Название статьи: MLP-Mixer: An all-MLP Architecture for Vision
Автор исследования: Иван Сафонов, БПМИ 182

1. Работа написана весной 2021 года.

   Опубликована на arXiv 4 мая 2021 (первая версия), 10 июня 2021 (четвертая версия); на OpenReview 21 мая 2021 года.

   Была постером на NeurIPS 2021.

2. Авторы статьи: Google Research, Brain Team

   Более конкретно (жирным выделены основные авторы):

   - **Ilya Tolstikhin**: CV, GAN.
     Most cited: «Wasserstein auto-encoders»
   - **Neil Houlsby**: CV, NLP, Bayesian methods.
     Most cited: «An image is worth 16x16 words: Transformers for image recognition at scale»
   - **Alexander Kolesnikov**: CV.
     Most cited: «An image is worth 16x16 words: Transformers for image recognition at scale»
   - **Lucas Beyer**: RL, CV.
     Most cited: «In Defense of the Triplet Loss for Person Re-Identification», «Revisiting Self-Supervised Visual Representation Learning»
   - Xiaohua Zhai: RL, CV.
     Most cited: «An image is worth 16x16 words: Transformers for image recognition at scale», «Revisiting Self-Supervised Visual Representation Learning»
   - Thomas Unterthiner: CV, GAN, Bioinformatics.
     Most cited: «Fast and accurate deep network learning by exponential linear units (ELUs)», «GANs trained by a two time-scale update rule converge to a local nash equilibrium», «Self-normalizing neural networks»
   - Jessica Yung: CV (transfer learning, representation learning)
     Most cited: «Big Transfer (BiT): General Visual Representation Learning»
   - Andreas Peter Steiner: CV, Bioinformatics.
     Most cited: «MLP-Mixer: An all-MLP Architecture for Vision»
   - Daniel Keysers: CV
     Выпускал работы начиная с 2001, много популярных старых работ.
   - Jakob Uszkoreit: CV, NLP
     Most cited: соавтор «Attention is all you need»
   - Mario Lucic: CV, GAN
     Most cited: «Are GANs Created Equal? A Large-Scale Study»
   - Alexey Dosovitskiy: CV
     Один из основных авторов Visual Transformer (статья «An image is worth 16x16 words: Transformers for image recognition at scale»)

3. Всего работа ссылается на 60 статей.

   Наибольшее влияние, как мне кажется, оказала работа «An image is worth 16x16 words: Transformers for image recognition at scale», в которой описывается архитектура Visual Transformer. Множества авторов очень сильно пересекаются (работа также сделана в Google Research, Brain Team), архитектуры ViT и MLP-Mixer различаются только тем, что посередине модели вместо N блоков трансформера применяется N блоков Mixer Layer.

   Остальные ссылки не являются такими интересными в контексте работы.

4. Всего у статьи пока что 114 цитирований.

   Интересные статьи, ссылающиеся на нашу:

- «Do Vision Transformers See Like Convolutional Neural Networks?» by Google Research, Brain Team: сравнение Visual Transformer и CNN, того как они решают задачу классификации, представлений на промежуточных слоях. Рассуждение об архитектурах и связях с MLP-Mixer
- «Pay Attention to MLPs» by Google Research, Brain Team: также исследуется тема замены трансформеров на MLP, на этот раз сравнивают не только в CV, но и в NLP с BERT
- «A Battle of Network Structures: An Empirical Study of CNN, Transformer, and MLP»: сравнение трех архитектур, рассуждение о гибридных архитектурах

Часто упоминают MLP Mixer в статьях про Visual Transformer.

5. Прямых конкурентов у статьи не было.

6. Возможная идея для исследования: объединение CNN, ViT, MLP-Mixer слоев в одной модели.

   - Давайте вместо обучаемой линейной проекции патча $B \times B$ изображения использовать CNN, в конце которой получится вектор из C каналов, который уже будет передаваться в MLP Layers. Интересно посмотреть на разные значения B. Также возможно это поможет легко решить проблему применимости модели к картинкам разного размера (потому что сначала применяется CNN, которая может быть применена к любым размерам).
   - Что будет, если в модели использовать и слои транформера и Mixer Layers? Несколько возможных вариаций: по очереди в некотором порядке, либо каждый слой это сумма слоя транформера и Mixer Layer.
   - Размещать CNN блоки можно не только в начале, также можно: после любого слоя модели представить, что у нас изображение размера $\frac{H}{B} \times \frac{W}{B}$ с C каналами, применять CNN слой.

   Интересно провести подобное Architecture Study, понять поможет ли какое-нибудь такое объединение улучшить точность.

   Еще одна возможная идея: применение MLP-Mixer к задаче сегментации. Можно получившиеся в конце S векторов размера C также обучаемой линейной проекцией перевести в блоки размера $B \times B$ и соединить в картинку исходного размера.

7. Авторы показывают, что MLP-Mixer обучается эффективнее на TPU, чем другие модели.

   Возможно, у нее также более эффективный inference. Возможно можно немного пожертвовав в качестве ускорить время применения в задачах классификации на CPU/мобильных устройствах.