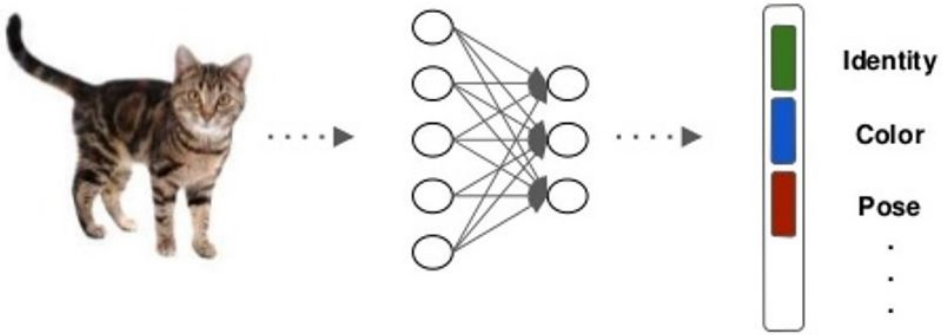
Bootstrap Your Own Latent

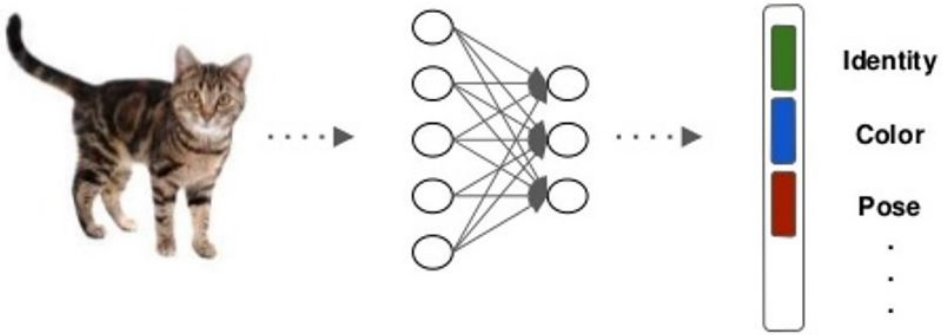
Станкевич Матвей
Чураков Игорь
Сухоросов Алексей
Колесников Георгий

Self-supervised representation learning



- Хотим получить векторные представления картинок

Self-supervised representation learning



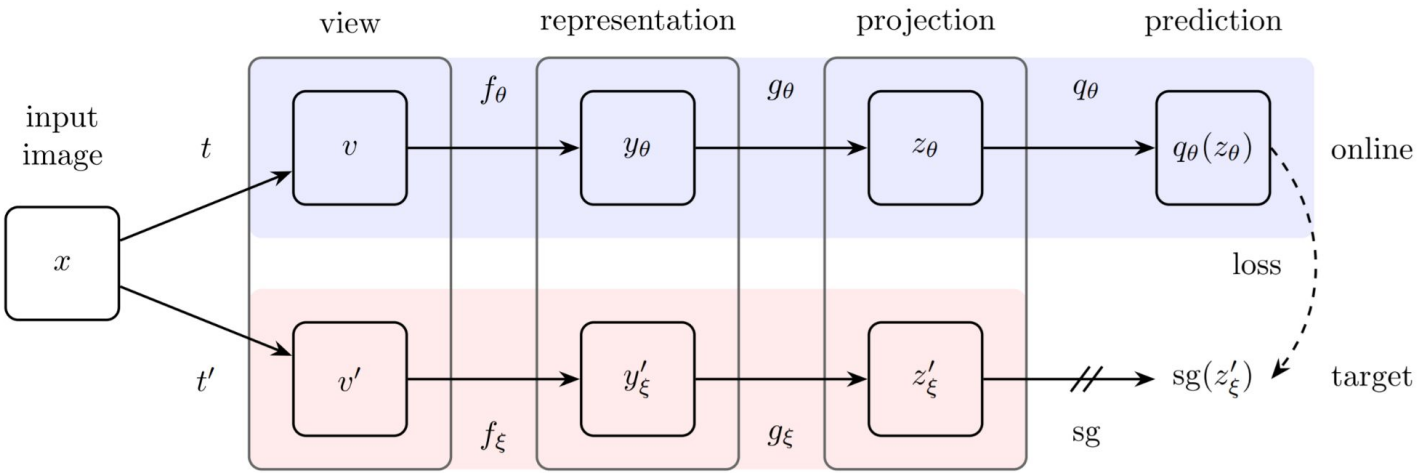
- Хотим получить векторные представления картинок
- Так как данные не размечены или размеченных данных мало, не можем обучиться на какую-то задачу (например, классификацию) и использовать внутренние представления

Существующие методы

Большинство существующих методов обучения представлений можно разделить на 2 группы - генеративные и дискриминативные:

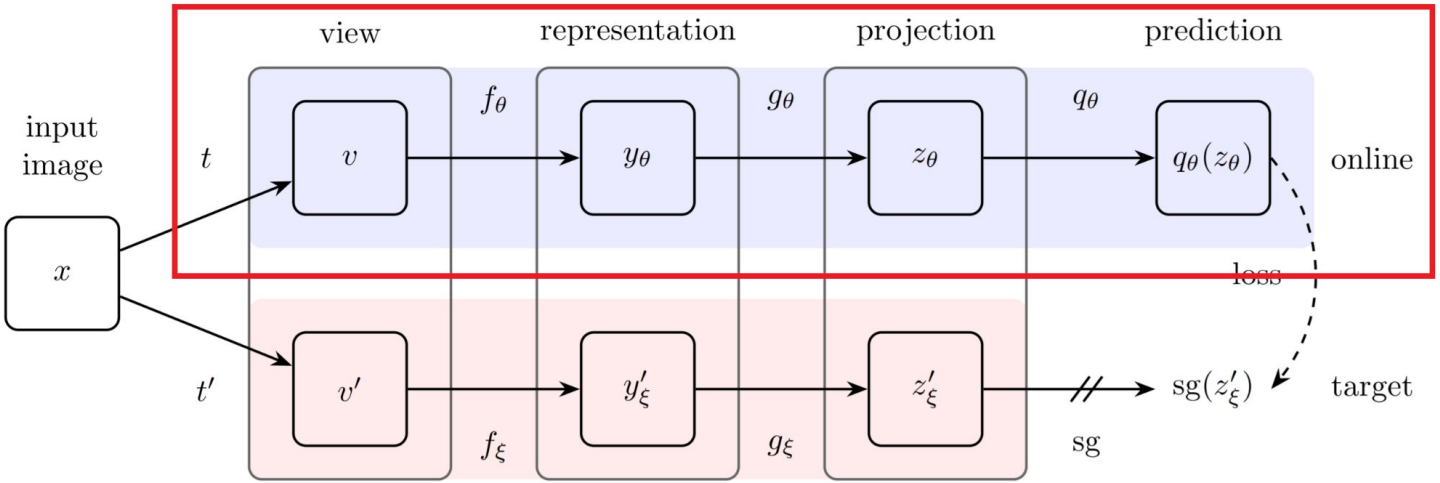
- Генеративные модели пытаются выучить распределение на данных и их представлениях
- Среди дискриминативных методов наиболее популярны контрастивные методы, которые пытаются выучить такие представления, что представления аугментаций одной картинки будут близки, а разных - нет. Примеры: SimCLR, MoCo

BYOL: архитектура



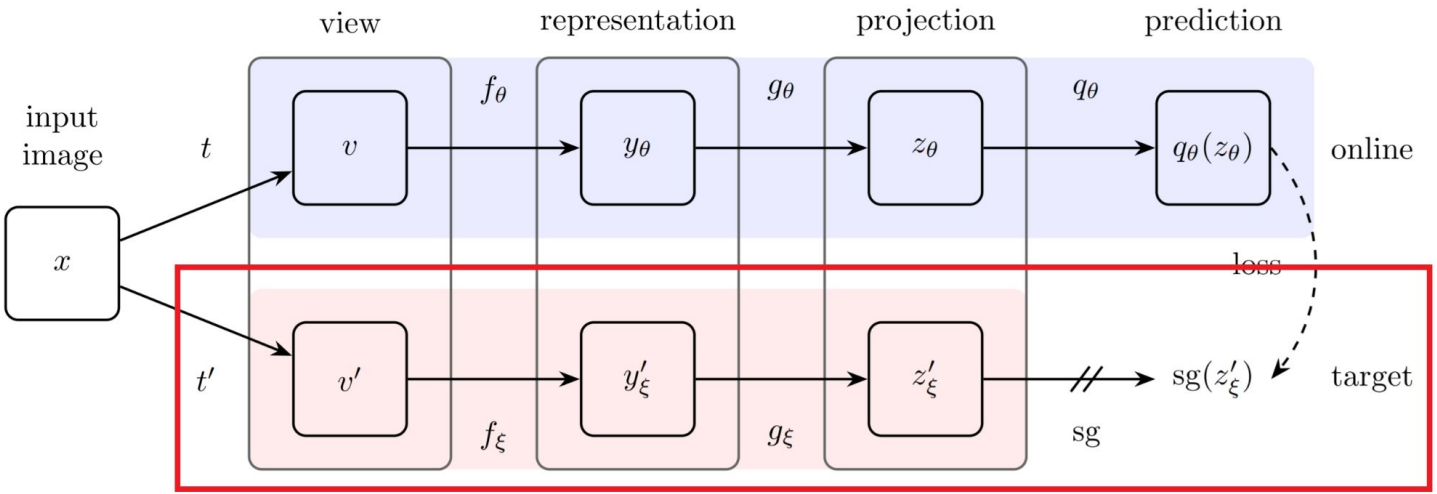
В BYOL используются 2 сети — **online** и **target**. Суть метода в том, чтобы online сеть по аугментации исходного изображения v обучалась предсказывать представление другой аугментации - v’.

BYOL: архитектура



Online сеть состоит из **энкодера f_θ**, **проектора g_θ** и **предиктора q_θ.** На вход энкодеру подаётся t(x) — аугментированное исходное изображение, t ~ T

BYOL: архитектура



Target сеть имеет ту же архитектуру, что и online сеть, но без предиктора.
Аугментация t’ выбирается из своего множества аугментаций T’.

BYOL: функция потерь

$$\mathcal{L}_{\theta,\xi} \triangleq \left\| \overline{q_\theta}(z_\theta) - \overline{z}'_\xi \right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\left\| q_\theta(z_\theta) \right\|_2 \cdot \left\| z'_\xi \right\|_2}.$$

$$\overline{q_\theta}(z_\theta) \triangleq q_\theta(z_\theta) / \left\| q_\theta(z_\theta) \right\|_2$$

$$\overline{z}'_\xi \triangleq z'_\xi / \left\| z'_\xi \right\|_2$$

Итоговый лосс:

$$\mathcal{L}_{\theta,\xi}^{\texttt{BYOL}} = \mathcal{L}_{\theta,\xi} + \widetilde{\mathcal{L}}_{\theta,\xi}$$

BYOL: обучение

$$\theta \leftarrow \text{optimizer}\left(\theta, \nabla_\theta \mathcal{L}^{\texttt{BYOL}}_{\theta,\xi}, \eta\right),$$
$$\xi \leftarrow \tau\xi + (1-\tau)\theta,$$

где θ - веса online сети, ξ - веса target сети, η - learning rate.

После обучения для получения представлений оставляется только энкодер f_θ

BYOL: имплементация

Используемые аугментации - такие же, как в SimCLR:

- random crop
- random flip
- color distortion
- grayscale
- gaussian blur
- solarization: $x \mapsto x \cdot \mathbf{1}_{\{x<0.5\}} + (1-x) \cdot \mathbf{1}_{\{x \geq 0.5\}}$ - только для target сети

BYOL: имплементация

Используемые архитектуры:

- f_θ и f_ξ - ResNet (ResNet-50, ResNet-101, ResNet-152, ResNet-200)
- g_θ, g_ξ и q_θ - MLP

Используемый оптимизатор:

- LARS - Layer-wise Adaptive Rate Scaling - адаптивный метод оптимизации для работы с большими батчами

BYOL: результаты

Method	Top-1	Top-5
Local Agg.	60.2	-
PIRL [35]	63.6	-
CPC v2 [32]	63.8	85.3
CMC [11]	66.2	87.0
SimCLR [8]	69.3	89.0
MoCo v2 [37]	71.1	-
InfoMin Aug. [12]	73.0	91.1
BYOL (ours)	**74.3**	**91.6**

(a) ResNet-50 encoder.

Method	Architecture	Param.	Top-1	Top-5
SimCLR [8]	ResNet-50 (2×)	94M	74.2	92.0
CMC [11]	ResNet-50 (2×)	94M	70.6	89.7
BYOL (ours)	ResNet-50 (2×)	94M	77.4	93.6
CPC v2 [32]	ResNet-161	305M	71.5	90.1
MoCo [9]	ResNet-50 (4×)	375M	68.6	-
SimCLR [8]	ResNet-50 (4×)	375M	76.5	93.2
BYOL (ours)	ResNet-50 (4×)	375M	78.6	94.2
BYOL (ours)	ResNet-200 (2×)	250M	**79.6**	**94.8**

(b) Other ResNet encoder architectures.

Table 1: Top-1 and top-5 accuracies (in %) under linear evaluation on ImageNet.

Для оценки self-supervised представлений обучим линейный классификатор поверх фиксированных представлений

BYOL: результаты

Method	Top-1		Top-5	
	1%	10%	1%	10%
Supervised [77]	25.4	56.4	48.4	80.4
InstDisc	-	-	39.2	77.4
PIRL [35]	-	-	57.2	83.8
SimCLR [8]	48.3	65.6	75.5	87.8
BYOL (ours)	**53.2**	**68.8**	**78.4**	**89.0**

(a) ResNet-50 encoder.

Method	Architecture	Param.	Top-1		Top-5	
			1%	10%	1%	10%
CPC v2 [32]	ResNet-161	305M	-	-	77.9	91.2
SimCLR [8]	ResNet-50 (2×)	94M	58.5	71.7	83.0	91.2
BYOL (ours)	ResNet-50 (2×)	94M	62.2	73.5	84.1	91.7
SimCLR [8]	ResNet-50 (4×)	375M	63.0	74.4	85.8	92.6
BYOL (ours)	ResNet-50 (4×)	375M	69.1	75.7	87.9	92.5
BYOL (ours)	ResNet-200 (2×)	250M	**71.2**	**77.7**	**89.5**	**93.7**

(b) Other ResNet encoder architectures.

Table 2: Semi-supervised training with a fraction of ImageNet labels.

Для оценки semi-supervised представлений дообучим энкодер на части данных вместе с линейным классификатором

BYOL: результаты

Method	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
BYOL (ours)	**75.3**	91.3	**78.4**	**57.2**	**62.2**	**67.8**	60.6	82.5	75.5	90.4	94.2	**96.1**
SimCLR (repro)	72.8	90.5	74.4	42.4	60.6	49.3	49.8	81.4	**75.7**	84.6	89.3	92.6
SimCLR [8]	68.4	90.6	71.6	37.4	58.8	50.3	50.3	80.5	74.5	83.6	90.3	91.2
Supervised-IN [8]	72.3	**93.6**	78.3	53.7	61.9	66.7	**61.0**	**82.8**	74.9	**91.5**	**94.5**	94.7
Fine-tuned:												
BYOL (ours)	**88.5**	**97.8**	86.1	**76.3**	63.7	91.6	**88.1**	**85.4**	**76.2**	91.7	**93.8**	97.0
SimCLR (repro)	87.5	97.4	85.3	75.0	63.9	91.4	87.6	84.5	75.4	89.4	91.7	96.6
SimCLR [8]	88.2	97.7	85.9	75.9	63.5	91.3	88.1	84.1	73.2	89.2	92.1	97.0
Supervised-IN [8]	88.3	97.5	**86.4**	75.8	**64.3**	**92.1**	86.0	85.0	74.6	**92.1**	93.3	**97.6**
Random init [8]	86.9	95.9	80.2	76.1	53.6	91.4	85.9	67.3	64.8	81.5	72.6	92.0

Table 3: Transfer learning results from ImageNet (IN) with the standard ResNet-50 architecture.

Также BYOL показывает хорошие результаты не только на ImageNet, но и на других датасетах

BYOL: результаты

Method	AP_{50}	mIoU
Supervised-IN [9]	74.4	74.4
MoCo [9]	74.9	72.5
SimCLR (repro)	75.2	75.2
BYOL (ours)	**77.5**	**76.3**

(a) Transfer results in semantic segmentation and object detection.

	Higher better			Lower better	
Method	pct.<1.25	pct.$<1.25^2$	pct.$<1.25^3$	rms	rel
Supervised-IN [83]	81.1	95.3	98.8	0.573	**0.127**
SimCLR (repro)	83.3	96.5	99.1	0.557	0.134
BYOL (ours)	**84.6**	**96.7**	**99.1**	**0.541**	0.129

(b) Transfer results on NYU v2 depth estimation.

Table 4: Results on transferring BYOL's representation to other vision tasks.

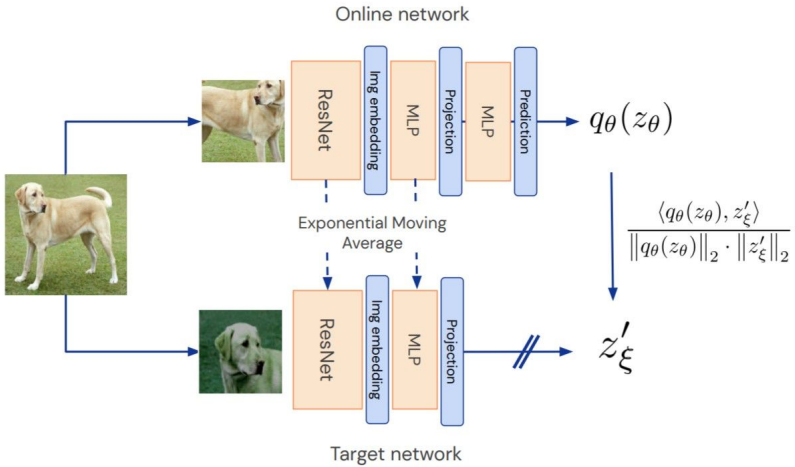
Результаты BYOL также превосходят результаты других методов в других задачах CV, таких как семантическая сегментация, детекция и оценка глубины

Заключение

Статья представляет self-supervised метод обучения представлений, который не опирается на наличие негативных пар в отличие от предыдущих SOTA методов. Предложенный метод показывает SOTA результаты для self-supervised методов и также практически догоняет по результатам supervised методы.

Рецензия. Содержание

- Метод обучения представлений без учителя
- Не опирается на негативные примеры, решается задача предсказания а не дискриминации
- Две сети, одна предсказывает выходы другой и оптимизируется напрямую. Вторая обновляется скользящим средним



Рецензия. Сильные стороны

- Простота и качество: BYOL по своему устройству проще других на момент его выхода SOTA (MoCo v2, SimCLR), устойчивей к выбору аугментаций и размеру батча. С помощью него можно получить лучшее качество

- Метод не опирается на негативные примеры: нет необходимости думать о том, как их выбирать, не нужно хранить в памяти большой набор (как например в MoCo)

- Алгоритм и ход экспериментов описаны ясно и однозначно, есть псевдокод, есть код на гитхабе

- Есть сравнение с текущими SOTA, с другими популярными метода для обучения представлений без учителя

Рецензия. Слабые стороны

Нет строгих теоретических выкладок почему метод не может коллапсировать к вырожденному решению, но авторы объясняют интуицию, которая стоит за методом и опираются на результаты экспериментов.

Также можно отметить что по сути мы обучаем две модели, что затратно по памяти.

Рецензия. Оценка

Статья хорошо написана. Сам метод отлично иллюстрирован и прост для понимания.

Есть код и подробное описание экспериментов. Был опыт использования для обучения представлений облаков точек: BYOL работал немного лучше MoCo.

Моя оценка 9/10, уверенность 5/5

Исследование контекста статьи

- Когда появилась статья?
- Авторы
- Важные источники
- Дальнейшие исследования
- Конкуренты

Когда появилась статья

Препринт опубликован в середине 2020, статья прошла на NeurIPS 2020.

Авторы

Основные авторы:

- Jean-Bastien Grill
- Florian Strub
- Florent Altché
- Corentin Tallec
- Pierre H. Richemond

Связанная статья от похожих авторов:
Bootstrap Latent-Predictive Representations for Multitask Reinforcement Learning

Важные источники

- **Continuous Control With Deep Reinforcement Learning** и **Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results** как источники идеи обновления параметров target сети.
- **Bootstrap Latent-Predictive Representations for Multitask Reinforcement Learning** как источник идеи использования прошлого состояния сети в качестве target’а.

Дальнейшие исследования

- Дальнейшее исследование от тех же авторов: BYOL works even without batch statistics.
- Применение похожей идеи к обработке видео: Broaden Your Views for Self-Supervised Video Learning

Конкуренты

- MoCo (2020 март)
- SwAV (2021 январь)
- EsViT (2021 июнь)