MLP-Mixer: An all-MLP Architecture for Vision
(Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer,
Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers,
Jakob Uszkoreit, Mario Lucic, Alexey Dosovitskiy)

Автор исследования: Алексей Цеховой.

1. Работа свежая, опубликована на arXiv'e в мае этого года, были правки в июне, ещё не была представлена на конференциях.
2. Основной вклад имеют I. Tolstikhin, N. Houlsby, A. Kolesnikov и L. Beyer - активные исследователи Google Brain. N. Houlsby, A. Kolesnikov и L. Beyer состоят в числе авторов статьи, представившей в прошлом году Vision Transformer [1] с текущим вторым местом на ImageNet. Там они отказались от свёрток, показав state of the art результаты с трансформером, исследуемая же статья продолжает идею и предлагает отказаться и от них.
3. Предыдущая работа [1], получив отличные результаты и широкую гласность (1062 цитирований на момент написания), очевидно, является главным входновлением. Авторы сами упоминают публикации по сверточным сетям [2] и трансформерам [3] как источник некоторых своих решений.
4. Статья свежая, цитирования имеются, но их важность пока неизвестна.
5. Касательно конкуренции, заявлено, что модель является частным случаем архитектуры Synthesizer [4], представленной годом ранее, где так же рассматривался отказ от механизма внимания. Также можно проследить аналогию смешивания каналов и патчей с обычными свертками при определённых параметрах (1x1 и большой шаг соответственно).
6. Была изучена лишь классификация изображений. Интересным является модифицирование для возможного применения в других задачах зрения. Так, например, нет очевидных решений для передачи видеоряда. Так как обучение и тесты проводились лишь большими моделями на огромных комбинированных наборах данных, неизвестна применимость в меньших задачах.
7. Показано, что архитекрута имеет отличный компромисс между качеством и вычислительной сложностью, что важно для любого потенциального массового продукта.

Список источников:
[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby - "An image is worth 16x16 words: Transformers for image recognition at scale", in ICLR, 2021.
[2] A. Krizhevsky, I. Sutskever, G. E. Hinton. - "ImageNet classification with deep convolutionalneural networks", in NeurIPS, 2012.
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin - "Attention is all you need", in NeurIPS, 2017.
[4] Yi Tay, D. Bahri, D. Metzler, D. Juan, Z. Zhao, C. Zheng. - "Synthesizer: Rethinking self-attention for transformer models", ICML 2021.