Комбинирование сверточных слоев и трансформеров в генеративно-состязательных сетях

Демин Дмитрий Андреевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: dmitriy.demin.98@mail.ru

Научный руководитель *— Назаров Леонид Владимирович*

Одной из популярных на сегодняшний день архитектур нейронных сетей в задачах компьютерного зрения являются визуальные трансформеры [1]. Существуют работы, посвященные созданию генеративно-состязательных сетей с помощью такой архитектуры, например, [2]. Сильным местом данной архитектуры является возможность видеть зависимость между удаленными пикселями изображения.

Одним из ключевых слабых мест трансформеров является их высокая требовательность к ресурсам. Например, авторы [2] использовали для своих исследований 16 GPU.

Другая популярная архитектура генеративно-состязательных сетей основана на свертках, например, [3]. Данные элементы не позволяют видеть зависимости между удаленными пикселями, но занимают меньше места в памяти и требуют меньше времени для обучения.

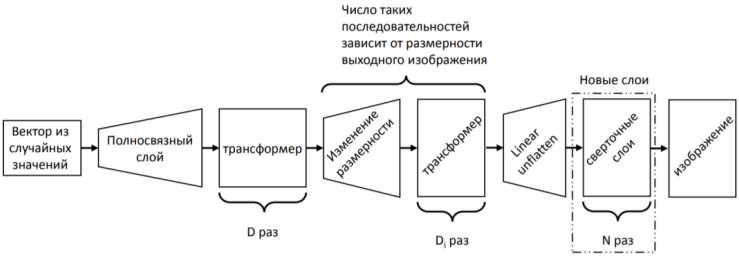

Схема сети-генератора с предлагаемой архитектурой.

Была проведена попытка совместить 2 данных подхода. Предлагается для первых слоев генератора и последних дискриминатора использовать трансформеры, а для более крупных слоев - свертки. В случае данной работы слои до размерности 16×16 сети реализо-

ваны с помощью трансформеров, а переход от 16×16 к 32×32 и наоборот - с помощью сверточных слоев.

После обучения сетей с 3 упомянутыми архитектурами для изображений одинаковой размерности 32×32 из одного набора данных (предподготовленные изображения собак из stanford dogs dataset) и с одинаковой размерностью входного вектора равной 256 были получены следующие результаты в плане качества изображений (FID) и потребляемых вычислительных ресурсов:

Архитектура	Размер генератора	Время обучения, ч	FID
DCGAN	58,1 мб	около 16	135.17
TRANSGAN	32,1 мб	около 54	98.64
смешанная	28,8 мб	около 24	123.02

Как видно, смешанная архитектура является сопоставимой с трансформерами с точки зрения занимаемой памяти и компромиссной с точки зрения времени обучения сети. Это важно в случае обучения сети, созданной для задач генерации изображений большой размерности или при наличиии ограниченных вычислительных ресурсов.

Литература

1. Dosovitskiy A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv 2020, arXiv:2010.11929
2. Yifan Jiang, Shiyu Chang, Zhangyang Wang TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up, arXiv 2021, arXiv:2102.07074
3. Radford A., Metz L., Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015, arXiv:1511.06434