Анализ статьи “Learning Transferable Visual Models From Natural Language Supervision”

Во-первых, у данной статьи почти 7.500 цитирований, так что так что на неё нельзя было не обратить внимание. Авторы предлагают довольно таки легковесную модель (всего 63M параметров) для широкого спектра задач CV.

Пройдемся по пунктам эксперимента:

- **Окружение.**
  Библиотеки
  Тут всё понятно и прозрачно - есть github репозиторий. В requirements.txt всего 5 библиотек - помимо обычных torch/torchvision только tqdm, ftfy и regex.
  Если библиотеки поставятся некорректно: без tqdm`a можно и обойтись, а вот ftfy и regex нужны для корректного распознавания описаний, так что без них моделька сломается.
  Ни о каких фреймворках для управления конфигурацией в репозитории речь не идет - так что поставить ту же Hydra не помешает.
  Железо
  CLIP позиционируется как предобученная модель, которая не требовательна к объему данных на target-задаче. Однако если данных много и хочется обучать модель на них всех, то процесс обучения может затянуться: авторы указывают, что обучали CLIP на датасете из 400M объектов *в течение 2х недель, используя 256 GPU V100*.
  Данные
  Сылка на пост - здесь хорошо описаны проблемные места CLIPa
  У модели явно есть проблемы с некоторыми задачами, если на MNISTe даже линрег работает лучше
- **Воспроизведение результатов**
  Статья просто изобилует экспериментами, наборами данных и моделями. Жаль, что без собственного дата-центра половину экспериментов необходимо будет урезать в количестве датасетов и тестируемых моделях.
  Из одного из обзоров: “The largest ResNet model, RN50x64, took 18 days to train on 592 V100 GPUs while the largest Vision Transformer took 12 days on 256 V100 GPUs.” И это в самом интересном эксперименте, - в статье это фигуры 10 и 12, - сравнение CLIPa с другими State-of-the-Art-моделями CV.
  Я уже работал с BYOL, MoCo и SimCLRv2 - эти модели также позиционируются как прорывные в CV. В качестве альтернативы эксперименту из статьи можно провести аналогичный, но не на 12/27 датасетах как авторы - у каждой из модели выше есть данные по времени обучения и требуемым вычислительным ресурсам. Плюсом как минимум все модели тестируются на CIFAR100. Так что в целом можно почти везде воткнуть альтернативные эксперименты с меньшими вычислительными затратами.
- **Сферы применения**
  CLIP будет полезен дизайнерам, маркетологам, SMM-специалистам благодаря гибкости выполнения задач по созданию визуального контента.
  Например, CLIP будет очень хорош для подбора разнообразного контента в соцсетях за счет этих преимуществ.