Efficient Visual Pretraining with Contrastive Detection

Ваньков Тимур
Цыганов Артем
Булатова Екатерина
Першин Максим

План презентации

1. Докладчик
   1.1. Проблематика и связанные работы
   1.2. Система контрастного обнаружения
   1.3. Unsupervised mask generation
   1.4. Эксперименты и сравнения
2. Рецензент
   2.1. Плюсы статьи
   2.2. Минусы статьи
3. Практик-исследователь
   3.1. Публикация
   3.2. Авторы
   3.3. Связанные статьи
   3.4. Цитаты
   3.5. Исследования
4. Хакер
   4.1. Собрать colab-ноутбук с инференсом pretrained resnet-50, resnet-200 DetCon моделей.
   4.2. Пример transfer learning classification на cifar/mnist/fashoinmnist с помощью pretrained DetCon модели.

Докладчик

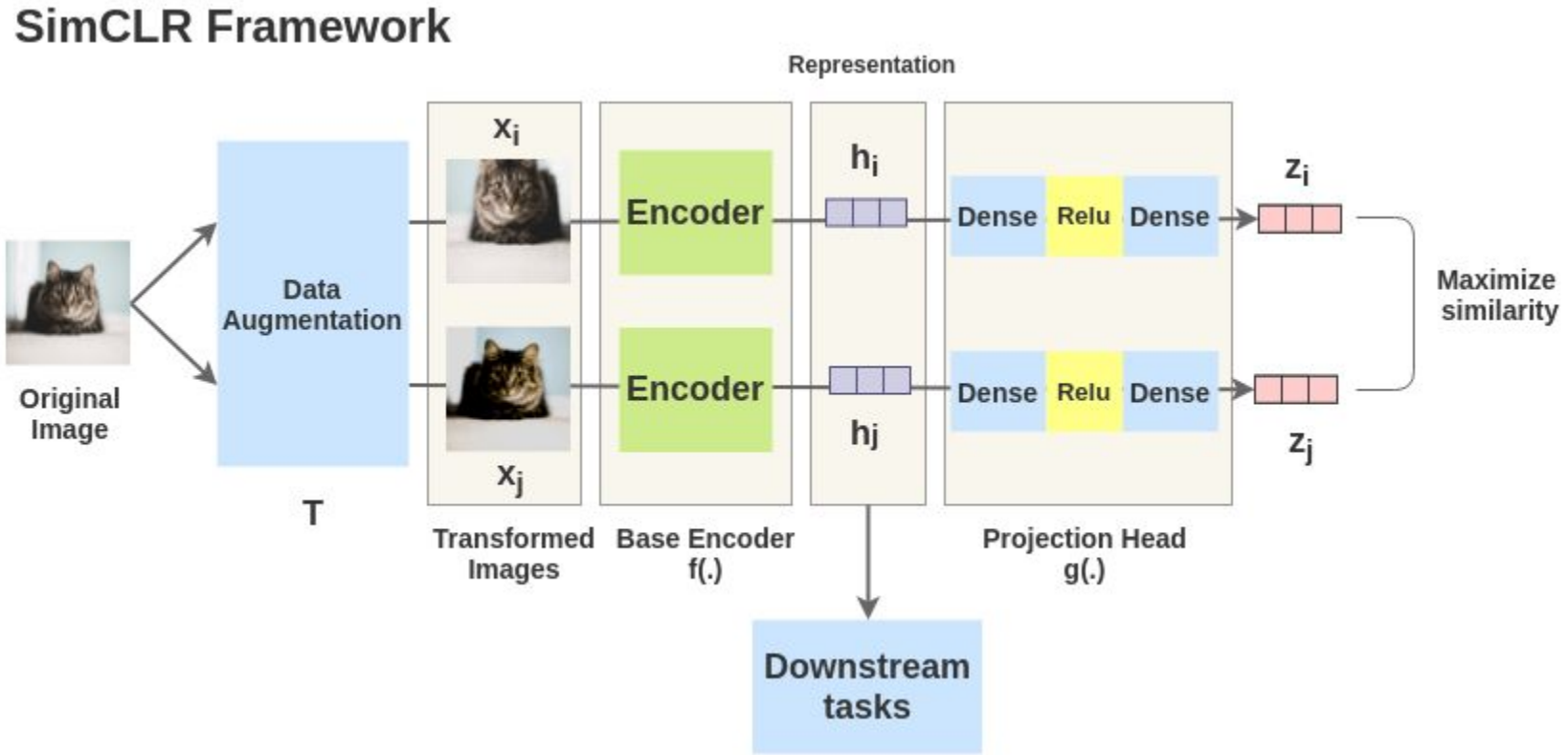
Проблематика и связанные работы

SimCLR

BYOL

DetCon

Проблематика и связанные работы



Проблематика и связанные работы

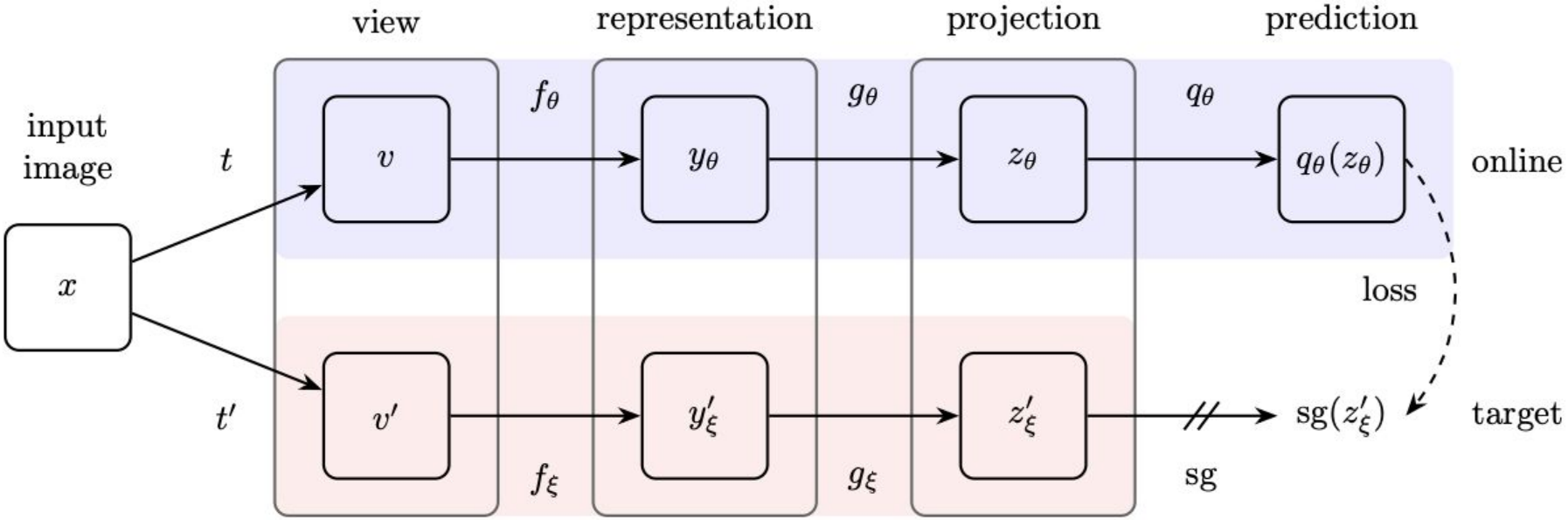


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation.

Система контрастного обнаружения

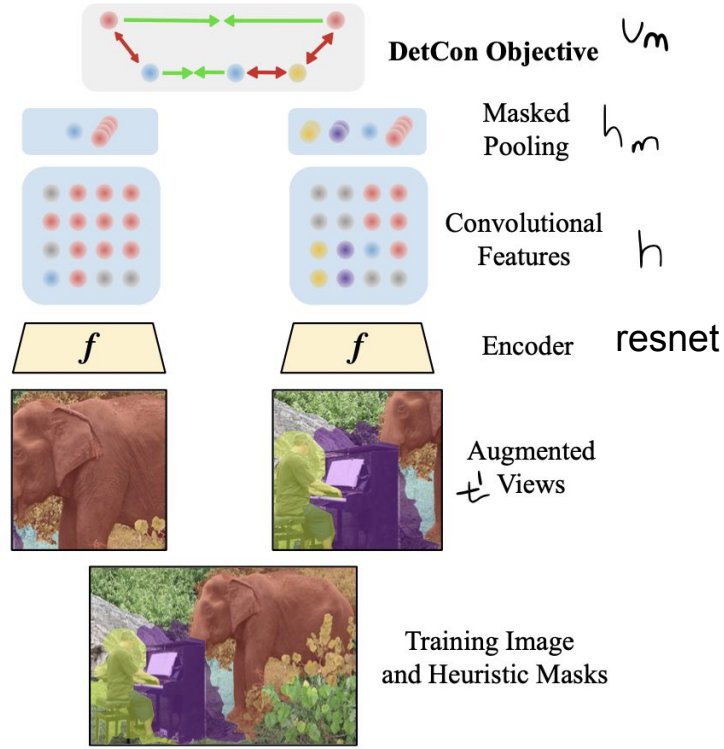
$$\boldsymbol{h_m} = \frac{1}{\sum_{i,j} m_{i,j}} \sum_{i,j} m_{i,j}\, \boldsymbol{h}[i,j],$$

$$\boldsymbol{v_m} = g_\theta(\boldsymbol{h_m}), \quad \boldsymbol{v'_{m'}} = g_\theta(\boldsymbol{h'_{m'}})$$

$$\boldsymbol{v_m} = q_\theta \circ g_\theta(\boldsymbol{h_m}), \quad \boldsymbol{v'_{m'}} = g_\xi(\boldsymbol{h'_{m'}})$$

$$\ell_{m,m'} = -\log \frac{\exp(\boldsymbol{v_m} \cdot \boldsymbol{v'_{m'}})}{\exp(\boldsymbol{v_m} \cdot \boldsymbol{v'_{m'}}) + \sum_n \exp(\boldsymbol{v_m} \cdot \boldsymbol{v_n})}$$

$$\mathcal{L} = \sum_m \sum_{m'} \mathbb{1}_{m,m'} \ell_{m,m'}$$



Unsupervised mask generation

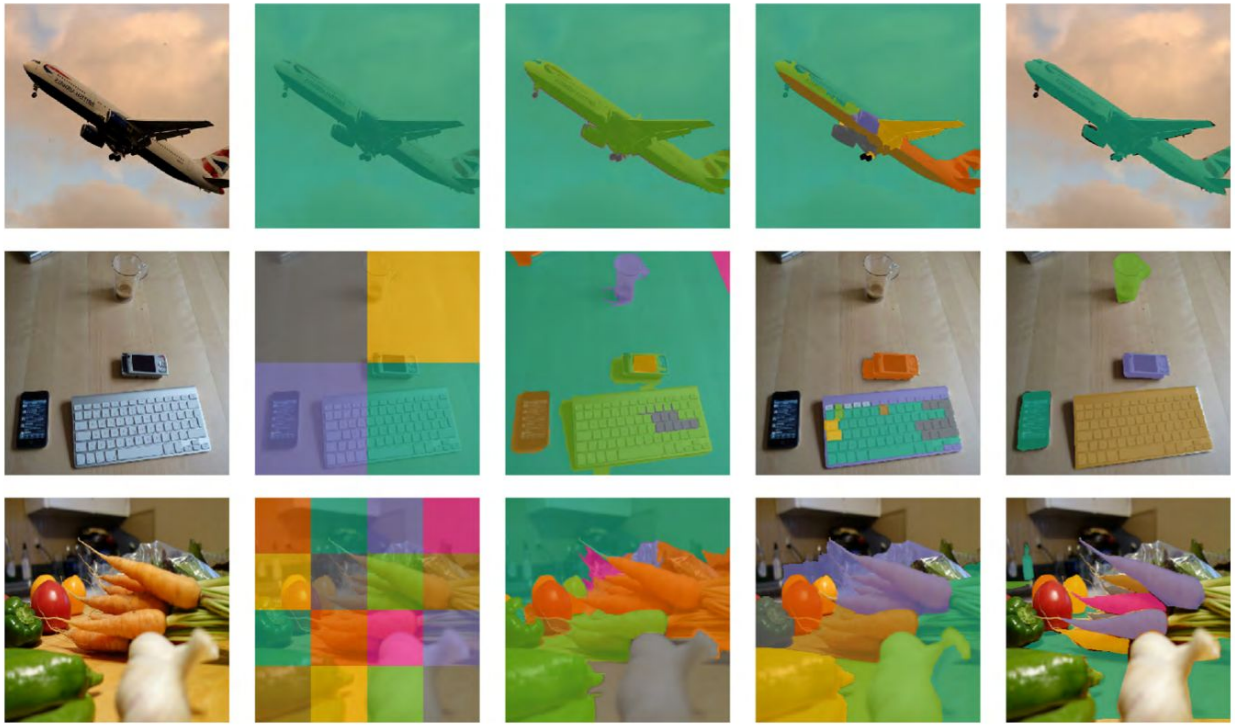


Figure 3. **Example masks used by the DetCon model.** 1st column: random images from the COCO training set. 2nd column: masks based on spatial proximity only. Global masks (top) are implicitly used by methods such as SimCLR, MoCo, and BYOL. 3rd column: image-computable masks obtained from the Felzenszwalb-Huttenlocher (FH, [17]) algorithm, with $s = 500$. 4th column: image-computable masks inferred using Multiscale Combinatorial Grouping (MCG) [2]. 5th column: "oracle" masks used to assess potential improvements from higher-quality segmentations.

Докладчик

Эксперименты и сравнения

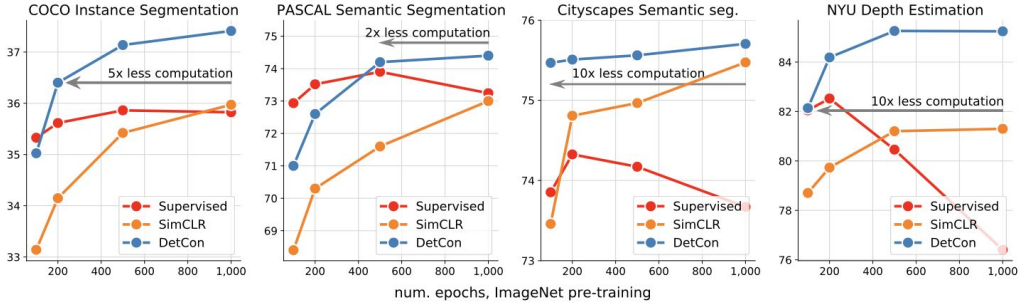
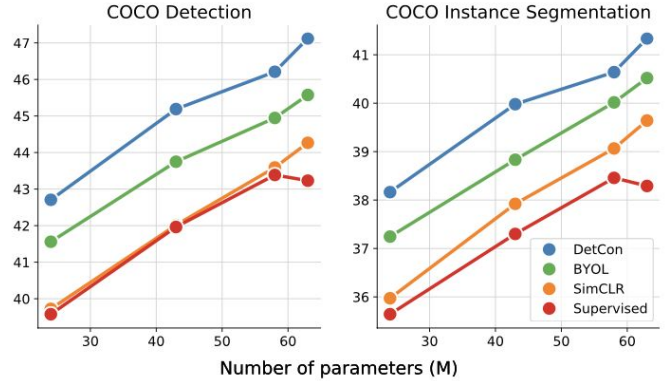


Figure 4. **Efficient ImageNet pretraining with DetCon$_S$.** We pretrain networks with SimCLR, DetCon$_S$, or supervised learning on ImageNet for different numbers of epochs, and fine-tune them for COCO detection and instance segmentation (for 12 epochs), semantic segmentation on PASCAL or Cityscapes, or depth estimation on NYU v2. DetCon$_S$ outperforms SimCLR, with up to 10× less pretraining.

pretrain	Data	Params	AP^{bb}	AP^{mk}
Supervised [20]	IN-1M	250 M	45.9	41.0
SEER [20]	**IG-1B**	**693 M**	48.5	**43.2**
DetCon$_B$	IN-1M	250 M	**48.9**	43.0



method	Fine-tune 1× AP^{bb}	Fine-tune 1× AP^{mk}	Fine-tune 2× AP^{bb}	Fine-tune 2× AP^{mk}
Supervised	42.0	37.3	43.4	38.4
SimCLR [9]	42.0	37.9	43.8	39.3
InfoMin [54]	42.9	38.6	44.5	39.9
BYOL [21]	43.7	38.8	44.3	39.4
DetCon$_B$	**45.2**	**40.0**	**45.7**	**40.4**

(a) **ResNet-101** feature extractor

method	Fine-tune 1× AP^{bb}	Fine-tune 1× AP^{mk}	Fine-tune 2× AP^{bb}	Fine-tune 2× AP^{mk}
Supervised	43.4	38.5	43.4	38.5
SimCLR [9]	43.6	39.1	44.9	40.0
BYOL [21]	44.9	40.0	45.7	40.6
DetCon$_B$	**46.0**	**40.6**	**46.4**	**40.7**

(b) **ResNet-152** feature extractor

method	Fine-tune 1× AP^{bb}	Fine-tune 1× AP^{mk}	Fine-tune 2× AP^{bb}	Fine-tune 2× AP^{mk}
Supervised	43.2	38.3	43.5	38.5
SimCLR [9]	44.3	39.6	45.3	40.3
BYOL [21]	45.6	40.5	45.9	40.5
DetCon$_B$	**47.1**	**41.3**	**47.2**	**41.5**

(c) **ResNet-200** feature extractor

Рецензент

Плюсы статьи

- Большое количество экспериментов
- Сравнения со state of the art подходами
- Все исследования сопровождаются графиками
- Имеется ссылка на код на языке Jax

Минусы статьи

- Скудное теоретическое обоснование
- Тяжела для чтения и понимания

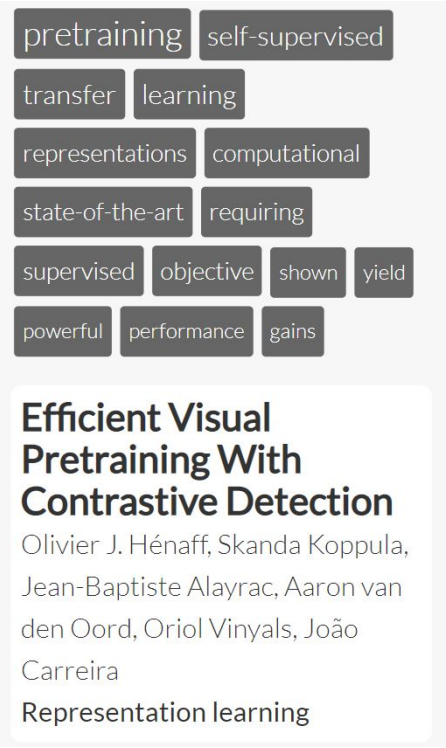
Оценка 7

Уверенность 3

Практик-исследователь

Публикация

- Была опубликована на arxiv 19 марта 2021
- Принята на ICCV21 oral, второй день, 13 октября в 9 утра. Вся конференция проходила онлайн
- Не получила наград



Авторы

Olivier J. Henaff

- Data-efficient image recognition with *contrastive* predictive coding
- 4 статьи за 2021, 2 из них про комбинацию contrastive learning и self-supervised методов

Skanda Koppula

- 3 статьи в 2021

Jean-Baptiste Alayrac

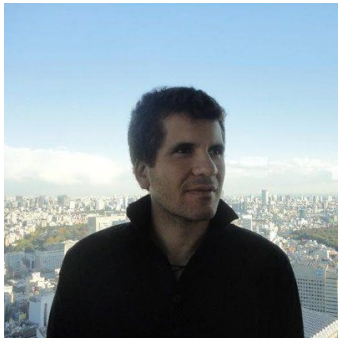
- Self-Supervised MultiModal Versatile Networks
- 9 статей в 2021, много про трансформеры и аудио

Авторы

Aaron van den Oord

- *Representation learning* with *contrastive* predictive coding
- Data-efficient image recognition with *contrastive* predictive coding
- 18 статей за 2021 год, много про self-supervision, генерацию, обработку последовательностей

Oriol Vinyals



- 17 статей за 2021, в основном языковые модели и биология

Joao Carreira

- CPMC: Automatic object segmentation using constrained parametric min-cuts
- 7 статей за 2021, в основном видео и аудио

Связанные статьи

- Базовые:
  - SimCLR и BYOL, из них берут архитектуры
- С похожими идеями:
  - “Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals” (Wouter Van Gansbeke et al): ставится self-supervised задача с масками, полученными unsupervised методами. Акцент не на pretraining, специальные архитектуры для задач COCO
  - “Self-Supervised Visual Representation Learning from Hierarchical Grouping” (Xiao Zhang, Michael Maire): учимся накладывать маски, их используем в self-supervised задаче
- Предшественник:
  - “Data-Efficient Image Recognition with Contrastive Predictive Coding” (Olivier J. Hénaff et al): пересечение в 2 автора: как уменьшить количество размеченных данных

Цитаты

- Ссылаются в основном как на pixe-level метод получения представлений с помощью contrastive loss
- Статьи, занимающиеся похожими темами (“конкуренты”):
  - “Object-Aware Cropping for Self-Supervised Learning” (Shlok Mishra et al): contrastive representations, улучшение cropping для обучения на менее систематизированных наборах данных
  - “DETReg: Unsupervised Pretraining with Region Priors for Object Detection” (Amir Bar et al): предобучают специальную сеть на обнаружение объектов

Исследования

- Уже проведено обширное исследование
- Можно попробовать заменить ResNet на другую архитектуру
- Об отсутствии теоретического обоснования: постараться подтвердить предположение о том, почему метод сработал так хорошо
- (предложено авторами) Насколько хорош наш метод как решение задачи instance segmentation, можно ли использовать его как unsupervised segmentation вместо того, что использовали мы?