Following the ACL rolling review form here is my review for the paper, "Learning Transferable Visual Models From Natural Language Supervision" by Radford et. al.

- **Paper Summary**
  This paper introduces a multi-modal, zero-shot model called CLIP (Contrastive Language-Image Pretraining) that utilizes Natural Language supervision for learning visual representations. In this work authors create a new 400 million dataset and use a simplified version of Contrastive learning training from ConVirt which they suggest is an efficient way for learning from natural language supervision. In the paper Figure 1. illustrates the CLIP approach, on how during training contrastive pre-training is utilized to match <image,text> pairs in embedding space to perform well on downstream tasks. Authors have created a large dataset of 400 million image-text pairs and during contrastive pre-training stage they jointly train image and text encoders to maximize the cosine similarity between image-text pairs in the latent space. They employ symmetric cross-entropy loss for aligning in both directions text->image, and image->text. Authors come up with several neat engineering solutions to scale the overall pipeline (some of it is described in section 2.4, 2.5, 3.1.4 in paper). This paper also highlights a clever use of prompt template (e.g. a photo of a {label}) in helping the zero-shot performance and alleviating issues due to polysemy when matching text-image pairs. Addition to prompt engineering authors show ensembling different prompts (e.g a photo of a big {label}, a photo of a small {label}), over embedding space shows good performance gain over baselines (Figure. 4). The authors evaluate their method on several benchmark datasets and show that their model outperforms existing state-of-the-art methods in zero-shot and few-shot (using linear probe) settings. Even though the model performs well on most of the common benchmarks, it fails on certain some simple classification benchmarks like MNIST, and more details are given in Figure 5. The reasoning behind it is since the model never learnt those concepts during training this behavior is expected. Overall, the paper demonstrates the effectiveness of using natural language supervision to train transferable visual models. Also, authors of this paper have gracefully cited all the relevant prior work on top of which their approach is based on.

- **Summary of Strengths**
  - The CLIP model achieves state-of-the-art performance on a range of visual recognition benchmarks, demonstrating its effectiveness and potential for real-world applications. This performance is achieved by using several techniques like prompt-engineering and ensembling (Figure 4 in paper), simplifying tasks from matching exact words to pairing text to image (Figure 2 in paper). They highlight that ensembling in embedding space (instead of probability space) is helping the model learn better representations.
  - Authors show the effectiveness of their approach by carrying out several ablations across different benchmarking datasets (e.g Figure 5, 6), and across different SoTA models (Figure. 10).
  - In section 3.3 authors does thorough analysis of CLIP model robustness to natural distribution shift in dataset. Figure 13. highlights that even though not ideal it is still better than the existing standard and existing robustness techniques

- The authors have made the code for the CLIP model available on GitHub, along with pre-trained models and tools for fine-tuning and evaluation (except for training pipeline and the dataset used for training). So it is useful for the community to explore their work and benefit from it.

- **Summary of Weaknesses**
  - Authors neither release the dataset nor even comment about the sources from where they get it for training. Just saying downloaded from internet is limited information, and somehow restricts the community to experiment on top of it. However considering its ethical implications it could be a reasonable thing to do.
  - The reasoning provided for low performance on some dataset like MNIST, EuroSAT (referring Figure. 5) even though it appears to be sound, some experiments done to observe the actual representation might have been better I think.
  - Comparing human performance in section 4 of the paper seemed very limited as just 5 humans were involved in comparing it with CLIP model's zero-shot performance

- **Overall Assessment**
  - Accept. I believe research community as a whole will benefit from this work and I can see its potential application for many tasks like image search, encoder in generative models, etc

- **Reviewer confidence**
  - I am confident that I could understand the approach mentioned in the paper, and the method described in paper will be helpful to research community
- **Limitations** (refer section 6. in paper)
  - Authors mention that the comparison they did with the supervised training (using Resnet models) is not always with an existing SoTA model for the downstream task, which requires 1000x compute (described in section 6). They rightly suggest on doing further research for computational and data efficiency.
  - Paper also mentions that model has limited to poor performance on tasks like counting the number of objects in image, and with fine-grained classification like differentiating models of cars, species of flowers, variants of aircrafts.
  - Authors also discusses that CLIP generalizes poorly to truly out-of-distribution data, using example of OCR as an illustration
- **Biases/Ethical Concerns**
  - Since CLIP is trained on data from internet which are unfiltered and uncurated authors are aware that the model will have an implicit social bias present in it. They provide more detailed analysis and results, with some of the mitigation strategy in section 7 of the paper.
- Reproducibility
  - Authors have released the inference code and models (smaller version) for community to try, In a way one can use it to reproduce the results or try incorporating in the downstream task