Название статьи: Efficient Visual Pretraining with Contrastive Detection

Авторы статьи: Olivier J. Henaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, Joao Carreira

Автор текста: Чураков Игорь

Первый препринт статьи появился 23 марта 2021 года. Осенью этого же года статью приняли на ICCV2021. Статья написана шестью исследователями из DeepMind, это британская компания, занимающаяся искусственным интеллектом, в 2014 году она была приобретена Google.

Статья довольно новая, у нее всего 20 цитирований, из любопытных можно отметить, что она упоминается в другой статье [1], соавторами который являются Olivier J. Henaff и Aaron van den Oord, в ней предлагается комбинировать существующие методы self-supervised обучения, k-means и дистилляцию. Ее препринт также появился весной и ее также приняли на ICCV2021.

Оба этих автора в 2020-м году работали над Data-Efficient Image Recognition with Contrastive Predictive Coding [2], в ней также исследуются способы self-supervised обучения для задач компьютерного зрения, основная идея статьи заключается в том, чтобы выделять векторы признаков из небольших патчей изображений а потом контрастировать их. Также проводится большое исследование pretraining efficiency, то есть дообучение предобученной сети на малой доле размеченной выборки. Возможно идея с выделением признаков для масок, которая собственно используется для обучения в методе DetCon это развитие идей именно этой статьи.

Из связанных работ стоит отметить SimCLR [3] и BYOL [4], кроме того, что они решают ту же задачу, авторы DetCon используют их идеи для выбора архитектур сети и пайплайнов аугментаций.

Сами авторы в статье отмечают что их подход очень похож на Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals [5] и Self-Supervised Visual Representation Learning from Hierarchical Grouping [6]. Обе эти статьи также используют сегментацию изображений для self-supervised learning, отличия заключаются в том, что в них учится backbone предназначенный именно для решения задач сегментации и авторы этих статей не проводят эксперименты для измерения pretraining efficiency.

По поводу того что можно улучшить: для работы алгоритма необходимо разбиение изображения на маски, авторы предлагают несколько способов это делать (различные эвристики, ручная разметка). Логичным продолжением кажется попытаться сделать весь алгоритм end-to-end, то есть предложить какую-либо дифференцируемую сегментацию изображений.

Статьи:

[1] Tian, Y., Henaff, O. J., & Oord, A. V. D. (2021). Divide and Contrast: Self-supervised Learning from Uncurated Data. *arXiv preprint arXiv:2105.08054*.

[2] Henaff O. Data-efficient image recognition with contrastive predictive coding //International Conference on Machine Learning. – PMLR, 2020. – С. 4182-4192.

[3] Chen T. et al. A simple framework for contrastive learning of visual representations //International conference on machine learning. – PMLR, 2020. – С. 1597-1607.

[4] Grill J. B. et al. Bootstrap your own latent: A new approach to self-supervised learning //arXiv preprint arXiv:2006.07733. – 2020.

[5] Van Gansbeke W. et al. Unsupervised semantic segmentation by contrasting object mask proposals //arXiv preprint arXiv:2102.06191. – 2021.

[6] Zhang X., Maire M. Self-supervised visual representation learning from hierarchical grouping //arXiv preprint arXiv:2012.03044. – 2020.