A Review: Learning Transferable Visual Models From Natural Language Supervision

The paper "Learning Transferable Visual Models From Natural Language Supervision" outlines a novel method for pre-training a model to recognize both images and text simultaneously. The work addresses the issue of developing learning representations capable of capturing the rich and complicated interactions between visual and textual information in order to enhance the accuracy of downstream tasks (i.e. image classification and image retrieval). To solve this issue, the research proposes CLIP, a contrastive learning framework that trains a model to connect pictures and text in a shared embedding space. The paper's key contribution is the invention of a self-supervised learning strategy that allows the model to learn to align picture and text representations without requiring explicit alignment across the modalities.

The model is pre-trained on a vast dataset of picture and text pairings created by the authors known as the WebImageText (WIT) dataset. The abundance of publicly accessible data in this form on the internet serves as a significant driver for natural language supervision. The WIT dataset has over 400 million picture and text pairs for visual model training. The produced dataset has a total word count comparable to the WebText dataset used to train GPT-2. Over-fitting is not a serious worry because the pre-training dataset is so vast. CLIP learns a multi-modal embedding space by training an image encoder and a text encoder simultaneously to maximize the cosine similarity of the image and text embeddings of the batch's N actual pairs while minimizing the cosine similarity of the embeddings of the N^2-N incorrect pairings. During pre-training, natural language is utilized to refer to previously acquired visual ideas (or describe new ones), allowing for zero-shot model transfer to downstream tasks.

Understanding and interpreting the link between pictures and text is a key challenge in computer vision and natural language processing, hence the study is well-motivated. CLIP can increase the accuracy of a wide range of downstream tasks and allow novel applications that need a comprehension of both modalities by simultaneously pre-training a model on both visual and textual input. Since it does not need annotations, learning from natural language is significantly easier to scale than typical crowd-sourced labelling for picture classification. It also has a significant benefit over most unsupervised or

self-supervised learning systems in that it not only "learns" a representation but also relates that representation to language, allowing for flexible zero-shot transfer.

The ability of CLIP to do zero-shot image classification, in which the model can properly identify photos based on textual descriptions without requiring any particular training in those classes, is one of its distinguishing features. CLIP achieves SOTA performance on this, as well as other downstream tasks including object identification and visual question answering. The best CLIP model improves accuracy on ImageNet significantly and further matches the performance of the original ResNet-50. This demonstrates that CLIP is a big step toward adaptable and effective zero-shot computer vision classifiers. CLIP-trained models scale very well, and the largest trained model marginally exceeds the highest-performing existing model on both total score and computation efficiency. The article also shows that the model is resistant to task and distribution alterations, which is significant for real-world applications where data distribution may vary over time.

Although CLIP is typically good at detecting common objects, it struggles with more abstract or systematic tasks like counting the number of objects in a picture and more complicated tasks like determining how close the nearest object is in a picture. CLIP also does not effectively capture the hierarchical structure/semantics expressed in pictures and texts, which is crucial for vision-language comprehension and reasoning. The amount and quality of the pre-training dataset may restrict the model's capacity to catch fine-grained information in pictures and text. Further study is needed to determine how to enhance the model's performance on increasingly difficult visual and language tasks. CLIP's zero-shot classifiers have shown to be sensitive to phrase or phrasing, necessitating trial and error "prompt engineering" to work properly.