Рецензия DINO

Константин Матвеев 181

В статье “Emerging Properties in Self-Supervised Vision Transformers” авторы предлагают фреймворк self-supervised обучения DINO (self-distillation with no labels). Они применяют его к Vision Transformer на картинках и к другим моделям и доказывают (качественно и количественно) превосходство этого метода и получаемых моделей над существующими self-supervised и иногда supervised моделями сразу в нескольких задачах. Метод заключается в обучении системы ученик-учитель, при этом ученик и учитель одной архитектуры и изначально рандомно инициализированы.

В процессе обучения каждую эпоху учитель фиксируется и в результате обновляется с помощью экспоненциального скользящего среднего лучшим учеником на эпохе. Во избежание коллапса, авторы статьи применяют к выходу учителя центрирование (своего рода нормировку) и “острый” софтмакс, чтобы балансировать между коллапсом в одинаковый вывод и в равномерное распределение.

Актуальность и новизну сложно переоценить: не сколько из-за того, что статья вышла в этом году, столько из-за результатов, которые достигаются авторами статьи. Они предлагают подход на основе современных BYOL и ViT, который побеждает self-supervised и supervised подходы на разных задачах, не требуя при этом сложных адаптаций под конкретную задачу.

Из сильных сторон, следует выделить результаты, которые достигает DINO на ViT, в особенности на классификации ImageNet, где при лучших результатах у обученной модели в 10 раз меньше параметров, чем у self-supervised SOTA. Также выдающиеся результаты получены на image retreival, copy detection, даже video instance segmentation и других задачах. В самой статье приводятся полноценные эксперименты по разным задачам, большой ablation анализ влияния разных аспектов подхода, что является существенной сильной стороной статьи. Выбор бейслайнов аргументирован. Авторы предлагают как количественную оценку метриками, так и качественную, через визуализацию различных датасетов. Таким образом, полученные результаты и их анализ — преимущество статьи.

Из слабых сторон статьи можно выделить неполноценное описание подхода. В частности, нет описания инициализации ученика и учителя, что, как мне кажется, представляет особый интерес в задаче: во-первых, есть крайние конфигурации, при которых метод не заведётся (константная инициализация), во-вторых, поскольку концептуально модели учатся улучшать изначальный “шум” от рандомных моделей, инициализация должна играть весомую роль. Вместе с этим в статье не только не приводится сравнение способов инициализации, но она вообще никак не упоминается.

Также неясное объяснение приводится в аугментациях, которые, как известно, могут сильно влиять на качество self-supervised. В частности, авторам следует более детально описать выбор размеров глобальных и локальных кропов (сейчас в статье приводятся только разделение на >50% и <50% площади, что слишком общо). Мини-эксперимент в аппендиксе статьи не считаю достаточным. Из-за этого описание моделей слабо из-за отсутствия конкретики по аугментациям, размерам кропов, инициализации.

Теоретическую обоснованность нельзя назвать сильной, поскольку авторы не приводят математических доказательств корректности обучения. Это компенсируется большим количеством экспериментов и наглядных демонстраций работы. Помимо этого, есть ряд стилистических недочётов: неясная формулировка выбора “N = 16 (“/16”) or N = 8 (“/8”)”. Из-за этого в статье случаются проблемы с доходчивостью.

Воспроизводимость не идеальна из-за перечисленных недостатков. При этом авторы приводят репозиторий со своим кодом, в т. ч. экспериментов, и предобученные модели.

В целом, недостатки в описании и теоретической обоснованности работы метода перебиваются выдающимися результатами, которые приносят статье оценку 9.

Оценка: 9
Уверенность: 3
Презентация:
https://docs.google.com/presentation/d/1PNrjqDRZP4fF5NMGfHyZc0mb2T9rDKY9h1bGgVD0d7o/edit?usp=sharing