Week 1 Report

Romouald Dombrovski

August 23, 2024

1. Abstract

Radford, A. *et al*. (2021) "Learning Transferable Visual Models From Natural Language Supervision". In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748-8763). PMLR. https://proceedings.mlr.press/v139/radford21a.html

Abstract: State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study the performance of this approach by benchmarking on over 30 different existing computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet-50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on.

Summary (GPT-4o)**:** The paper *Learning Transferable Visual Models from Natural Language Supervision* introduces CLIP (Contrastive Language–Image Pre-training), a model that leverages large-scale natural language supervision to learn visual concepts that generalize across a wide range of tasks. The authors train CLIP using a dataset of 400 million image-text pairs and show that it can perform zero-shot transfer—where the model is tested on tasks it has never explicitly been trained on—achieving strong performance on standard benchmarks. The key insight is using contrastive learning to align images and corresponding text descriptions in a shared

embedding space, enabling the model to understand visual content in a more general and scalable way. The results demonstrate that combining vision and language during pre-training leads to highly transferable representations across different visual tasks.

2. Scripts and Code Blocks

Not a lot of script changes to report, as this is the first week and I was just familiarizing myself with the code base and setting it up on local dev containers. There were several hiccups with the process, so I began working on improving the initial docker-compose process. I still do not have write access to the NFHM git repository so I wasn't able to push any changes.

This coming week we will have a meeting with Dr. Porto and the collaborators from University of Florida, which will possibly change the scope of the project especially in terms of our hosting infrastructure.

In the meanwhile we will be continuing to familiarize ourselves with the existing pipeline, and checking where improvements can be made. For now, there are improvements to be made within the startup pipelines.

3. Next Week Proposal

Several possible areas of improvement for next week:

- Fix postgres docker image startup.
- Entry scripts for ingestor pipelines.
- Debug embedder pipeline, begin monitoring resource usage during embedding process.
- Attempt to fix multithreaded approach for embedding process.