International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

Review of Tencent ML-Images Large-Scale Multi-Label Image Database

Luke Breitfeller[1], Sahib Singh[2], Abhinav Reddy Chamakura[3]

[1]Language Technology Institute, Carnegie Mellon University
[2]Heinz College, Carnegie Mellon University
[3]Heinz College, Carnegie Mellon University

Abstract - *The paper examines and re-implement the paper Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning by Wu et al [1]. The paper implements a fine-tuned Resnet visual representation model, trained on a novel 10M image dataset of the authors' invention.*

Key Words: Visual Representation Learning, ResNet, Multi-Label Image Database.

1. INTRODUCTION

The paper is re-implementation of the paper Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning by Wu et al. [1], which seeks to define and test a large-scale image database annotated with multiple labels per image. The database, known as Tencent ML- Images, contains 10M images with 14k possible image labels. The paper tests this database using the ResNet image classifier model, modifying the model to run on a distributed multi-GPU system, getting improved efficiency and accuracy of 79.2 top-1 accuracy when used in transfer learning to single-label image prediction tasks.

1.1 Tencent ML-Images

The Tencent ML-Images databases contains 10M total images. These images were assembled through scraping of existing ImageNet and OpenImage databases. As ImageNet is a single-label database, Wu et al. extrapolated multi-label annotations using entailment within a semantic multi-label hierarchy (where the image label "animal" might be a parent of "dog", and thus all images labeled "dog" must also be labeled "animal") and also predicting labels based on the co-occurrence matrix of multi-labeled OpenImage images.

1.2 Implementation of ResNet

To test the database, Wu et al. train a ResNet visual representation model on their database and test the accuracy for all 14k labels. They cite prior SOTA usages of ResNet [2] which were run sequentially for two months to train over 50 epochs. Wu et al. Instead implement the model distributively on a system of 128 linked GPUs. With this system, they were able to run the training model over 60 epochs for 90 hours.

1.3 Fine-tuned checkpoints

The paper produced five checkpoint weights with their implementation of ResNet. One checkpoint, trained only on ImageNet data, serves as their baseline. The remaining checkpoints are trained on the Tencent ML-Images database and fine-tuned on ImageNet, as previous work has demonstrated this technique improves accuracy (Sun et al.). These checkpoints either used fixed learning rate (checkpoint 2) or adaptive learning rates of different types (3-5).

1.4 Evaluation

We et al. used instance-level metrics to measure the the results of their trained model. This included instance-level precision, recall, and F1. Wu et al. note their results are not particularly high (F1 scores of 23.3 for top-5 prediction and 28.1 for top-10 prediction) and cite missing examples of certain labels in their validation data as one significant gap in their evaluation.

1.5 Paper's impact in the field

We et al. present their work as filling in an existing gap in the field of visual representation. Large- scale, publicly available image databases that had existing prior used single-label annotations; while multi-label image databases did exist before, they were relatively small, limiting their usefulness as training tools for complex machine learning models.

The paper asserts that the reliance on single-label image annotations prevents visual representation models from being able to analyze images with more than one object, or from being able to make inferences about an image's content based on other

factors of the image (for example, one may conclude that if an image contains a doctor, it is highly likely to be located inside a hospital).

An additional contribution of the paper is implementing the commonly used ResNet model in a distributed framework (as discussed in "Implementation of ResNet"), significantly cutting down on the time required to run the model on large-scale databases.

2. RESULTS

2.1 Implementation Challenges

A significant challenge in re-implementing the paper was working around comparatively limited resources. The distributed GPU framework run by Wu et al. required over 1000 GPU-hours to train the data, which we determined would require a prohibitively large budget to run using AWS.

An additional challenge was the size of the database. The full Tencent ML-Images database consists of 10M images. Each image varies in storage size, but many take over a MB of storage space on their own. We did not have the storage space necessary to access the full dataset.

2.2 New Code

Though Wu et al. published the code used to download Tencent ML-Images, establish the RestNet model, and train the data sequentially, they did not publish the parallelized version of the code. In this paper we converted the code to a parallel format utilizing TensorFlow.

We also developed additional tools based on the image download code provided to extract only excerpts of the code.

3. RESULTS

3.1 Training Data

A significant hurdle in training on the Tencent ML-Images dataset was the size of the data. We found that every 500 images from the source ImageNet and OpenImage datasets took a full GB of memory, which prohibited us from loading the full dataset into our implementation of the ResNet training. We ran with a significantly reduced training set, preventing us from matching the F1 scores of the full dataset.

Model Ckpt	Our Top1Train224	Paper Top1-Train224	Our Top-1Val224	Paper Top-1Val224
!Ckpt2 ImageNet	76.7	78.80	92.92	94.50
Cpkt3 ImageNet 224x229	75.61	78.3	90.5	94.3

3.2 Fine Tuning Results

The fine-tuning part includes training of the Model trained previously on Tencent ML-Images database. The checkpoint was provided for public use by the authors. The model is then trained on entire ImageNet with an image size of 224 by 229. Our re-implementation results are listed below.

4. CONCLUSION

4.1 Literature Review

Though the database is quite large, many image categories are barely represented among the data, and the average number of images per label is 1447.2. Given that the paper puts so much emphasis on the size of the dataset, it seems that for any single label a smaller, more specialized dataset would perform the task equally well and at lower time cost.

Another note is that the dataset contains very few human-annotated images, and in fact the process of scraping images from ImageNet includes additional machine annotations that may not be correct, like annotating images with one label with a label that had high co-occurrence in the OpenImages models. A pre-processing step also selects random bounded boxes from the original image for use in training, meaning that certain image labels may be rendering entirely incorrect once they are used for training.

Possibly as a result of the above, the paper presents low scores for multi-label prediction. Though in some respects even those scores are impressive for 11k possible labels, it is a far cry from the informed visual representation that can learn from other labels that the authors originally envisioned for the database.

4.2 Our Work

An unexpected hurdle was the size of the database given the naturally large size of image data. We think, knowing what we know now about working with image files, we would have chosen instead a challenge analyzing text data to lower storage concern.

Another concern was that the paper did not implement anything particularly novel in its model–rather, the model and evaluation were designed to demonstrate the usefulness of the new dataset. As our limited resources prevented us from utilizing the full dataset, this by extension limited our ability to gain useful data from the re-implementation.

REFERENCES

[1] Wu, Baoyuan, et al. "Tencent ML-Images: A large-scale multi-label image database for visual representation learning." arXiv preprint arXiv: 1901. 01703 (2019).

[2] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in ICCV. IEEE, 2017, pp. 843–852.