# Reviewer

Name: Akshay Raman
Net ID: ar8692

# Paper 1

## Title

Learning Transferable Visual Models From Natural Language Supervision

## Rating

*10: strong accept, should be highlighted at the conference*

## Summary

The authors aim to find scalable pre-trained methods for computer vision. To this end, they demonstrate that the task of predicting which caption goes with which image is an efficient pre-training method. Traditional computer vision systems relied on supervised learning approaches trained on limited labelled datasets. In contrast, CLIP takes advantage of large-scale publicly available web data by constructing a custom dataset of over 400M (image, text) pairs. The pre-trained models show impressive zero-shot transfer performance on over 30 computer vision datasets. Moreover, the authors also show that CLIP learns good image representations which can be useful for downstream tasks.

## Strengths

The authors use a simplied version of ConVIRT trained from scratch and successfully demonstrate that methods like contrastive learning and whole caption embedding enable efficient and scalable pre-training. The paper includes a comprehensive benchmark, spanning over 30 datasets, to measure the zero-shot performance of the model. CLIP shows impressive zero-shot performance on a variety of datasets and matches accuracy of previous baselines without additional training (Ex. Resnet-50 on ImageNet)

## Weaknesses

The paper doesn't share the WIT dataset (400M) which was used for pre-training and also fails to provide details about the data curation process. While they do compare the pretraining performance of WIT with an open-source dataset (YFCC100M), the comparison is unfair since the model was only trained on a small subset of 15M images, text pairs. CLIP performance is also poor on fine-grained and OOD tasks, which underscores the importance of training data used. Additionally, the training process is computationally expensive requiring multiple days of GPU training.

## (Optional) Follow-up Questions

I would be interested to learn more about the dataset used for pre-training and the data curation process uses to contruct the WIT dataset.

## (Optional) Follow-up Ideas

A natural extension would be to extend this pre-training approach to other modalities like video, audio. Another research direction would be to develop strong few-shot learning approaches that can utilize CLIP efficiently.

# Paper 2

## Title

Sigmoid Loss for Language Image Pre-Training

## Rating

*8: accept, good paper*

## Summary

The paper aims to improve upon the softmax-based contrastive loss which is common in language image pretraining. They propose a novel sigmoid loss function that is more efficient and scalable than previous methods. This loss function allows scaling up the batch size and and supports efficient distributed training. The results show that SigLIP achieves superior performance over other prior work (like CLIP). The authors also tested the model on large batch sizes (upto 1M), and found that 32k is an optimal batch size for pre-training.

## Strengths

The paper proposes a new sigmoid loss function that performs better than softmax baselines, particularly for small batch sizes. The disentangling of batch from the loss allows efficient loss implementation requiring less memory transfers in distributed training. SigLIP outperformed all prior work in imagenet zero-shot accuracy. Additionally, SigLIP is robust on noise or corruption in the data. The loss function improves efficiency for small batch sizes while also allowing the models to be trained with very large batch sizes (1M).

## Weaknesses

The models are trained using the WebLI dataset which is not publicly available and does not mentions details about the data curation process. The sigmoid loss consists of a learnable bias term which helps to reduce imbalance better positive and negative pairs. It is observed that the performance of the loss function is sensitive to the initialization of the bias term.

## (Optional) Follow-up Questions

Why does the performance drop for very large batch sizes in SigCLIP? The paper does mention that the grad norm values spike as the batch size increases but is still missing an explanation for the spike. Additionally, I would also like to understand why SigLiT doesn't suffer from performance drop and saturates instead.

## (Optional) Follow-up Ideas

A possible research direction would be to explore new loss function that are efficient like sigmoid and can also eliminate the bias terms to stabilize training. Another future work would be to test efficent masking algorithm to fix imbalance of the loss function.