Название статьи (авторы статьи):
Are Large-scale Datasets Necessary for Self-Supervised Pre-training? (Alaaeldin El-Nouby et. al)
Автор рецензии: Каратаева Екатерина

Краткое содержание и вклад авторов:
В данной статье авторы изучили методы self-supervised pretraining для 2D CV на примере denoising autoencoders. В ходе своего исследования авторы пришли к выводу, что этот тип моделей более устойчив к типу и размеру данных для предобучения, что позволяет предобучаться на меньших объемах данных. Также авторы предложили свой метод для self-supervised pretraining - SplitMask.

Сильные стороны:

1) Статья в целом написана доходчиво, простым языком. Серьезных затруднений при прочтении не вызывает.
2) Авторы провели достаточно полный эмпирический анализ self-supervised pretraining, сравнивая различные модели на датасетах разного типа и размера. Есть ablation study предложенного авторами метода. Также отмечу, что теоретического обоснования утверждений практически в статье не приводится, то есть эта статья больше про грамотные/разноплановые эксперименты.
3) Значимость этой статьи, по моему мнению, заключается в том, что авторы развенчивают миф, что нужно предобучаться на больших наборах данных, и чем больше набор данных, тем лучше. Авторы показывают, что для denoising autoencoders в self-supervised pretraining можно потратить меньше ресурсов для предобучения и получить качество не хуже, а в некоторых случаях лучше.

Слабые стороны:

1) На данный момент отсутствует репозиторий с кодом, но так как статья достаточно свежая (ей 1,5 месяца), то это можно простить, потому что нередко авторы выкладывают код позже статьи.
2) Упущены детали в статье для воспроизводимости такой же модели. Приведенное в статье описание предложенных методов и процедуры обучения достаточно для общего понимания происходящего, но этого недостаточно для воспроизведения такой же модели самостоятельно, опираясь лишь на статью.

Рецензий на эту статью не нашла.

Оценка: 8
Уверенность: 3

Я поставила такую оценку, потому что авторы статьи предложили новый метод, который показывает очень хорошие результаты, провели качественный эмпирический анализ на множестве различных датасетов и сделали много интересных выводов в ходе своего исследования, но я не могу оценить, насколько новы полученные знания, к тому же я плохо знакома с related works, поэтому я уверена в своей оценке на 3.