Исследование контекста статьи

"Are Large-scale Datasets Necessary for Self-Supervised Pre-training?"

Автор исследования: Болотин Арсений

Основные сведения.

Статья выложена на arxiv 20.12.2021 (arXiv preprint). Пока что присутствует только одна версия.

Авторы статьи:

- **Alaaeldin El-Nouby*:** Facebook AI Research and INRIA[1]
  Most cited: LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference
- **Gautier Izacard*:** Facebook AI Research and INRIA[1]
  Most cited: Leveraging passage retrieval with generative models for open domain question answering
- **Hugo Touvron:** Facebook AI Research, Sorbonne University
  Most cited: Training data-efficient image transformers & distillation through attention
- **Ivan Laptev:** Research director, INRIA[1]
  Most cited: On space-time interest points
- **Hervé Jégou:** Facebook AI Research
  Most cited: Aggregating local descriptors into a compact image representation
- **Edouard Grave:** Facebook AI Research
  Most cited: Enriching word vectors with subword information

У Alaaeldin El-Nouby есть несколько работ по Vision Transformer:

- LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference
- Training vision transformers for image retrieval
- XCiT: Cross-Covariance Image Transformers

Каких-то статьей по self-supervised pre-training у авторов нет.

---

[1] INRIA - Institut national de recherche en informatique et en automatique

* equal contribution

Всего в статье приведено 70 источников.

Основные источники:

- An image is worth 16x16 words: Transformers for image recognition at scale

  Помимо того, что методы основаны на ViT, в статье про ViT предобучали модель на большом наборе данных. В исследуемой статье утверждается, что это необязательно.

- BEiT: BERT Pre-Training of Image Transformers

  Self-supervised pre-training для ViT. Denoising autoencoders. Предложенный метод SplitMask во многом основан на BEiT и с ним сравнивается.
- Emerging Properties in Self-Supervised Vision Transformers
- An Empirical Study of Training Self-Supervised Vision Transformers
  Moco V3, DINO - методы, приведённые в сравнении. Изучается self-supervised ViT training, стабильность обучения, self-distillation.

Цитирования и продолжения.
У статьи нет цитирований на данный момент, так как статья только вышла и ещё будет дорабатываться.

Другие работы.

How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers

(Google Research, Brain Team, 18.06.2021)
Одним из основных результатов статьи является то, что supervised pre-training на значительно меньших наборах данных с использованием продуманных аугментаций и регуляризаций можно добиться качества, которое сопоставимо с обучением на больших наборах данных (в 10 - 25 раз больше). Странно, что статья не упомянута в исследуемой. Более того, в этой статье предлагают сравнить supervised pre-training и self-supervised pre-training

Дальнейшие возможные исследования.

Стоит добавить сравнение с моделями, отличными от Vision Transformer, а также расширить эксперименты другими задачами.

Применение и практическое значение.

Полученные результаты в статье говорят о том, что можно предобучить denoising autoencoders на меньшем наборе данных без потери качества. Это позволяет уменьшить требуемые ресурсы.

Также удалось получить улучшение в качестве, предобучаясь сразу на наборе данных для которого решается задача без использования больших наборов данных.