Are MLPs all you need for computer vision?

Final Project for Intro to Computational Intelligence in Fall 2021

Alex Morehead

November 20, 2021

Contents

1 Abstract

Convolutional Neural Networks (CNNs) are to date a standard class of deep learning algorithms for learning from image-like data [1]. Recently, alternative types of neural networks such as Attentive CNNs [2], Vision Transformers [3], and MLP-Mixers [4] have been proposed for tasks in computer vision. Interestingly, MLP-Mixers have been succinctly designed to simultaneously incorporate the advantages of multilayer perceptrons (MLPs), CNNs, and Transformers (i.e., per-location features and patch-based spatial information) for image-based deep learning. However, there are currently few explorations in the literature on comparing these new MLP vision models to their standard MLP and shared weight neural network (SWNN) counterparts. In this work, I propose to investigate the effect of the location and patch-based feature mixing techniques that MLP-Mixers apply to see their effect on the model's image classification performance. My study will make use of benchmark computer vision datasets such as the MNIST and ImageNet datasets to allow meaningful comparisons of the classification results presented by the MLP-Mixer with those offered by standard MLPs and SWNNs.

2 Introduction

2.1 Motivation

Over the last few years, a litany of new neural network architectures have recently been proposed for computer vision tasks. Lagging behind these new models, however, are thorough case studies examining their relationship to historically-grounded neural networks such as MLPs and CNNs. Conducting more of such case studies could lead to an enhanced understanding of the true amount of progress that has been made in image-based deep learning.

3 Conclusion

Through this project, we will see that MLP Mixers achieve impressive results for image classification by distilling the advantages of contemporary computer vision learning algorithms such as CNNs and Transformers into a single simple network.

4 References

References

[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee, 2017.

[2] Michael Neumann and Ngoc Thang Vu. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*, 2017.

[3] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

[4] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.