AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Константин Чернис, группа M05-014

1. Введение

При решении почти любой задачи глубокого обучения, как в компьютерном зрении, так и в обработке естественного языка, для получения "удобного"для решения задачи представления входных данных применяется предобученный энкодер, который файнтюнится под текущую задачу. Долгое время стандартным решением в компьютерном зрении были архитектуры на основе свёрток (Kolesnikov et al., 2020 [5]), в то время как в обработке естественного языка уже долгое время доминирует архитектура Трансформер (Vaswani et al., 2017 [7]), которую отчасти можно назвать более перспективной в связи с меньшим inductive bias: вместо того, чтобы учить паттерны изображений, она позволяет обращать внимание на любую его часть при принятии решения, что в теории может позволить выучить более сложные закономерности.

Тем не менее, несмотря на перспективность архитектуры Трансформер для компьютерного зрения, на пути его успешного применения стояла следующая проблема: традиционным способом адаптации Трансформера к картинкам является вытягивание картинок в последовательность пикселей, что с учётом квадратичной сложности вычисления внимания относительно длины входа не позволяло использовать для обучения картинки достаточно высокого разрешения. В более ранних работах данную проблему пытались решить за счёт локального внимания (Palmar et al., 2018 [6]) и разреженных аппроксимаций (Child et al., 2019 [2]), но тут возникают проблемы с производительностью из-за неадаптированности графических ускорителей к подобным нестандартным архитектурам.

От себя добавлю, что недавно в обработке естественного языка появились модели с линейным вниманием (Wang et al., 2020 [8], Zaheer et al., 2021 [9]), но, как было показано в статье, это не лучше предложенного подхода, ибо возникают сложности аналогичные побуквенным языковым моделям в обработке естественного языка.

2. Метод

Для решения описанных выше проблем в статье предлагается разбить картинки на кусочки (patches), например, размера 16×16, что соответствует токенам в обработке естественного языка, после чего размер последовательности сокращается с HW до HW/P^2, что решает проблему длинных последовательностей. Данная идея позволяет авторам не отходить от стандартной архитектуры Трансформера (хоть они и

применили некоторые улучшения, появившиеся в последующие годы), что на правах господствующей архитектуры для обработки естественного языка открывает доступ к оптимизированному вычислению на графических ускорителях.

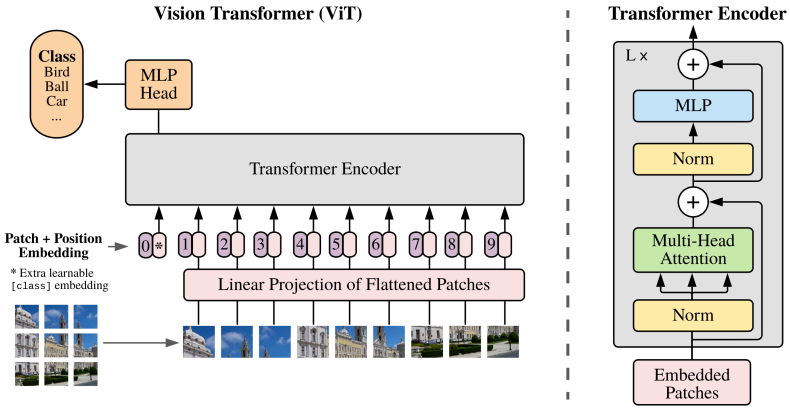

Рис. 1: Архитектура Трансформер

С точки зрения применения архитектуры, то, как видно из Рис. 1, авторы явно вдохновлялись подходом, предложенным в BERT (Devlin et al., 2018 [3]), самым известном энкодером для текста на архитектуре Трансформер. В частности, здесь применён токен для классификации в начале последовательности, из которого в дальнейшем и извлекается информация о классе.

Авторы приложили много усилий по искоренению inductive bias, чтобы позволить модели самой решать, что ей требуется для получения наилучшего результата, в частности, единственным местом, где он вообще возникает, являются позиционные эмбеддинги: на практике часто бывает полезно файнтюнить модели на большем разрешении, чем при предобучении, в таком случае при сохранении размера патча длина последовательности возрастает, то есть появляются позиции, для которых нет обученных эмбеддингов, в этом случае, аналогично обработке естественного языка, эмбеддинги экстраполируются до нужной длины.

3. Эксперименты

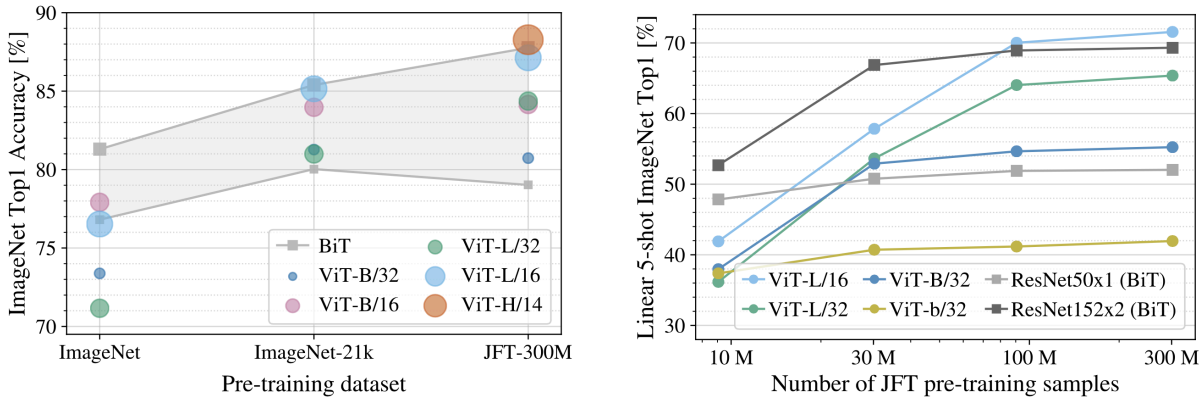

Рис. 2: Зависимость качества от количества данных

Главным численным результатом работы является доказанное превосходство Трансформера над свёртками на больших объёмах данных. Как видно из правого графика (Рис. 2), на датасетах стандартного размера (размер ImageNet 14 млн объектов) предложенная модель проигрывает традиционным ResNet на основе свёрток, но уже на 30 млн данных новая архитектура вырывается вперёд. Здесь суффиксы B и L соответствуют Base и Large версиям Visual Transformer (эти понятия аналогичны таковым в BERT), число обозначает размер кусочка.

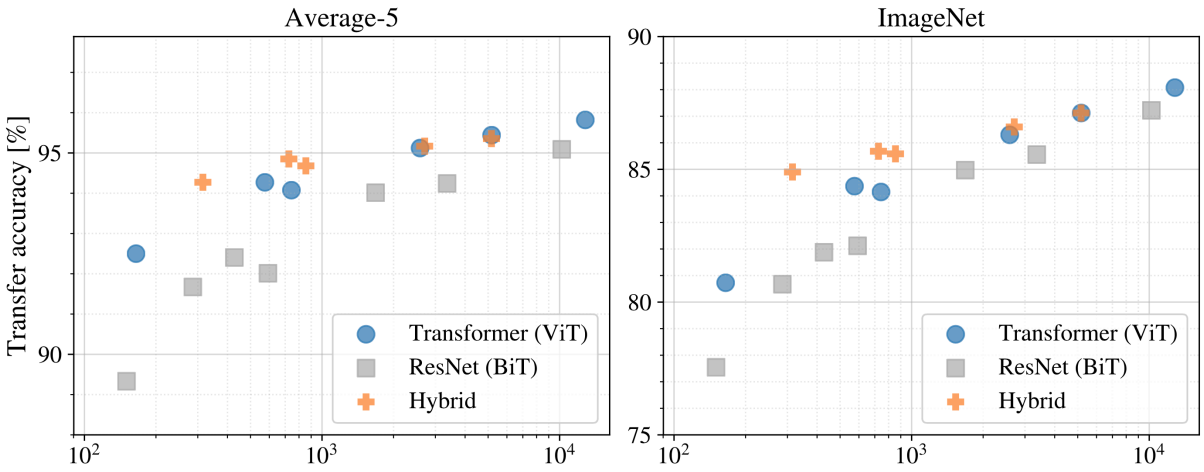

Рис. 3: Затраты ресурсов

Также интересно посмотреть на затраты ресурсов при обучении, из Рис. 3 видно, что новая архитектура требует в 2-4 раза меньше мощностей на получение того же качества, что и свёртки. Кроме того, можно отметить, что на начальных этапах модель выигрывает от inductive bias свёрток (что видно по результатам гибридных архитектур), но с ростом количества данных эта разница невелируется.

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Рис. 4: Результаты бенчмарков

Наконец, по результатам бенчмарков (Рис. 4) предложенная модель является state-of-the-art в классификации изображений, что особенно впечатляет, учитывая возможные в будущем улучшения за счёт ещё большего количества данных, чем не могут похвастать существующие свёрточные архитектуры.

4. Заключение

Оригинальная статья (Dosovitskiy et al. 2020, [4]) делает большой шаг вперёд в области компьютерного зрения, как давая новый виток развитию архитектур, так и предоставляя целую плеяду возможных подходов по их улучшению из соседней области. Кроме того, унификация архитектур может быть полезна при их дальнейшем использовании в продакшене, а также в смежных областях, таких как одновременная генерация текста и картинок (OpenAI, 2021 [1]).

Список литературы

[1] Dall·e, 2021.

[2] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[5] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2020.

[6] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[8] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[9] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2021.