



УНИВЕРСИТЕТ ИТМО

Задачи нейронных подходов в компьютерном зрении. Наборы данных для работы с изображениями

Ефимова Валерия Александровна

vefimova@itmo.ru

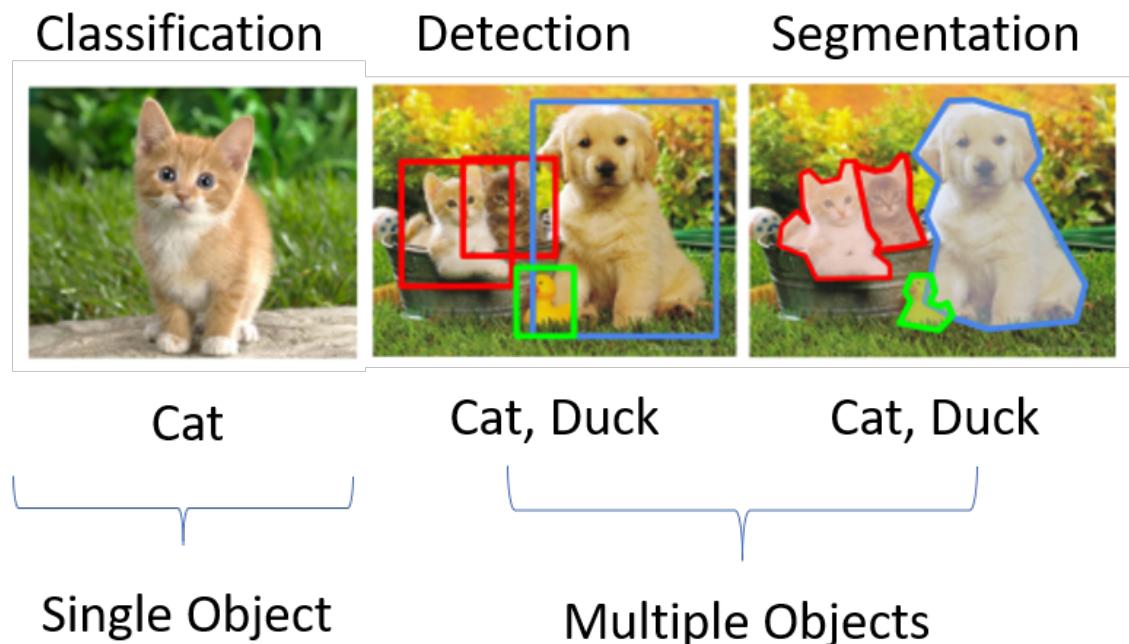
05.10.2022

ОБРАЗОВАТЕЛЬНЫЕ ПРОГРАММЫ В ОБЛАСТИ
ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

- 3 основные задачи обработки изображений
- Наборы реальных данных для работы с изображениями
- Наборы синтетических данных для работы с изображениями
- Аугментация
- Разметка данных
- Оценка качества классификации
- Известные архитектуры для классификации изображений (AlexNet, VGG16, ResNet, Inception, MobileNet, EfficientNet, Swin Transformer)
- Перенос знаний

Основные задачи анализа изображений:

- Классификация (classification)
- Детекция (detection)
- Сегментация (segmentation)
 - Семантическая сегментация (semantic segmentation)
 - Сегментация сущностей (instance segmentation)
 - Паноптическая сегментация (panoptic segmentation)



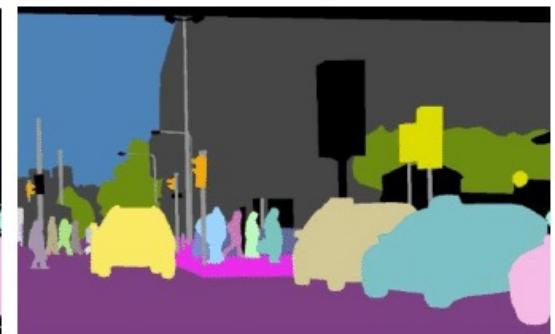
(a) Image



(b) Semantic Segmentation

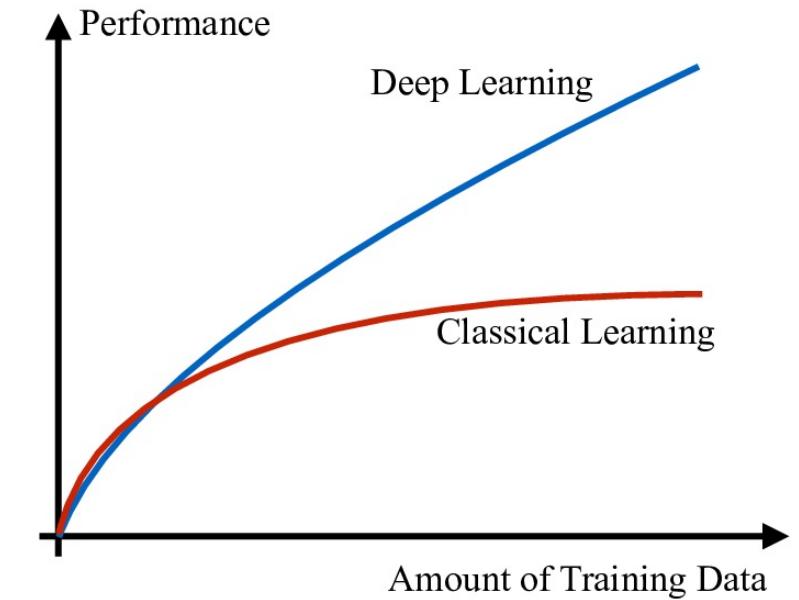
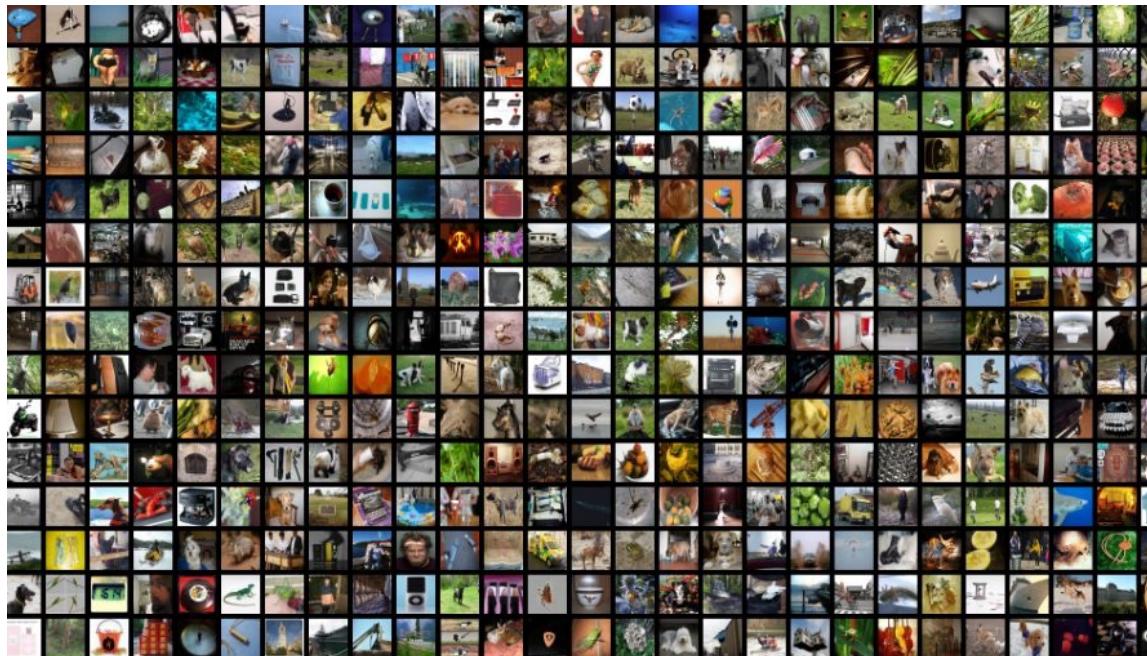


(c) Instance Segmentation

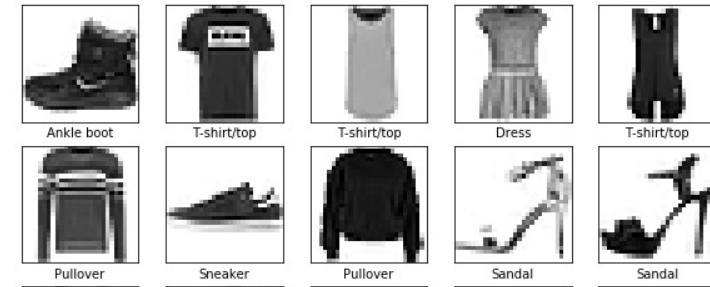
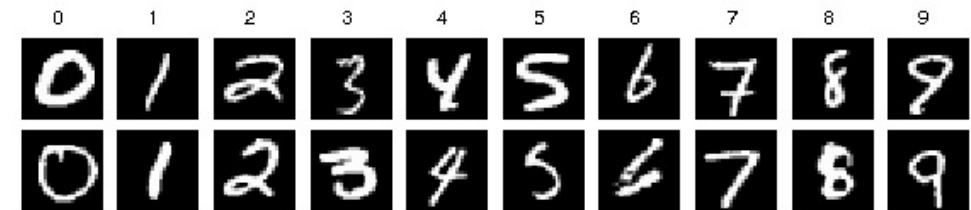


(d) Panoptic Segmentation

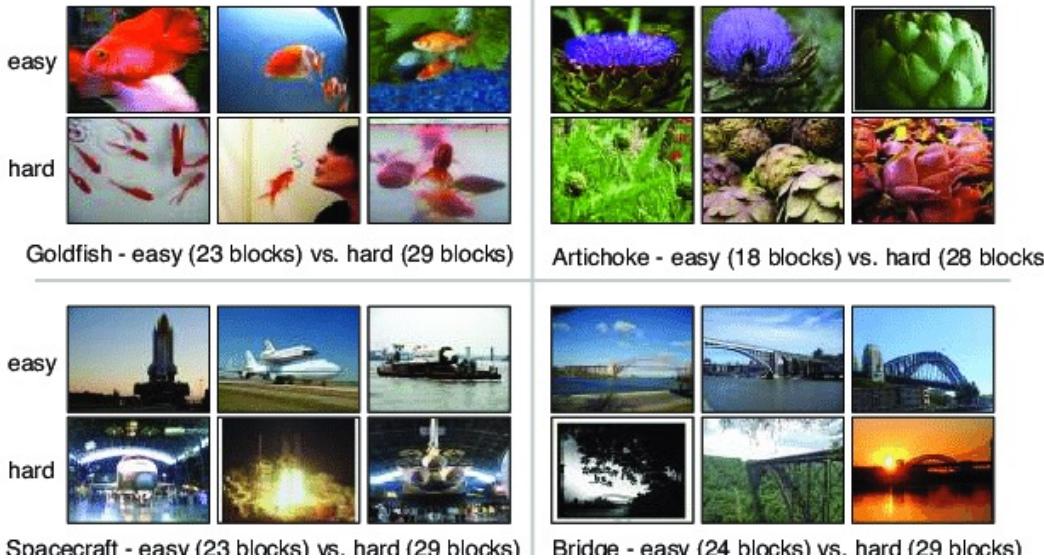
- В последнее время глубокие алгоритмы добились невероятных успехов в обработке и генерации изображений. Эти успехи основаны не только на архитектуре моделей, но и на размерах обучающих выборок и гигантских вычислительных мощностях.
- **Для обучения глубоких моделей нужно колоссальное количество данных.**



- MNIST – рукописные цифры, черно-белые изображения 28x28 пикселя, 70000 изображений.
- MNIST-Fashion – черно-белые фотографии различных видов одежды, 28x28 пикселей, 70000 изображений.
- CIFAR-10 – фотографии объектов 10 классов, цветные изображения 32x32 пикселя, 60000 изображений.



- Imagenet – фотографии с указанием классов объектов на изображении и их позиций, более 14 миллионов изображений.
- Imagenet начали собирать и размечать в 2007 году, а в 2009 он был представлен на CVPR.
- Вместе с публикацией набора данных стартовал конкурс ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

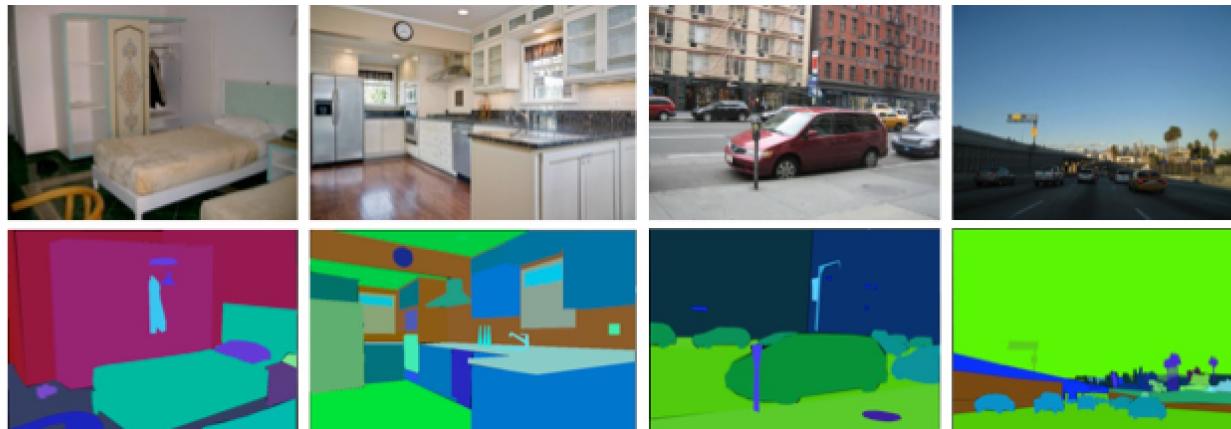


- MS COCO (Microsoft Common Objects in Context) – фотографии сложных повседневных сцен, содержащих объекты в их естественном окружении, 328000 изображений, 91 класс.
- Все объекты находятся в их естественном окружении. Изображения, как правило, содержат объекты разных классов (только 10% имеют единственный класс). Все изображения сопровождаются аннотациями, хранящихся в json формате.
- COCO имеет пять типов аннотаций для разных задач:
 - Задача нахождения объектов на изображении.
 - Обнаружение ключевых точек.
 - Сегментация окружения (англ. Stuff Segmentation). В отличии от задачи обнаружения объектов (человек, кот, машина), здесь внимание фокусируется на том, что его окружает (трава, стена, небо).
 - Паноптическая сегментация.
 - Аннотирование изображения (англ. Caption Evaluation). Генерация сопроводительной подписи к изображению.

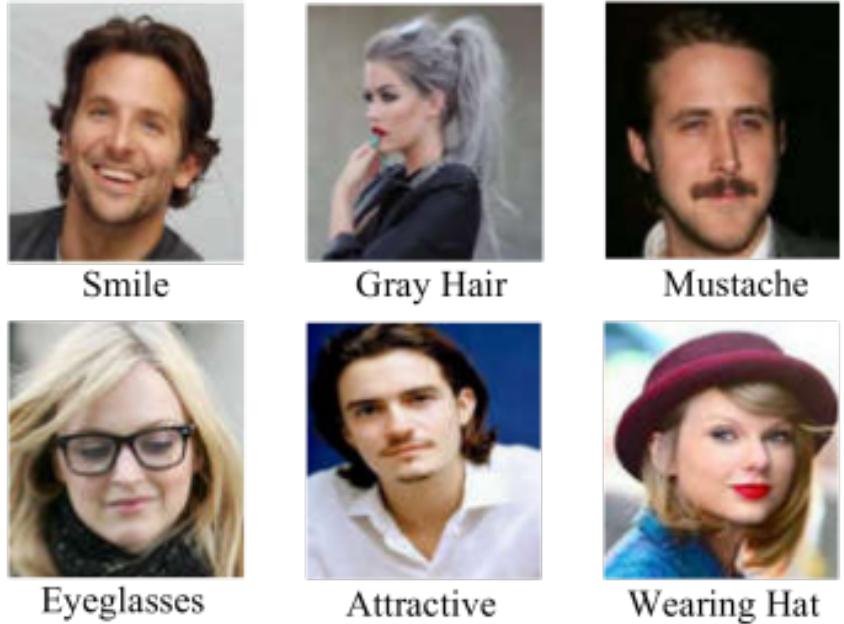


Наборы данных для работы с изображениями. Сегментация и генерация

- CelebA и CelebAHQ (CelebFaces Attributes Dataset) – фотографии лиц знаменитостей, охватывающие большие вариации поворотов головы, более 200000 изображений, 40 бинарных атрибутов.
- CityScapes – изображения городских улиц 50 городов с указанием семантической сегментации сущностей на них, 5000 изображений с разрешением 1024 * 2048.
- ADE20K – фотографии с указанием семантической сегментации сущностей на них. Для каждого объекта также приведена его сегментация на части, ~22 тысячи изображений, 3 169 классов.



ADE20K

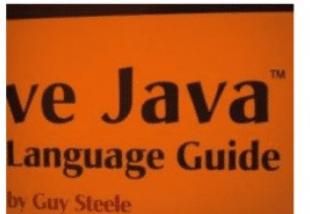


CelebA



CityScapes

- ICDAR 2003-2019 – семейство наборов фотографий с текстом на английском языке. Содержит баундинг боксы текста и сам текст.
 - 2003-2005 – 529 изображений, среди них обложки книг, таблички и другой крупный и хорошо читаемый текст.
 - 2011-2013 – 485+561 изображения, расширение предыдущего набора.
 - 2015 – фотографии торговых центров на гугл-очки.
 - MLT (2017) – текст на разных языках: на вывесках магазинов, объявлениях, табличках и прочее.
- COCO-Text – набор из 63686 изображений для распознавания объектов.
- Total-Text – набор из 1555 изображений, сфокусирован на распознавании изогнутого текста.



ICDAR



COCO-Text



Total-Text

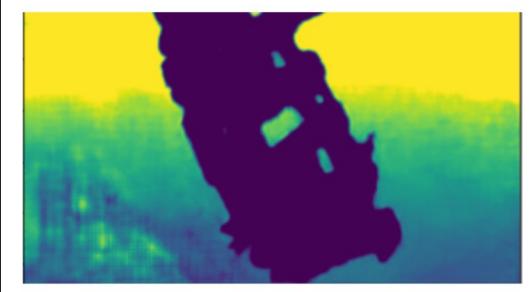
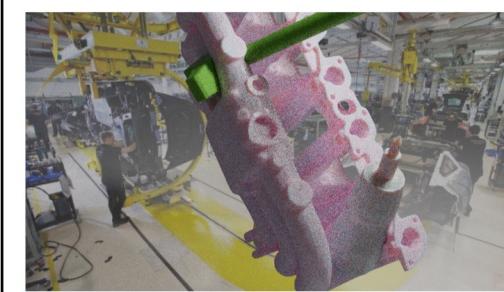
- DeepFashion2
- Cat 256
- MVS Data Set – multiple view stereo
- 3D-FRONT
- LSUN
- Oxford Flowers 102
- *И много наборов данных на Kaggle.
- **В интернете есть трансляции с CCTV камер, например с Times Square в Нью Йорке.



- SynthText in the Wild – 800 000 изображений, 8 000 000 слов



- Генерация в 3D – разместить объекты в 3D-пространстве = легко получить ground truth.



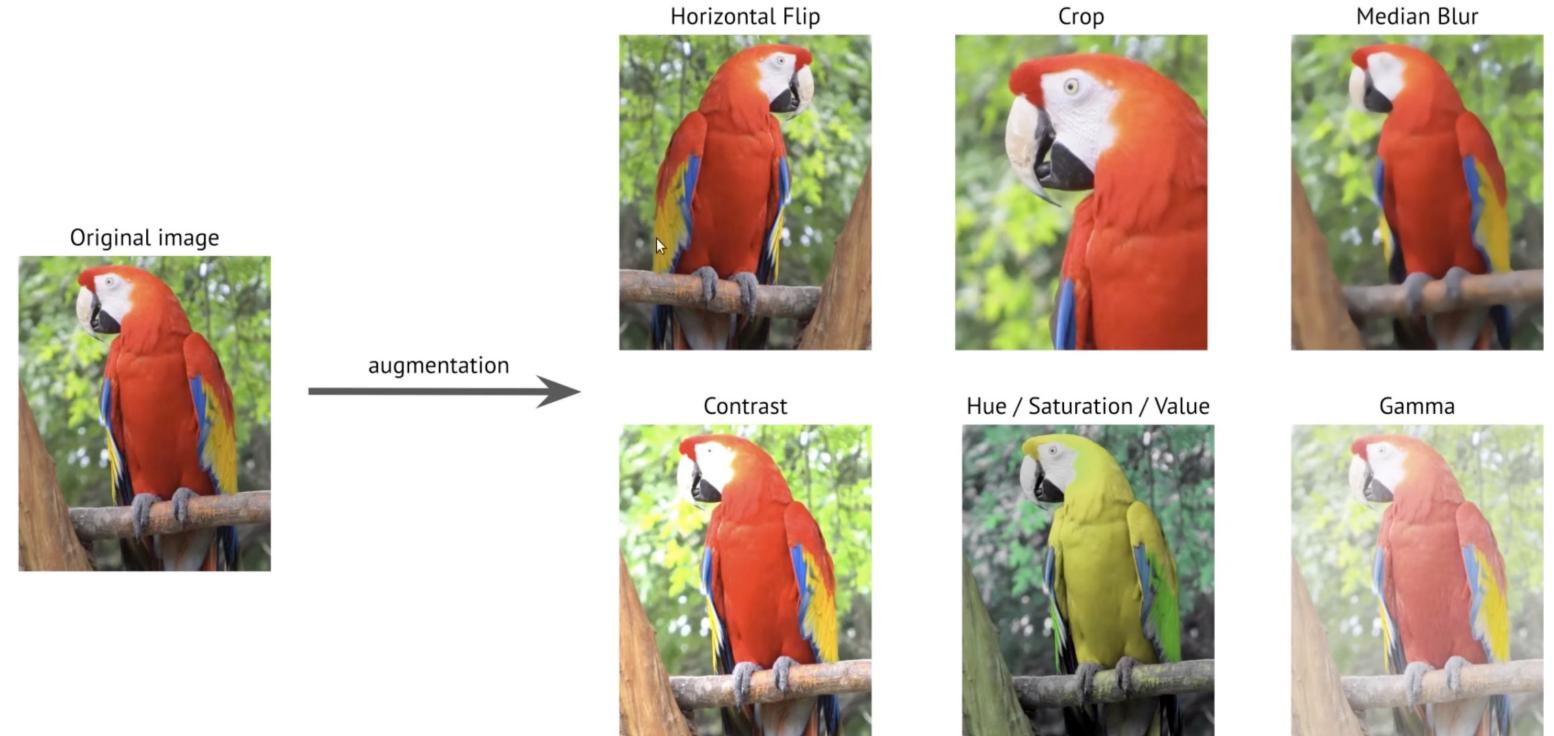
Увеличение обучающей выборки за счет модификации существующих данных.

Аугментация включает:

- операции с цветом;
- геометрические операции;
- операции с объектами.

Примеры аугментаций:

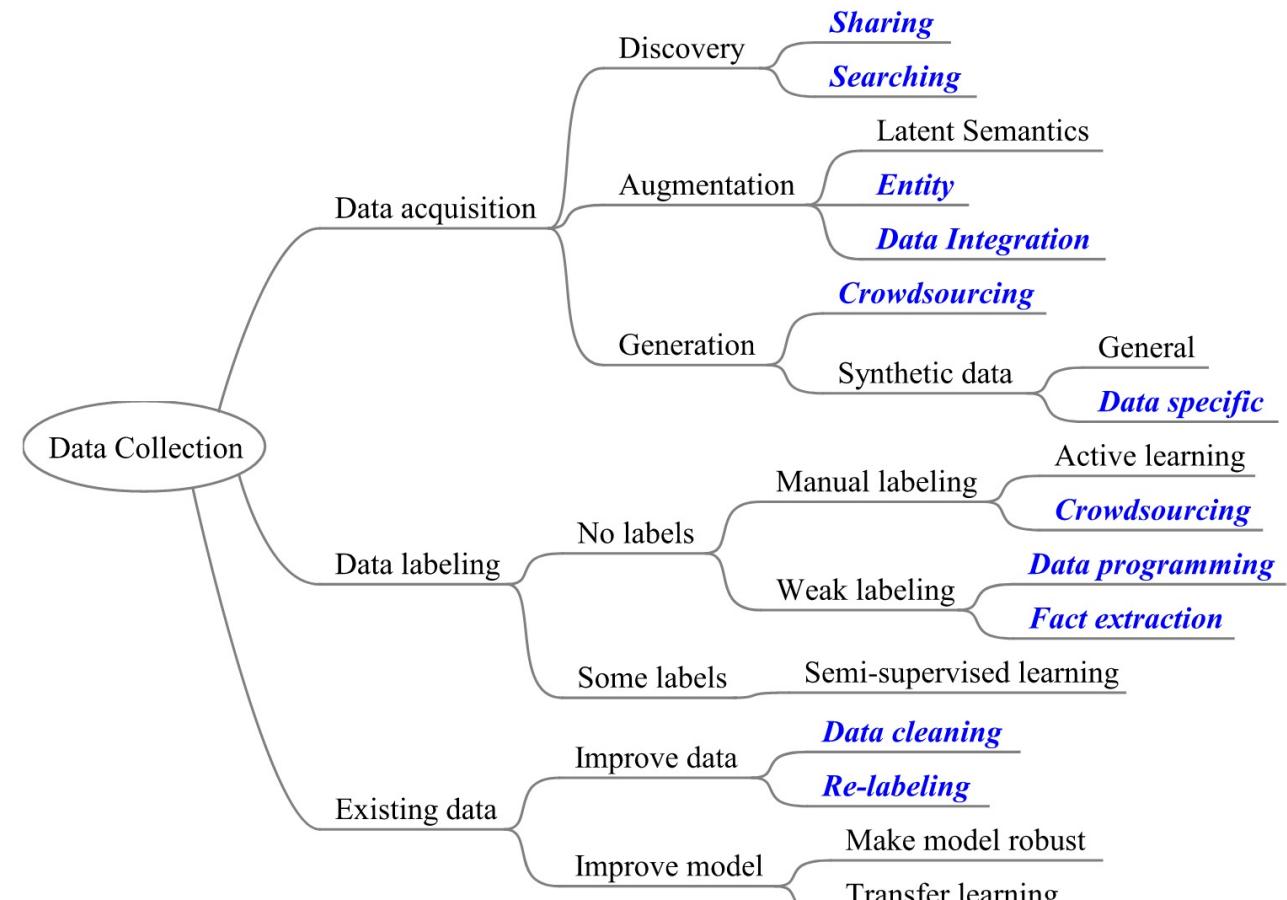
- поворот (flip & rotate);
- изменение цвета и контраста;
- размытие;
- сдвиг и вырезание;
- сжатие и растяжение.



- Часто задача новая, размеченных данных в общем доступе нет.
- Сначала разметить немного, но качественно.
- Как разметать легче?
 - Снизить стоимость разметки одного примера.
 - Разметать меньше примеров.

Инструменты разметки изображений

- CVAT <https://github.com/openvinotoolkit/cvat>
- LabelMe <https://github.com/CSAILVision/LabelMeAnnotationTool>
- LabelStudio <https://labelstud.io/>
- imglab <https://github.com/NaturalIntelligence/imglab>
- Semantic Segmentation Editor
<https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor>



Roh Y., Heo G., Whang S. E. A survey on data collection for machine learning: a big data-ai integration perspective //IEEE Transactions on Knowledge and Data Engineering. – 2019. – Т. 33. – №. 4. – С. 1328-1347.
<https://www.youtube.com/watch?v=ilOpUkt1x-I>



УНИВЕРСИТЕТ ИТМО

Спасибо за внимание!

ОБРАЗОВАТЕЛЬНЫЕ ПРОГРАММЫ В ОБЛАСТИ
ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА



УНИВЕРСИТЕТ ИТМО

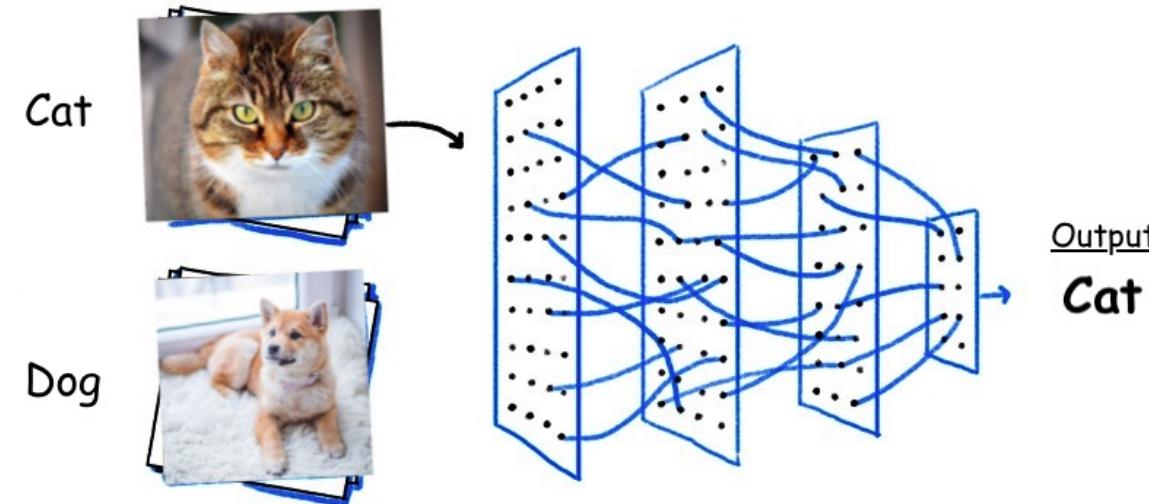
Задача классификации изображений

Ефимова Валерия Александровна

vefimova@itmo.ru

05.10.2022

Нужно отнести изображение целиком к одному из множества заданных классов.



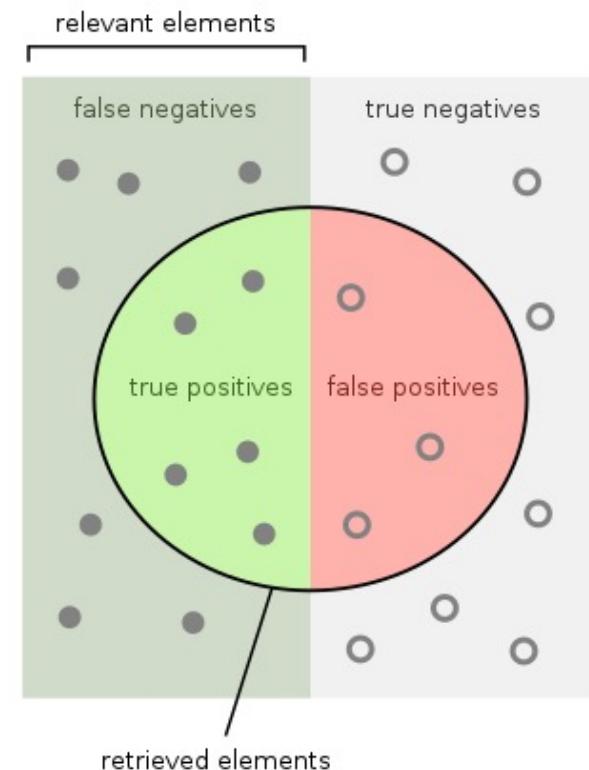
Те же метрики, что и для оценки обычной классификации.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = True positive; FP = False positive; TN = True negative; FN = False negative

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

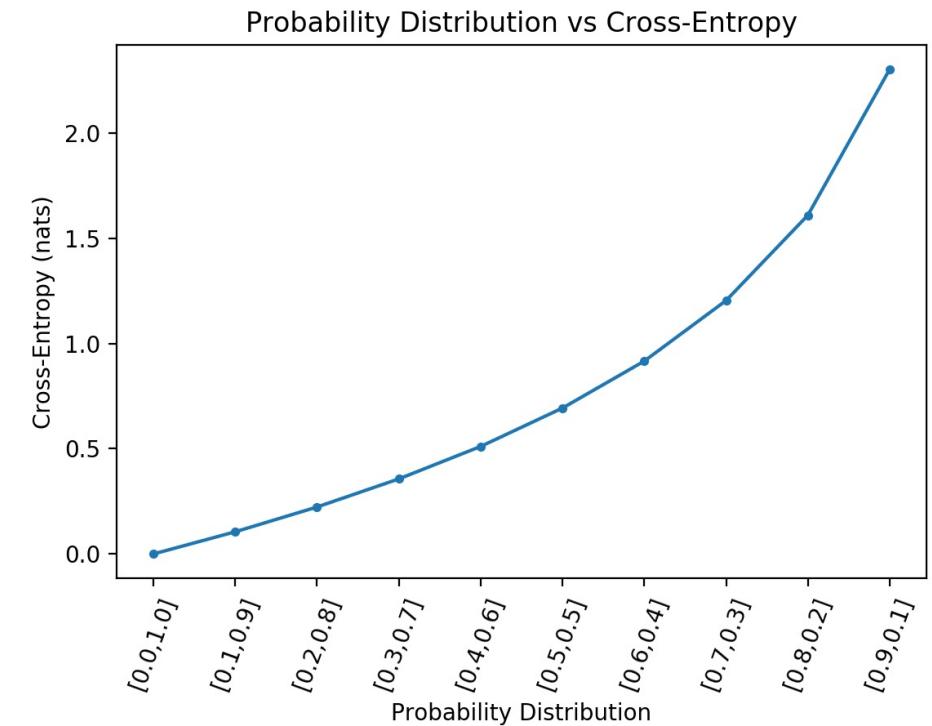
- Измеряет производительность модели классификации, выходной сигнал которой представляет собой значение вероятности от 0 до 1.
- Если классов $M = 2$, то используется двоичная кросс-энтропия:

$$-(y \log(p) + (1 - y) \log(1 - p))$$

- Если классов $M > 2$, то:

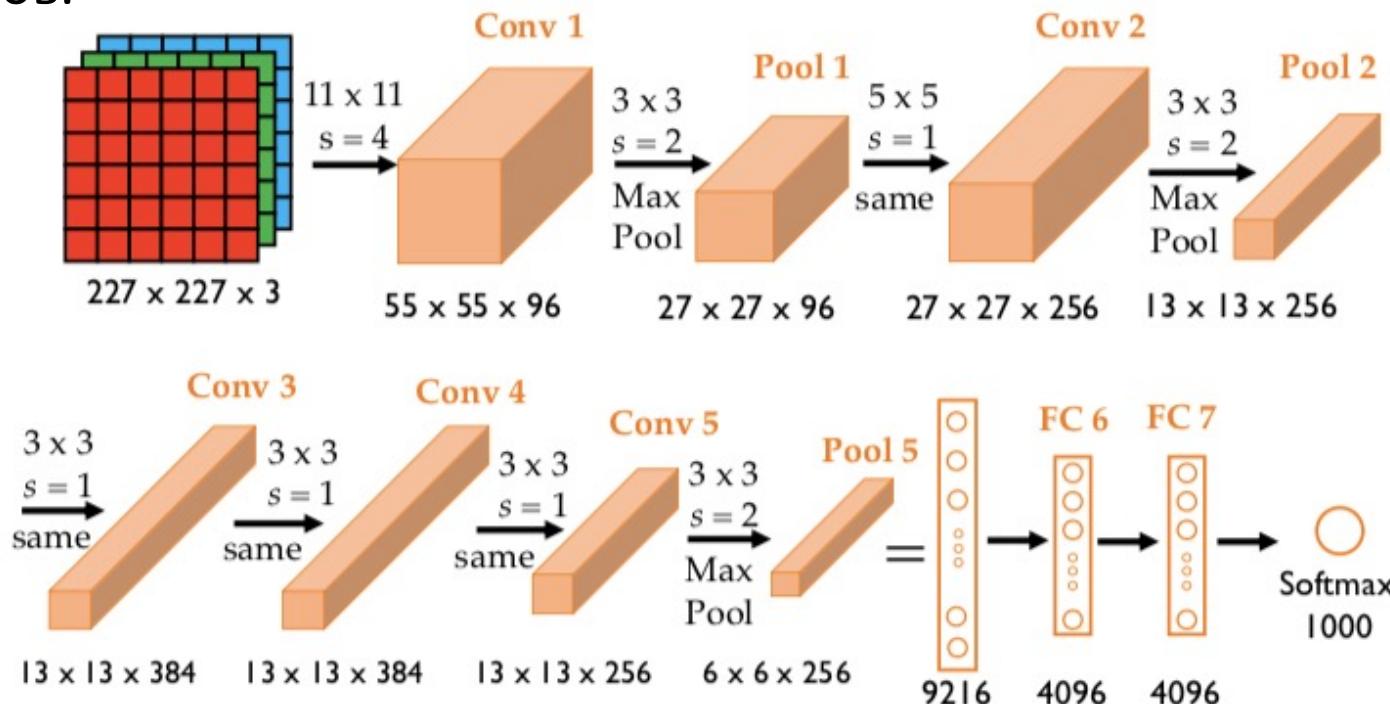
$$-\sum_{c=1}^M y_{o,c} \log(p_{o,c})$$

где M – число классов,
 y – индикатор (0 или 1) если предсказанная метка класса c корректна для объекта o ,
 p – предсказанная вероятность принадлежности объекта к классу c .

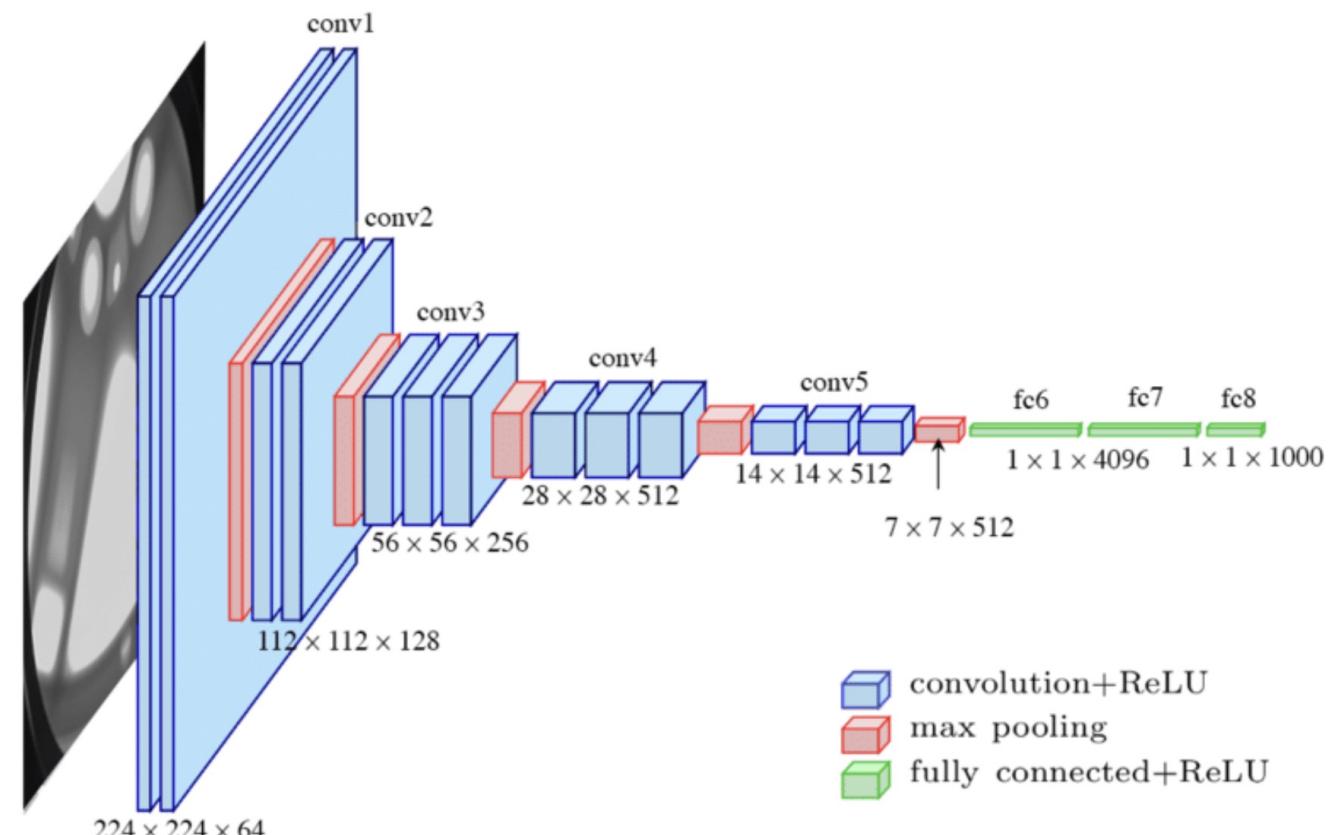


- Cross-Entropy = 0.00: Perfect probabilities.
- Cross-Entropy < 0.02: Great probabilities.
- Cross-Entropy < 0.05: On the right track.
- Cross-Entropy < 0.20: Fine.
- Cross-Entropy > 0.30: Not great.
- Cross-Entropy > 1.00: Terrible.
- Cross-Entropy > 2.00 Something is broken.

- Классификация изображений из Imagenet.
- Полносвязные слои в конце.
- ~6 млн параметров.
- ReLU активации.

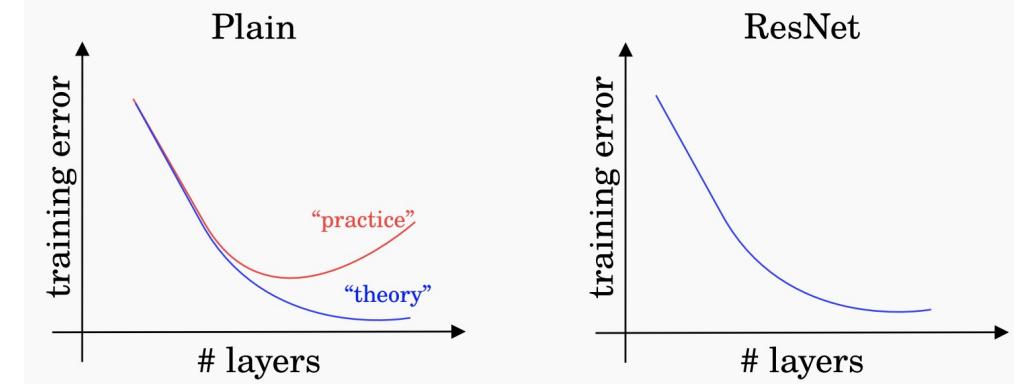
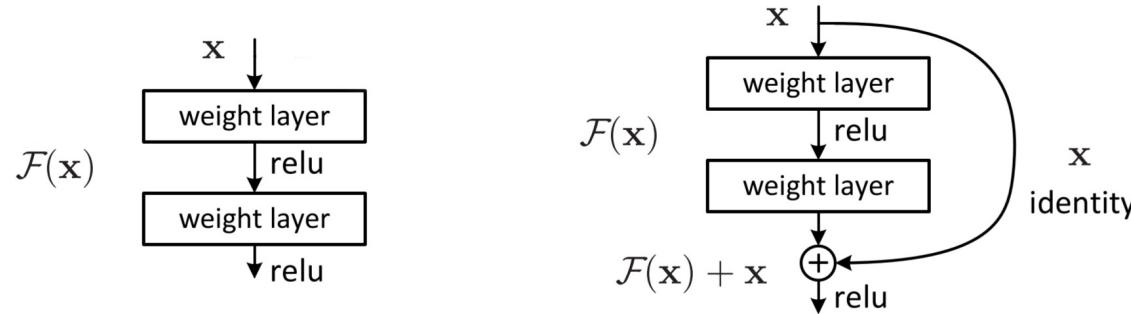


- Важная особенность: хорошо выделяет визуальные признаки изображения (используется для ошибки на схожесть).
- Все свертки с фильтром 3×3 , $s=1$.
- ~ 138 млн параметров.
- VGG19 – больше.
- Паттерн: $n_h, n_w \downarrow, n_c \uparrow$.

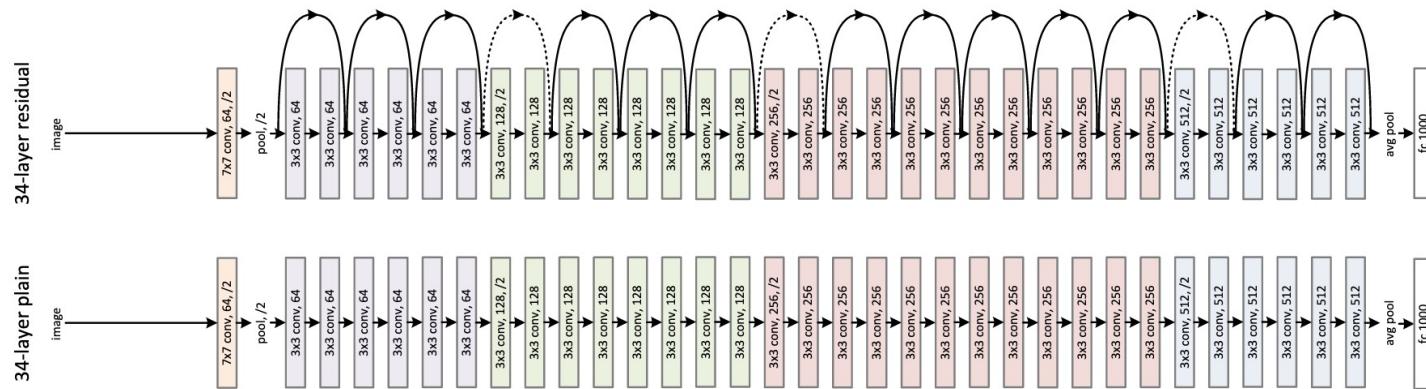


Семейство сетей ResNet (Residual Networks)

- Residual block – информация со слоя l передается на слой $l + 2$ с помощью skip connection = глубже.



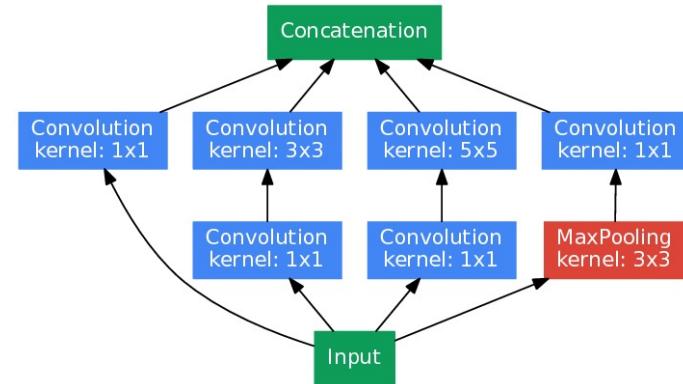
- ResNet 18, 34, 50, 101, 152.
- ResNeXt, SENet, SKNet, ResNeSt.



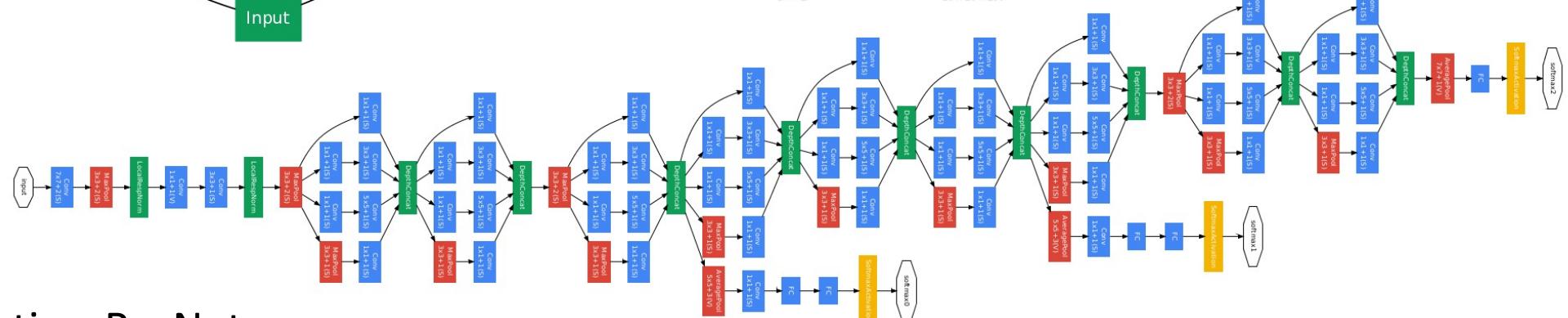
Свёртка 1×1 : $n_w \times n_h \times n_c * 1 \times 1 \times n_f = n_w \times n_h \times n_f$.

Позволяет сократить число каналов без потери качества.

Inception module:
вместо выбора
фильтра сделаем
все!



Inception модули +
max pool +
дополнительные
выходы

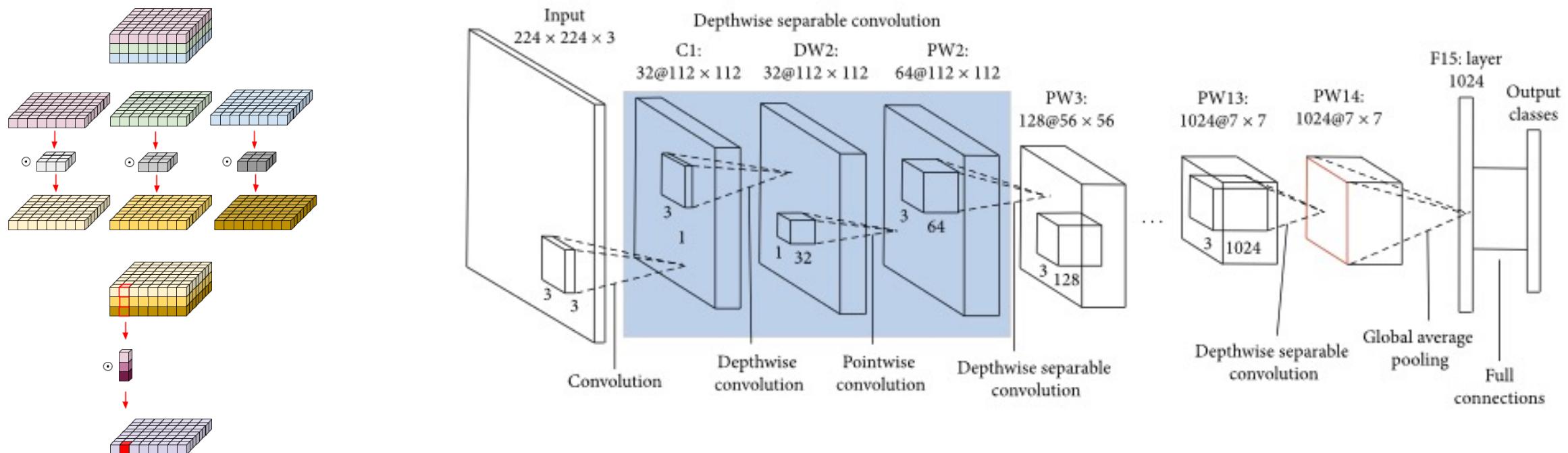


Inception v1-4, Inception-ResNet...

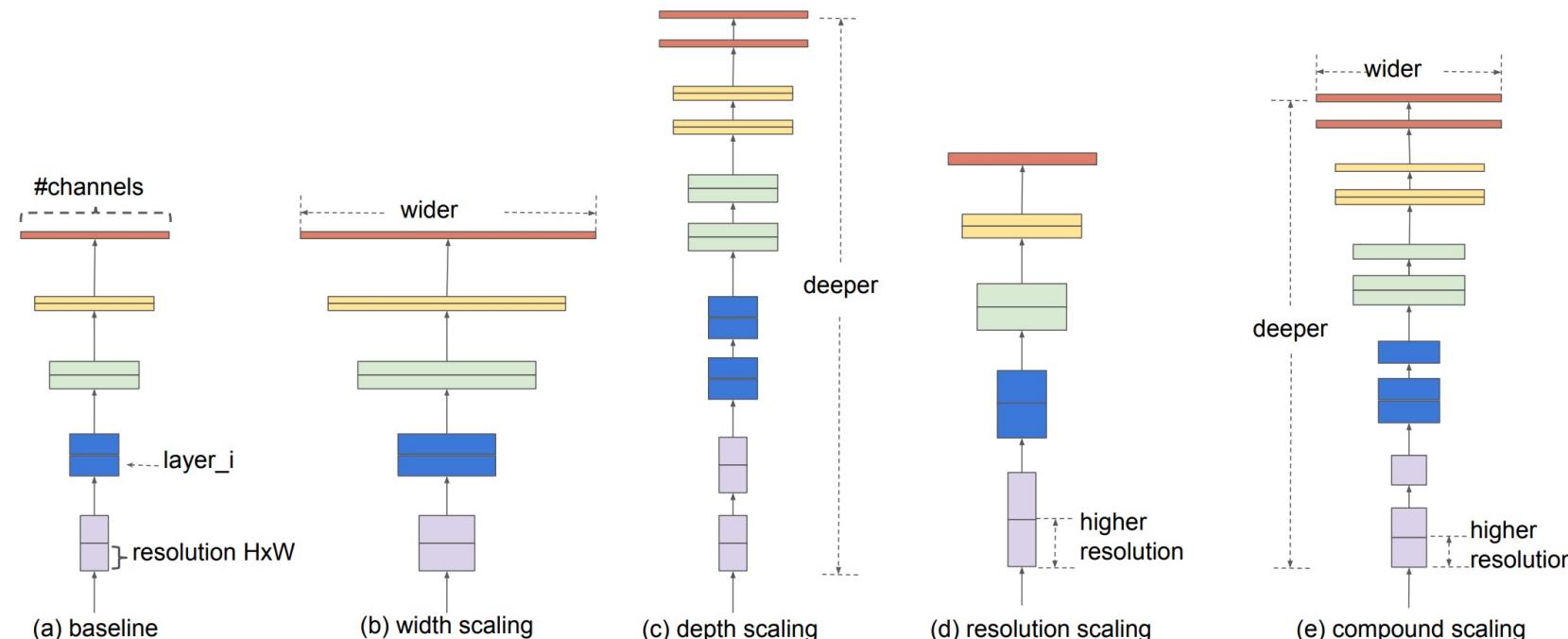
Используется как backbone и для подсчета метрик оценки качества генерации изображений.

1. Lin M., Chen Q., Yan S. Network in network //arXiv preprint arXiv:1312.4400. – 2013.
2. Szegedy C. et al. Going deeper with convolutions //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2015. – С. 1-9.

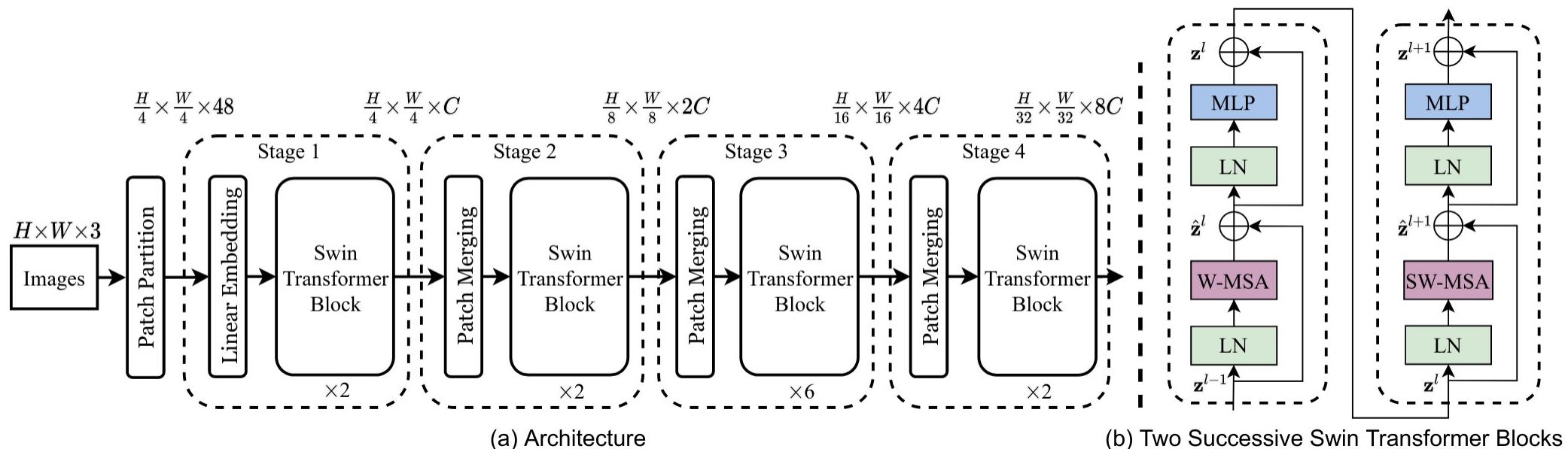
- Разработано для мобильных приложений.
 - Меньше размер и меньше операций за счет Depthwise Separable Convolution:
 - Свернем каждый канал сверткой 3×3 .
 - Свернем сконкатенированный тензор сверткой 1×1 .



- Обычный способ улучшить точность предсказаний – масштабирование сети. Устоявшиеся практики масштабирования: увеличении глубины или ширины CNN, использование большего разрешения входного.
- Предлагают новый метод масштабирования – составное (Compound Model Scaling).
- Архитектура базовой сети найдена с помощью поиска архитектуры нейронной сети (Neural architecture search, NAS).
- Цели масштабирования: увеличение точности и эффективности (FLOPS).

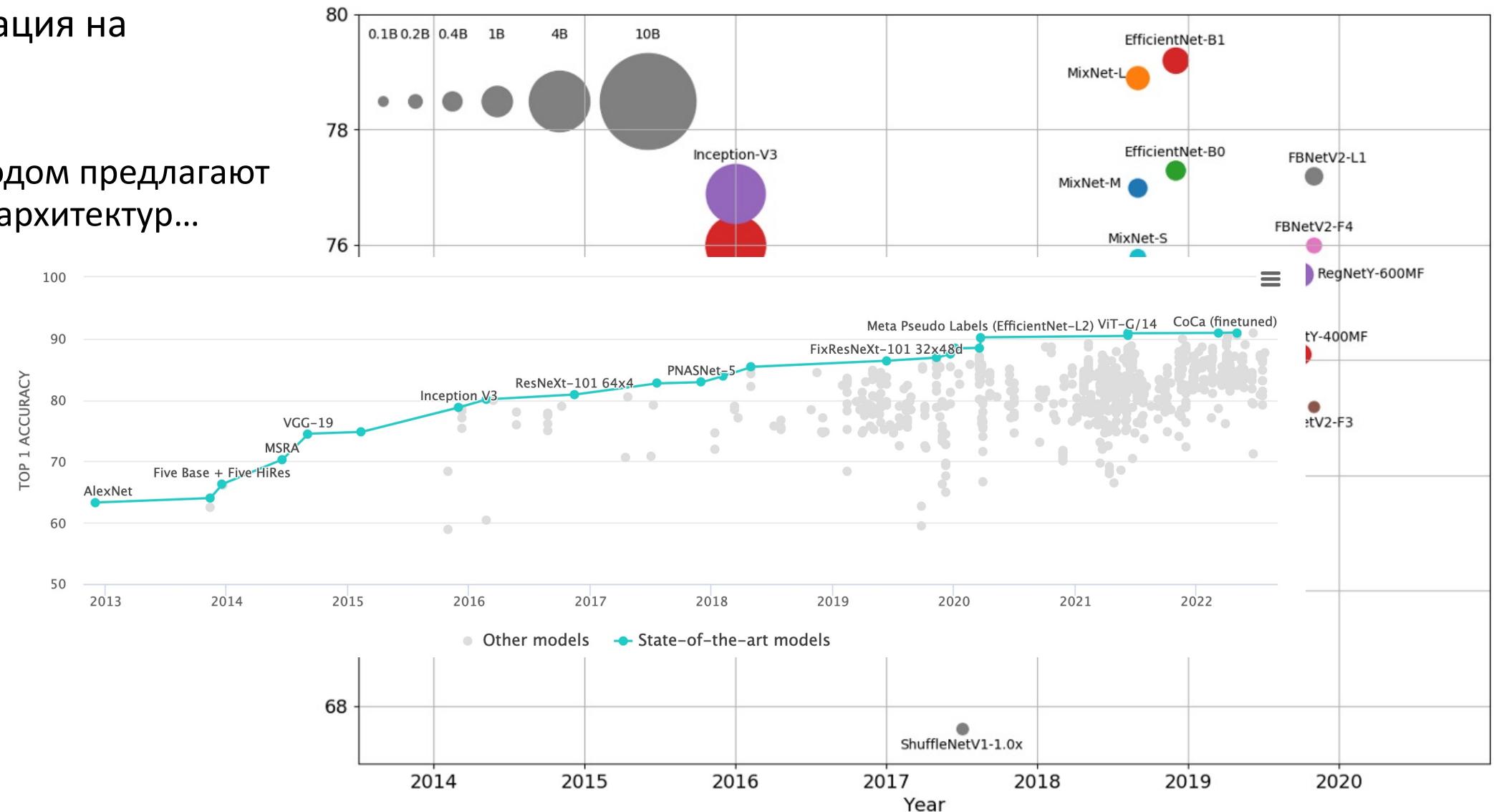


- С 2017 года трансформеры активно используются для обработки текста, с 2020 года появляются адаптации для CV.
- Swin Transformer – Адаптация архитектуры трансформеров для компьютерного зрения.
- Ограничение вычисления самовнимания в пределах сдвигаемого окна.
- Линейная сложность вычислений от размера изображения!
- Исходная картинка нарезается на патчи 4×4 и проецируется линейным слоем.

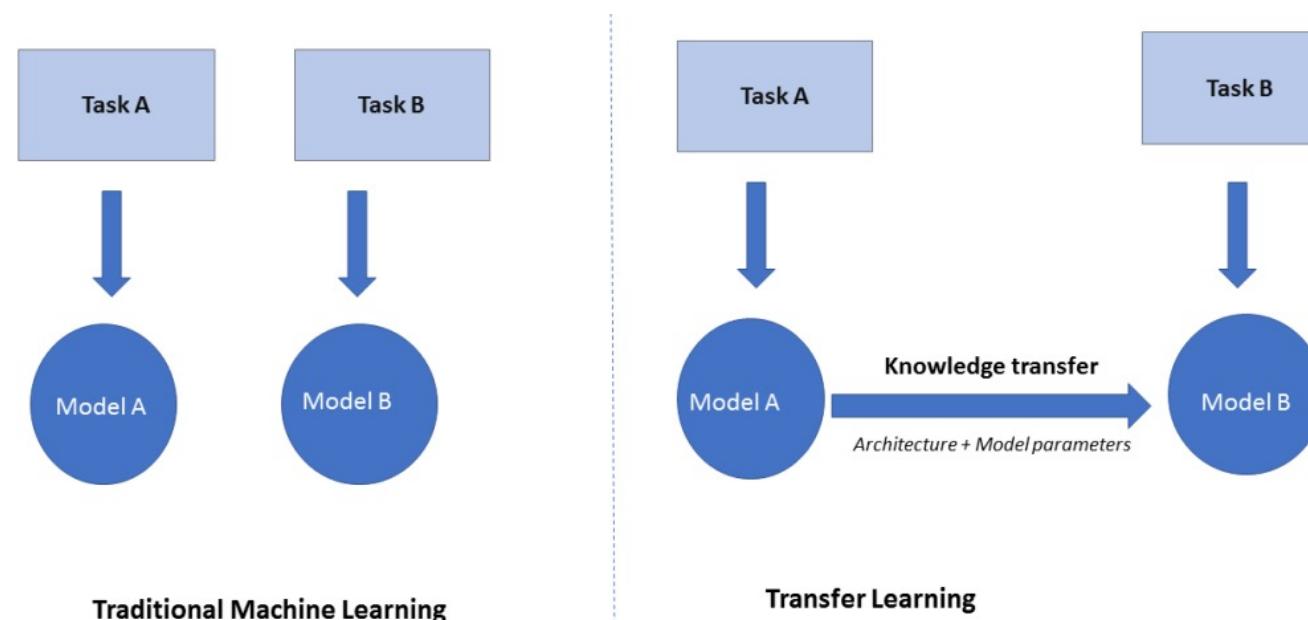


Классификация на ImageNet:

С каждым годом предлагают все больше архитектур...



- Для достижения хорошего качества тренировка сети на 1024+ GPU может занимать недели.
- Намного быстрее не тренировать сеть с нуля, а использовать предобученную сеть (тогда требуется намного меньше данных!)
- Перенос знаний или трансферное обучение – это применение знаний, полученных из одной задачи, к другой задаче.
- Обычно обучаю 1-2 последних слоя, замораживая остальные (freeze).



- 3 основные задачи обработки изображений: классификация, детекция и сегментация.
- Наборов реальных данных много, их можно искать в открытом доступе, например на [kaggle.com](https://www.kaggle.com) или в [OpenML](https://www.openml.org).
- Есть и наборы синтетических изображений, под конкретную задачу можно самостоятельно синтезировать набор данных (например, с использованием 3D).
- Набор данных можно увеличить в разы, используя аугментацию.
- Для новых задач может не быть набора изображений, тогда нужно его разметить.
- Качество классификации изображений оценивают как классическую классификацию.
- Рассмотрели архитектуры для классификации изображений (AlexNet, VGG16, ResNet, Inception, MobileNet, EfficientNet, Swin Transformer).
- Обученные классические сети есть в открытом доступе, для новой задачи для экономии ресурсов можно использовать перенос знаний.

- <https://www.kaggle.com/code/pmigdal/transfer-learning-with-resnet-50-in-pytorch>
- <https://www.pluralsight.com/guides/introduction-to-resnet>
- https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html



УНИВЕРСИТЕТ ИТМО

Спасибо за внимание!

ОБРАЗОВАТЕЛЬНЫЕ ПРОГРАММЫ В ОБЛАСТИ
ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА