Paper review

Paper title: Learning Transferable Visual Models From Natural Language Supervision

Paper link: <https://arxiv.org/pdf/2103.00020.pdf>

Paper 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. We release our code and pre-trained model weights at

Review:

1. Introduction: The paper begins with a brief note on the pretrained NLP models, which are trained on text data, and their advancement over the years. However, they also note the lack of bridge between state-of-the-art text-based models and image-based models. Prior work seems to suggest so. It then gives a brief history of the development of image-text/text-image models over the years. It is a good overview with interesting models worth looking into. It then goes into some limitations on the application of using Natural Language models for image representation. The paper then talks about some potential limitations in the technologies that compete with NLP. The main point of the current paper is stated in the next section – that the “scale” is the factor that differentiates the weakly supervised models and NLP for image generation. The paper claims are summarized in the last section but basically they claim that by simplifying ConVIRT and training it on 400 million image-text pairs is “an efficient method of learning from natural language supervision”. The rest of the section is some stats of how the model performs in comparison with state-of-the-art models.
2. Approach
   1. Natural Language Supervision – here is the core of the project – learning from natural language. The authors argue that this is what sets CLIP apart from other models, since they describe their approaches as unsupervised, weakly supervised etc. Whereas learning directly from natural language is a more scalable approach. It also has the benefit of learning not just the representation but connecting the representation to language.
   2. Creating sufficiently large dataset – The initial section talks about how previously used datasets are limited by several factors including poor labeling, smaller size or lack of natural language description. They attempted to fix this by constructing a new dataset of 400 million image-text pairs collected from variety of publicly available sources on the internet. Important to note that they tried to keep the data balanced and only used words that appeared at least 100 times on Wikipedia.
   3. Selecting an efficient pre-training method – The section start with a brief acknowledgement of the high computational requirements of NLPs and Image models. They also talk about their initial approach and the limitations of it. IMPORTANT: then they tried to redesign the exact nature of prediction models with a relaxed version where they take the text **as a whole** and match it with an image. I didn’t understand the next section but they go into more detail about how exactly they achieved the above goal. The last section go into more detail about the parameter optimizations and different concerns like over-fitting.
   4. Choosing and Scaling a Model – The first section talks about the image encoder and its ResNet-50 architecture. The second section talks about the text encoder and its Transformer architecture. Last section talks about scaling and optimizing the two different architectures.
   5. Training – This section contains the specifics on how the different parts were trained – exact sizes, epochs etc.
3. Experiments
   1. Zero shot Transfer
      1. Motivation – here the researchers define what zero-shot (generalization to unseen data) is and how they use it (as measurement of **task-learning capabilities**). They also talk about the first (and only one they are aware of) zero-shot transfer to standard image classification work – Visual N-Grams. The last section talks about what inspired this approach (how it was used to evaluate GPT-1 which served as the basis for GPT-2)
      2. USING CLIP FOR ZERO-SHOT TRANSFER – In this section the authors describe the way zero-shot was achieved using CLIP. In a nutshell, it embeds both the image and the text through the different encoders and then checks to see how much these categories match.
      3. INITIAL COMPARISON TO VISUAL N-GRAMS – In this section the authors just flex on Visual N-Grams on some common benchmarks like ImageNet and aYahoo. They do note that their model is much larger and requires more computing power to train.
      4. PROMPT ENGINEERING AND ENSEMBLING – Super important section which talks about words with double meaning and how the model cannot interpret in which sense the word is being used – examples given were “boxer”, which can refer to a bread of dog and an athlete. They also talk about how providing a longer and more detailed prompts can increase the accuracy of the model. For example, saying “A photo of a {label}” is a good default option instead of using just the word itself, since it provides more context to the model. Lastly, prompt engineering and ensembling improve accuracy by almost 5%.
      5. ANALYSIS OF ZERO-SHOT CLIP PERFORMANCE – This section talks about specific performance indicators in relation to ResNet-50. They also mention that they achieved state of the art in some of the tasks but the algorithm did poorly on other tasks – like lymph node tumor detection or German traffic sign recognition. The performance of the zero-shot CLIP model is also compared to models that used few-shot learning.
   2. Representation Learning – Here the authors also compare CLIP with a whole suite of other models (66 total) on 27 datasets. The rest of the section is a more detailed flex of how CLIP is the best model out of all.
   3. Robustness to Natural Distribution Shift – In this section the authors start by talking about machine learning performance in relation to human performance and how DL models still make simple mistakes – possibly due to patterns in their training data that do not hold true for larger datasets. In later sections the authors talk about how robust zero-shot models are by nature. The authors also pose some questions regarding a 9.2% increase in accuracy after the model has been trained on some of the images in ImageNet. The posed questions do not have concrete answers but some hypothesis are being mentioned. The gist of this section is that the model should not be trained on a specific dataset, rather for many to be really robust – who would have thought…
4. Comparison to Human Performance – In this section the authors explain an experiment they did for measuring Human Performance on One shot learning task about recognizing dog pictures. The paper notes how humans rapidly improved their accuracy from zero-shot learning to one-shot learning, suggesting that humans “know what they don’t know”. CLIP destroyed humans on this task as evident from their data. Also the paper points out that humans and CLIP both struggled on same data, which could be explained by mislabeling and noise.
5. Data Overlap Analysis – In this section the authors talk about how they try to avoid their validation set to be leaked into the training data via duplicates in the data. They use 3 different ways to account for potential overlap in the data. Their detector has a low false-positive rate. The rest of the section is a humble brag of how little improvement can be made to their awesome detector.
6. Limitations – The authors note that 1000x improvement in compute is needed to achieve state-of-the-art performance for zero-shot. The next section talks about how CLIP is not very good at counting things in an image and other specialized tasks. THE MODEL ALSO SUCKS AT MNIST HAHAHAHAHAHAHAHA WHAT A LOSER – 88% ACCURACY LMAO. The authors also talk about how DL are slower trainers and if each image was presented for a second to the AI it would take 405 years for it to be trained. There is a counterintuitive drop in performance from transitioning from zero-shot learning to few-shot learning.
7. Broader Impacts – This section opens with interesting applications to CLIP. Some of which, like identifying shoplifters could have serious social impact. The authors also talk about other important social aspects of their work and that the impact of CLIP needs to be further studied.
   1. Bias – In this section the authors ran CLIP (two different versions of it) on the FairFace dataset and found interesting misclassifications of some of the images. For example, when they included words like “crime” or “theft”, disproportionate number of people under the age of 20 were classified using these categories, which could signify an inherit bias within the model. Also, similar things were observed with words for animals like “Gorillas” and “orangutan” etc. The authors then use the finding for the people under the age of 20 to further experiment and add another label “child” which reduced the number of harmful misclassifications by the algorithm. This could signify that if the algorithm has a lot of harmful categories to choose from and not enough normal categories, it could be coaxed into having bias. The authors in one of the next sections discover a relationship between occupational roles and gender (females were associated with jobs like “nanny” or “housekeeper” and men with jobs like “mobster” and “prisoner”). Further bias was found when describing appearance – more pictures of females had words like “blonde” or “brown hair” etc. Even articles of clothing were impacted.
   2. Surveillance – The authors preface this section by saying that it is an important section to study in terms of CLIP’s impact, not in terms of promoting CLIP’s integration into it. Overall, the findings suggest that there are better alternatives to CLIP, given that it was originally not designed with surveillance tasks in mind. The authors believe it could potentially be used in very niche applications, given its good zero-short capabilities. Probably the skill requirements for such tasks could be lowered after the introduction of CLIP.
   3. Future work – here is important to note that one of the key areas is investigating bias in the model which is kind of what my project will be about.
8. Related work – Given how broad the definition of NLP is, the authors suggest a lot of literature supporting their research.
9. Conclusion – Reiteration of the connection between NLP and Computer Vision. Overall good conclusion.