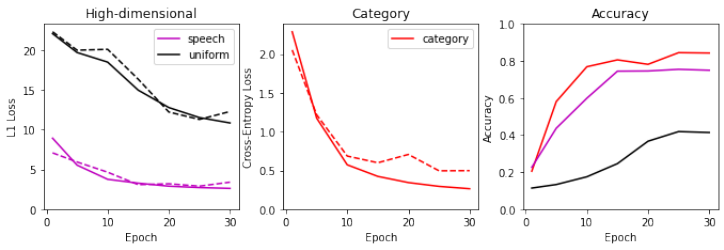
4 Results

The results of our experiments show an agreement with the paper's claims. We were able to accurately recreate the results of the researchers with a shallow training pipeline.

4.1 Results reproducing original paper

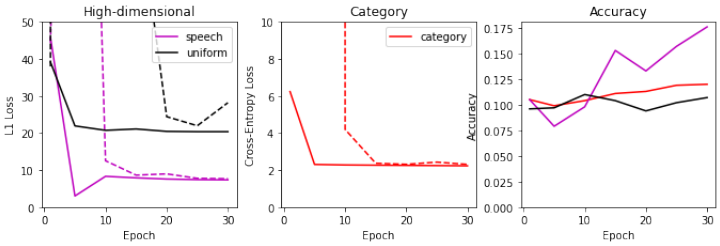
4.1.1 Result 1

In address to Claim 1, the high-dimensionality and high-entropy nature of the speech labels were similar in performance to the categorical label representation when trained on CIFAR-10's 10,000 samples. The fact that the uniform matrix labels performance worse than category and speech confirms that it is not only the high-dimension aspect of speech, but also its high-entropy nature.



4.1.2 Result 2

For Claim 2, when the image dataset was limited to 1%, the performance of the speech labels outdid that of the categorical and uniform labels. As such, the data efficiency property when using audio label representation is reproduced and the claim is verified.



5 Discussion

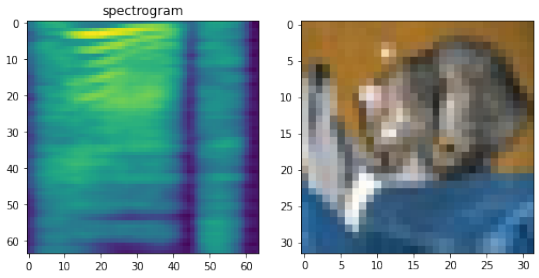
Though the experiment was run on a limited 30 epochs, the visualization and training log results do coincide with a reproduction of the original results. With further time and resources available, we perhaps would have verified the claims with another dataset, such as the MNIST (or the alphabet and number MNIST dataset used in mini project 3) to prove that the intuition behind speech labels vs. categorical labels proves true for other mediums of images. We further encourage to run experiments to verify the validity of the robustness claim in the original paper by running attacks on the chosen image datasets. Takeaways from this experiment in reproducibility are to explore other modalities of labels in image classification, and propose a future experiment to investigate the performances of different languages for generating speech labels.

5.1 What was easy

Even for one with limited experience dealing with scripts and environments, the installation instructions set up in their GitHub repository was very clear and easy. The file provided $train.py$ essentially made training, validation, and testing a one-shot process, with customizable flags for the options of models, datasets, etc. which made testing and reproducing results straightforward.

5.2 What was difficult

In trying to access parts of the pipeline/model, the black box nature of the study and codebase made it difficult to access resources that one desired. In trying to generate predictions, a bit of fiddling was required to access the our trained ResNet model, luckily the code was well-written and well-commented.



Class: cat