

Лекция 11. Vision Transformers. Self-supervised и contrastive Learning.

Глубинное обучение

Антон Кленицкий

Recap

Виды предобученных моделей

Encoder-only

- Классификация
- ELMo, BERT, RoBERTa

Decoder-only

- Text generation
- GPT family

Encoder-Decoder

- Translation, summarization
- BART, T5

BERT

Devlin J. et al. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding.

Transformer Encoder

Представление входных данных:

- Subword tokenization
- Специальные токены: [CLS], [SEP]
- Token + Position + Segment embeddings

Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	# #ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{##ing}$	$E_{[SEP]}$
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

BERT

Masked Language Modeling (MLM) + Next Sentence Prediction

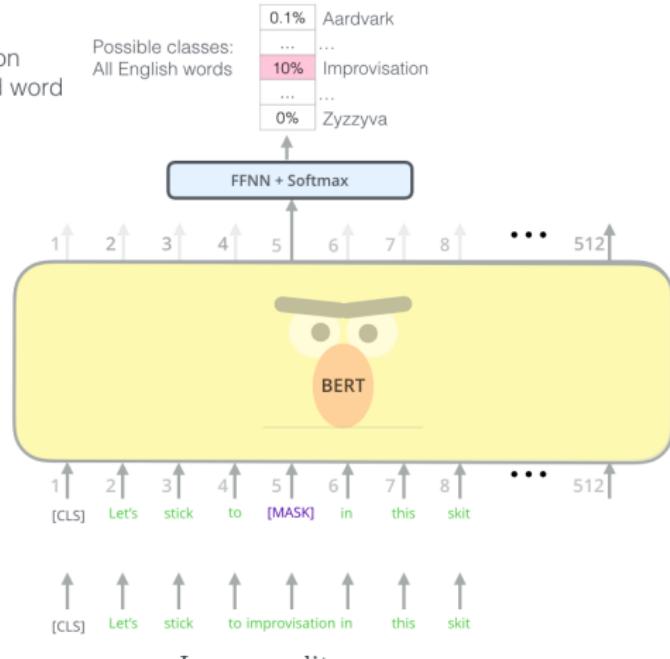
Use the output of the masked word's position to predict the masked word

Possible classes:
All English words



FFNN + Softmax

Randomly mask 15% of tokens



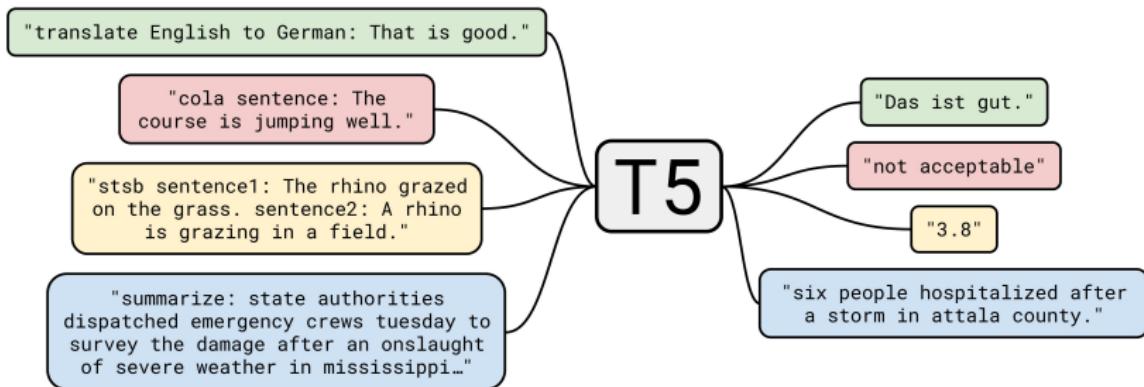
Input

Image credit

T5 (Text-to-Text Transfer Transformer)

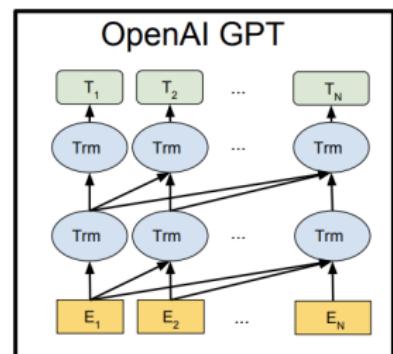
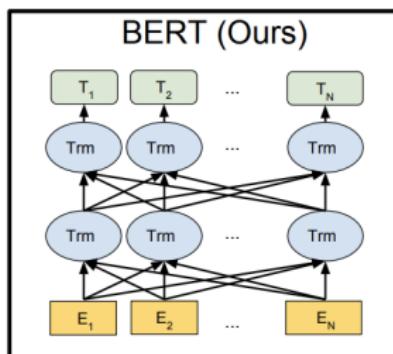
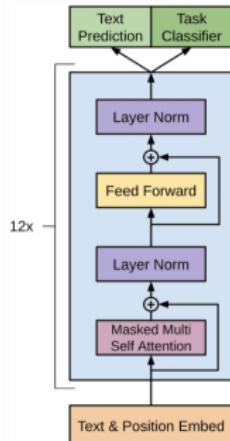
Raffel et al. (2019) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

- Encoder-Decoder
- Формулируем любую задачу как text-to-text
- К входным данным добавляется текстовый префикс конкретной задачи



Generative Pre-trained Transformer (GPT)

- Transformer Decoder
- Просто предсказываем следующий токен



GPT family

Последовательное увеличение количества данных, размера моделей и вычислений

GPT-1

- 117M параметров, L=12, H=768, A=12
- Размер контекста 512 токенов

GPT-2

- 1.5B параметров, L=48, H=1600
- Размер контекста 1024 токенов

GPT-3

- 175B параметров, L=96, H=12288, A=96
- Размер контекста 2048 токенов

GPT-4

- ???
- Размер контекста 8192 и 32768 токенов

Zero-shot и Few-shot learning

Prompt engineering - новая парадигма

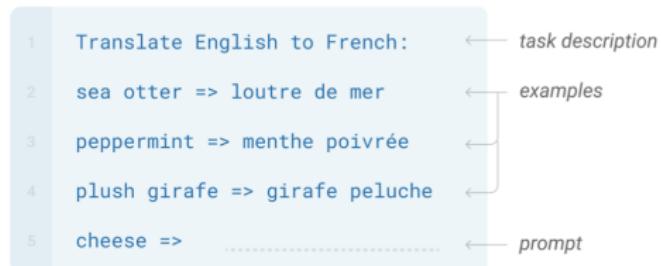
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



ChatGPT

Step 1

Collect demonstration data, and train a supervised policy.

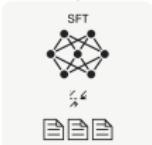
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



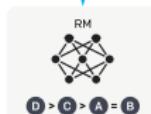
Moon is natural satellite of...

People went to the moon...

A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

r_k

Image credit

Vision Transformers

Transformers for Computer Vision

Можно рассматривать изображение как последовательность пикселей

Проблемы такого подхода:

- Большая размерность изображений

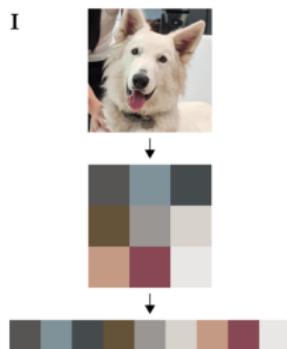


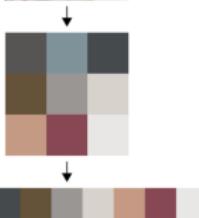
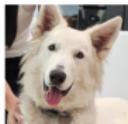
Image credit

Transformers for Computer Vision

Можно рассматривать изображение как последовательность пикселей

Проблемы такого подхода:

I



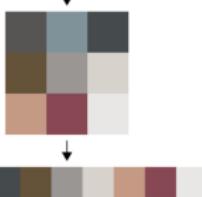
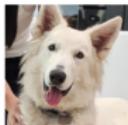
- Большая размерность изображений
-> Разбить изображение на куски (патчи) и рассматривать их как токены

Image credit

Transformers for Computer Vision

Можно рассматривать изображение как последовательность пикселей

I



Проблемы такого подхода:

- Большая размерность изображений
-> Разбить изображение на куски (патчи) и рассматривать их как токены
- В отличие от сверточных сетей трансформер не учитывает структуру данных

Image credit

Transformers for Computer Vision

Можно рассматривать изображение как последовательность пикселей

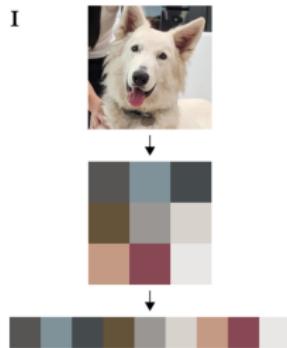


Image credit

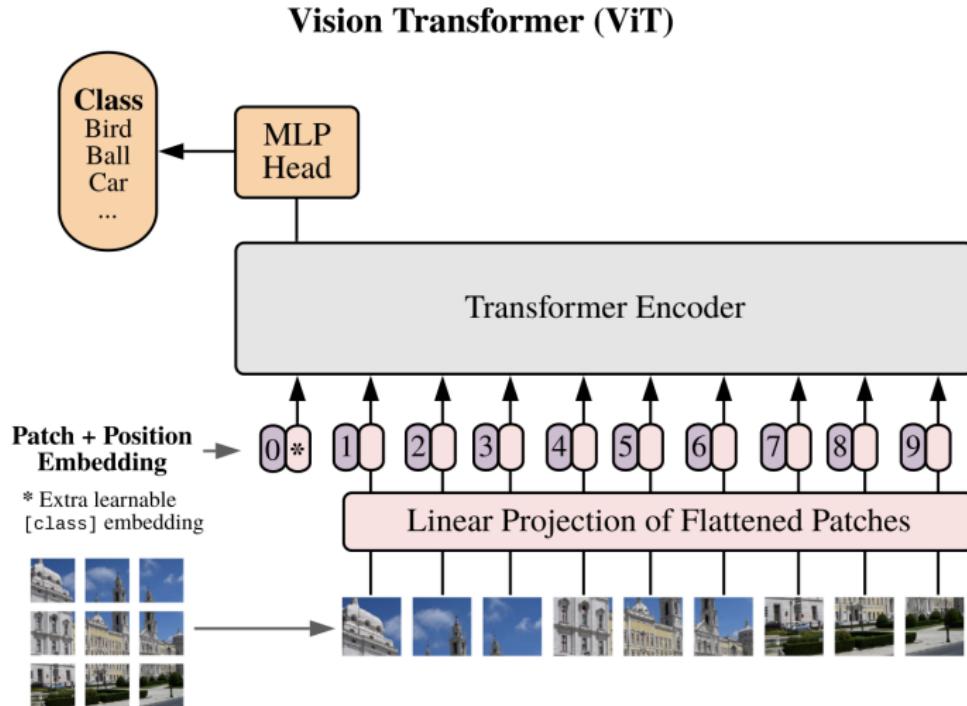
Проблемы такого подхода:

- Большая размерность изображений
-> Разбить изображение на куски (патчи) и рассматривать их как токены
- В отличие от сверточных сетей трансформер не учитывает структуру данных
-> Взять больше данных, чтобы модель смогла выучить структуру

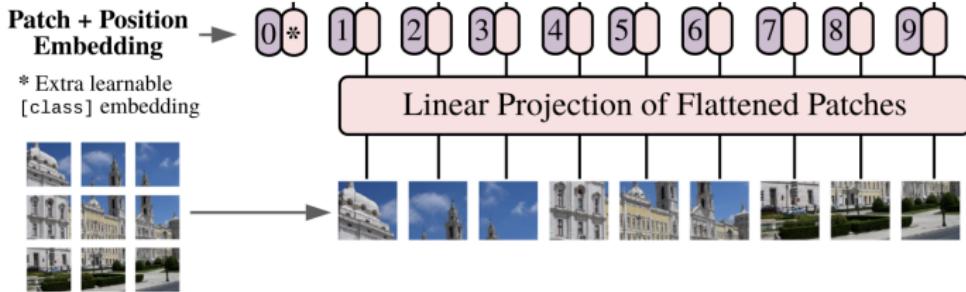
Трансформеры - data hungry, но лучше масштабируются, чем CNN

Vision Transformer

Dosovitskiy A. et al. (2020) *An image is worth 16x16 words: Transformers for image recognition at scale.*



Vision Transformer



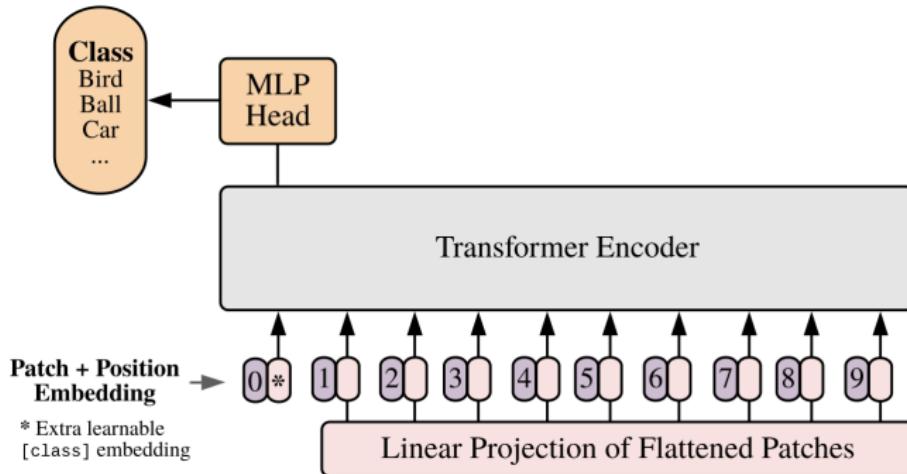
- Разбиваем изображение на патчи, разворачиваем их в последовательность
 - Flatten - превращаем патчи в одномерные вектора
 - Делаем их линейную проекцию - получаем эмбеддинги

Размер изображения $H \times W \times C$

Размер патча $P \times P$

Получаем HW/P^2 патчей размерностью P^2C

Vision Transformer



- Добавляем к последовательности эмбеддинг специального CLS токена для классификации
- Добавляем learnable position embedding
- Прогоняем через стандартный Transformer Encoder
- Поверх CLS токена - MLP Head для классификации

Vision Transformer

- Обучение - Supervised pretrain на огромном датасете
- После этого делаем Finetune на downstream task

Датасет JFT - 18k классов, 303M изображений

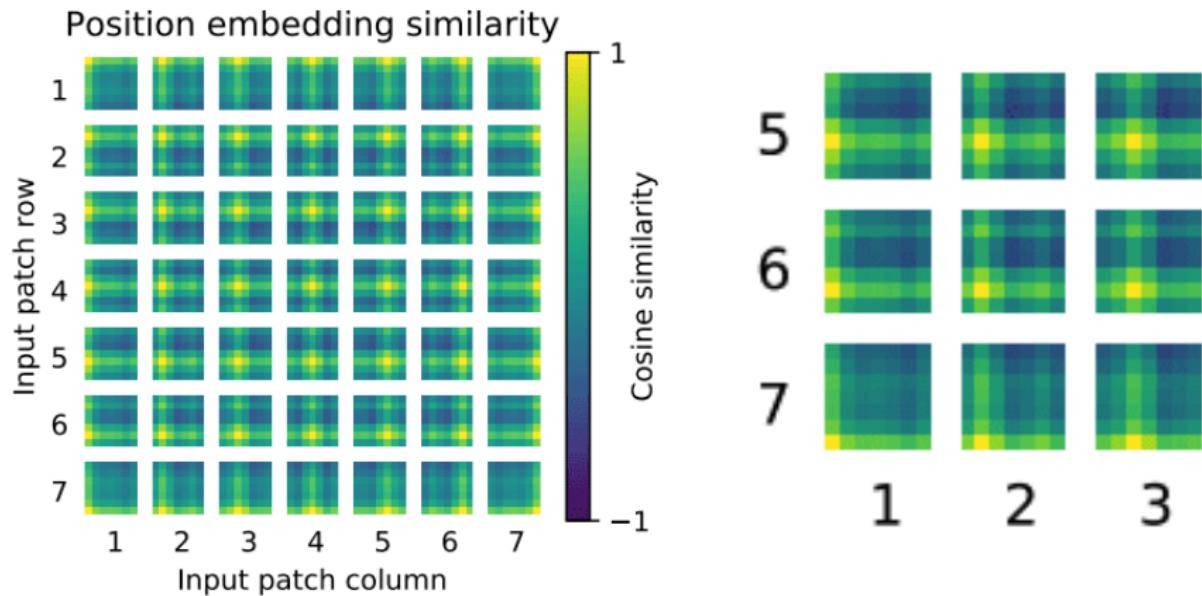
Для сравнения - в ImageNet-21k 14M изображений

Размеры моделей:

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

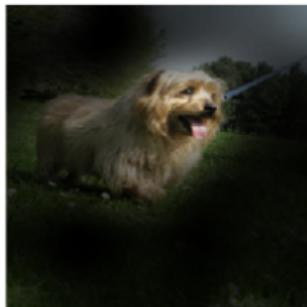
Vision Transformer

Выученные позиционные эмбеддинги



Vision Transformer

Input Attention



Swin transformer

Проблема ViT - низкое разрешение, плохо подходит для задач, где важны мелкие детали (object detection, segmentation)

Liu Z. et al. (2021) Swin transformer: Hierarchical vision transformer using shifted windows.

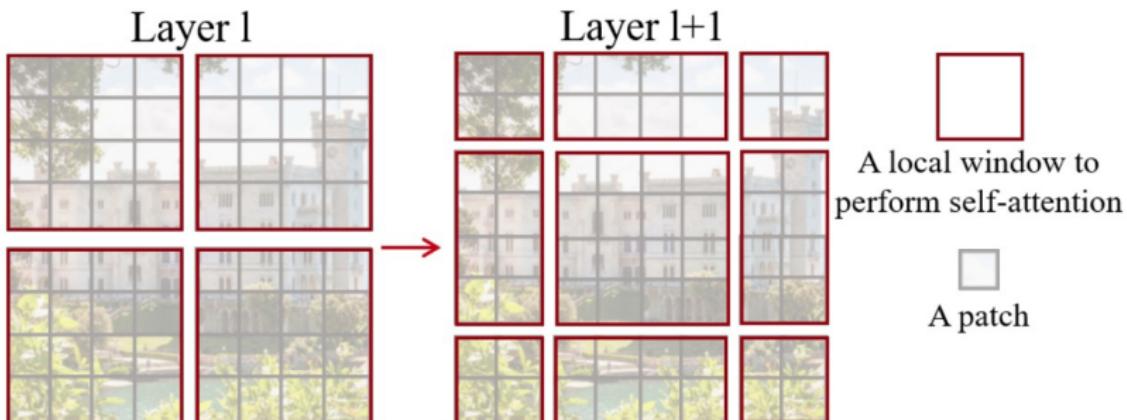
Swin (Shifted Windows) Transformer

- Local self-attention within window
- Окна сдвинуты относительно друг друга в разных слоях
- Построение иерархии представлений - маленькие патчи на нижних уровнях сливаются в большие патчи на верхних уровнях

Swin transformer

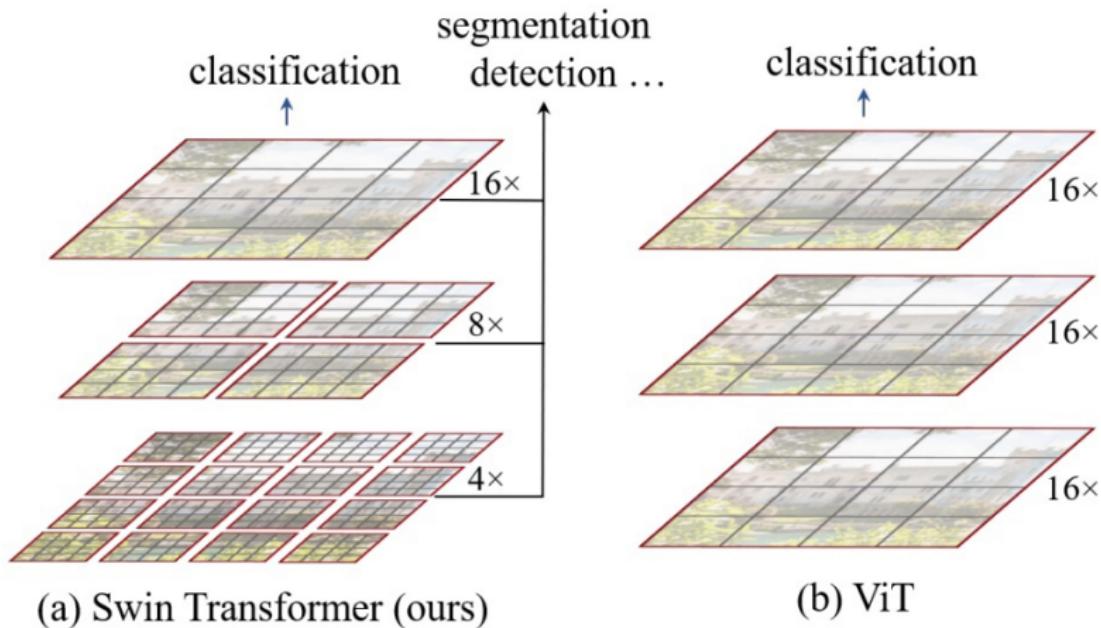
Shifted Window Multi-Head Attention

- Self-attention внутри блоков патчей
- Окна сдвигаются на следующем слое



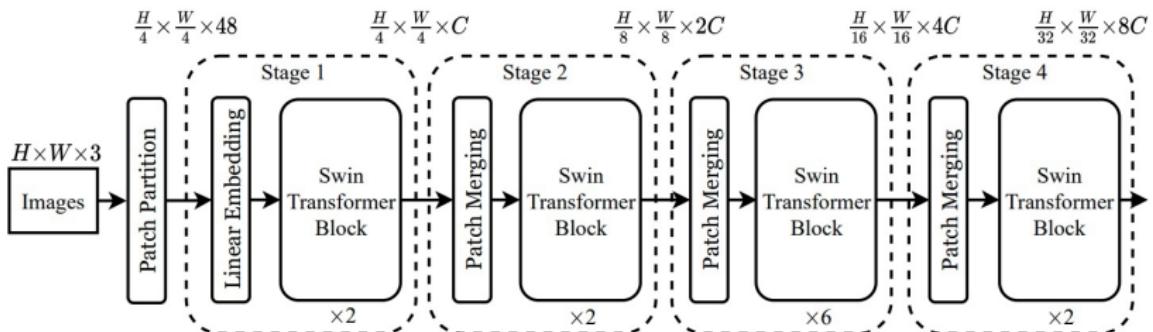
Swin transformer

Объединение патчей на верхних уровнях



Swin transformer

Итоговая архитектура



Размеры моделей

- Swin-T: $C = 96$, layer numbers = {2, 2, 6, 2}
- Swin-S: $C = 96$, layer numbers = {2, 2, 18, 2}
- Swin-B: $C = 128$, layer numbers = {2, 2, 18, 2}
- Swin-L: $C = 192$, layer numbers = {2, 2, 18, 2}

Self-supervised learning

Self-supervised learning

Если есть большая неразмеченная выборка, то можно

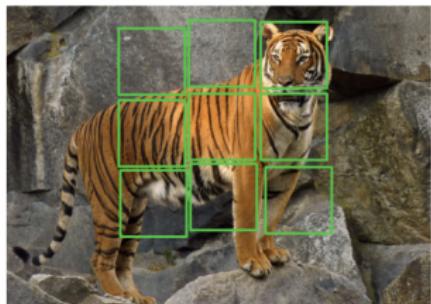
- Придумать вспомогательную задачу (pretext task), для которой можно автоматически получить разметку
- Обучить модель решать эту задачу, при этом интерес представляет не качество полученного решения, а полученные представления объектов (representation learning)
- Полученное представление можно использовать в любой другой задаче (downstream task)

Natural Language Processing

- Word2vec
- Masked Language Modeling (BERT)
- Causal Language Modeling (GPT)
- и другие задачи

Jigsaw (разгадывание паззла)

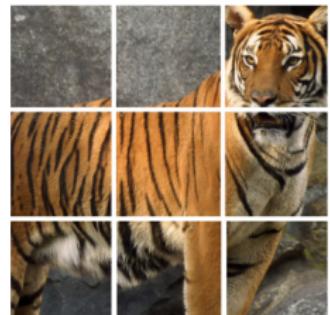
Norooz M. et al. (2016) *Unsupervised learning of visual representations by solving jigsaw puzzles.*



(a)



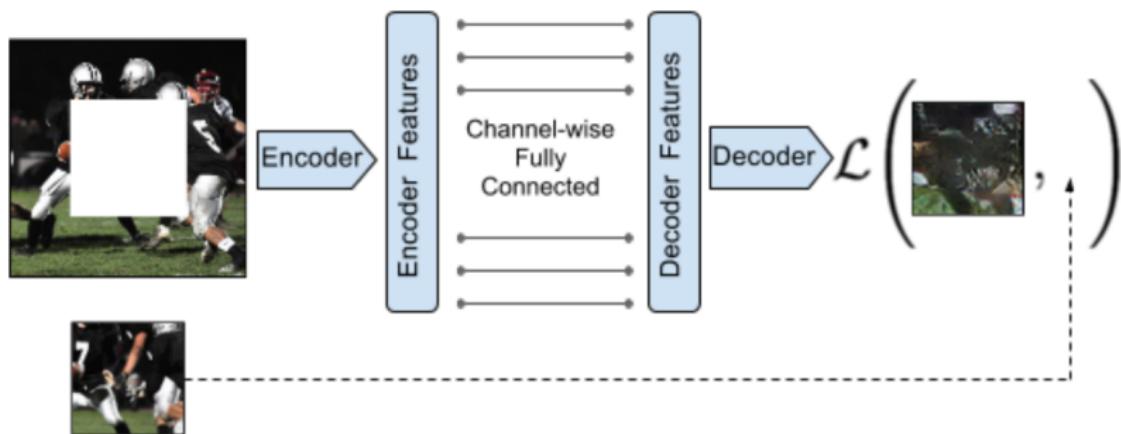
(b)



(c)

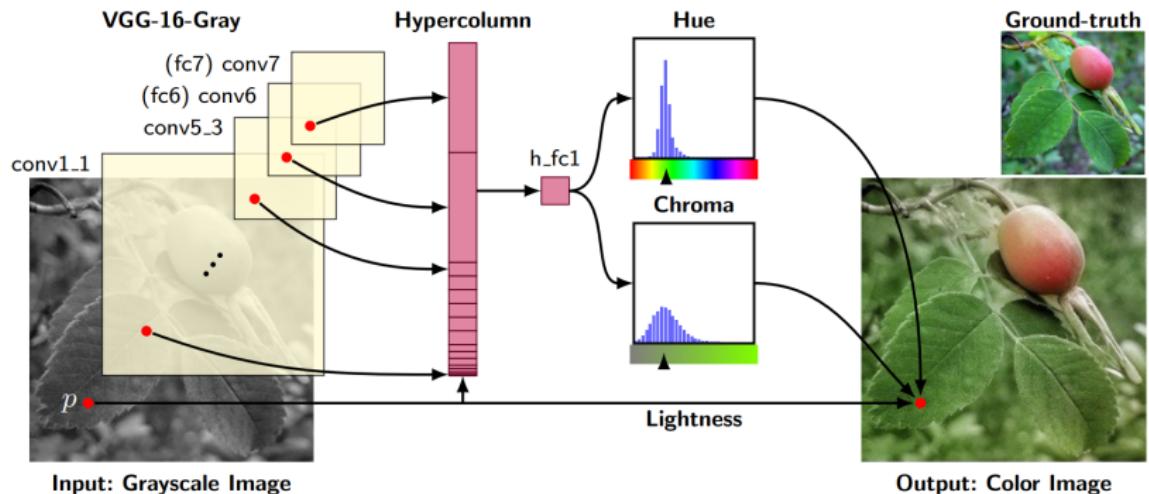
Inpainting

Pathak D. et al. (2016) Context encoders: Feature learning by inpainting.



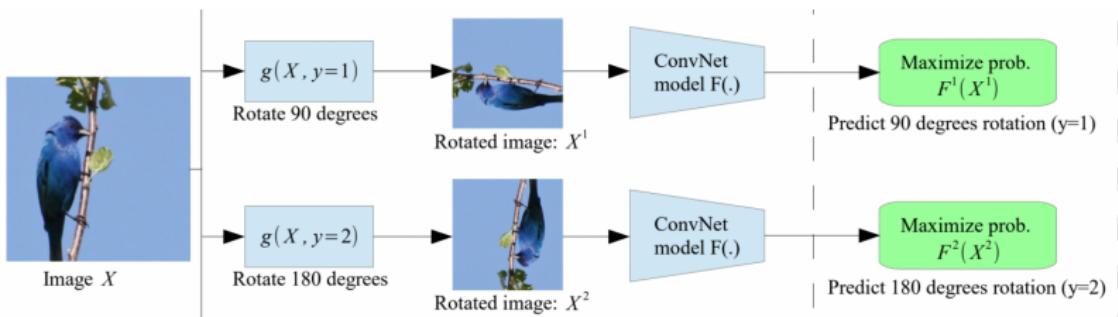
Colorization

Larsson G. et al. (2016). Learning representations for automatic colorization.



Predicting image rotation

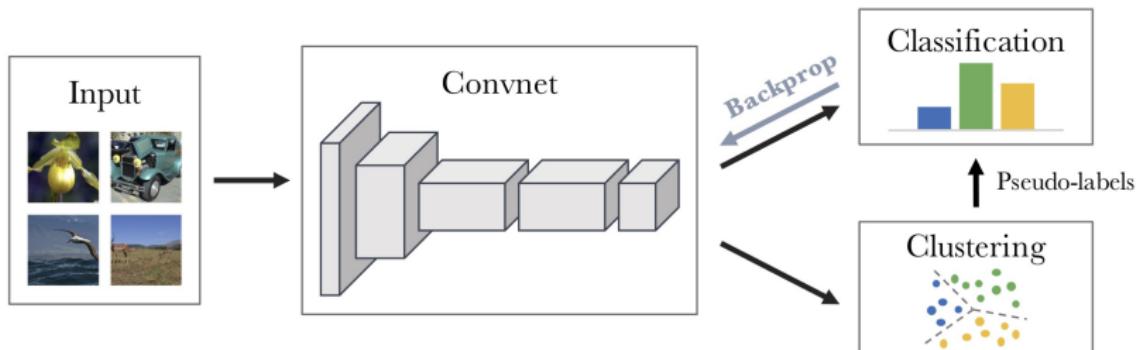
Gidaris S. et al. (2018) Unsupervised representation learning by predicting image rotations.



DeepCluster

Caron M. et al. (2018) Deep clustering for unsupervised learning of visual features.

- Прогоняем через нейросеть данные, чтобы получить эмбеддинги
- Кластеризуем эмбеддинги с помощью K-means
- Обучаем модель предсказывать полученные кластера
- Повторяем эту процедуру по кругу



Contrastive learning

Contrastive learning

Обучение таких представлений объектов, чтобы

- Похожие объекты / объекты одного класса имели похожие представления
- Непохожие объекты / объекты разных классов имели различные представления
- Функция потерь минимизирует расстояние между позитивными парами и максимизирует расстояние между негативными парами
- Можно использовать и в supervised, и в self-supervised режиме

Contrastive loss

$$L = (1 - Y)D^2 + Y \max(0, m - D)^2$$

$Y = 0$ если объекты похожи, $Y = 1$ если объекты непохожи

D - евклидово расстояние между эмбеддингами объектов

m - margin, минимальное расстояние, на которое хотим разнести непохожие объекты

Triplet loss

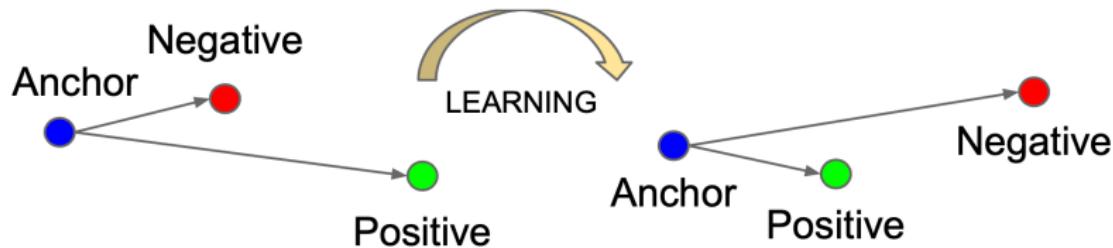
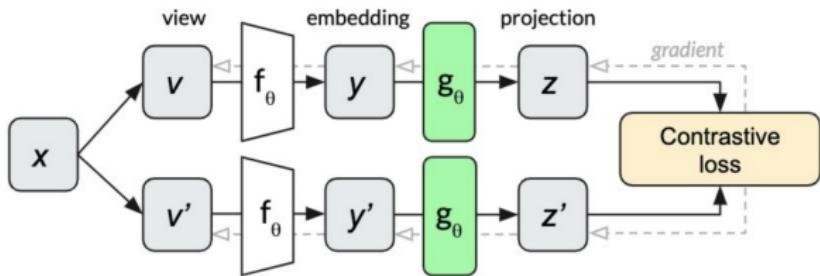


Image credit

$$L(x, x^+, x^-) = \max(0, \|f(x) - f(x^+)\|^2 - \|f(x) - f(x^-)\|^2 + m)$$

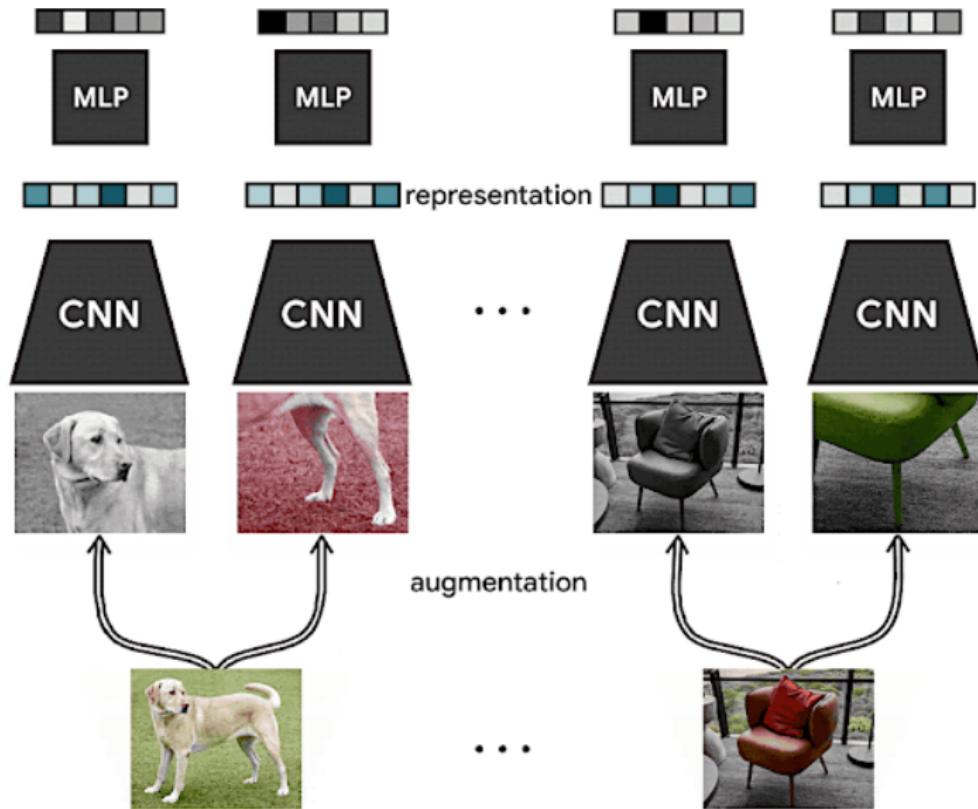
Важно выбирать hard negatives - сложные негативные примеры

Chen T. et al. (2020) A simple framework for contrastive learning of visual representations.



- Аугментация изображений
- Получение эмбеддингов с помощью ResNet
- Добавляется projection head - однослойный MLP
- Contrastive learning - аугментированные версии одного изображения должны быть похожи, разные изображения - непохожи

SimCLR



Аугментации данных



(a) Original



(b) Crop and resize



(c) Crop, resize (and flip)



(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$



(g) Cutout



(h) Gaussian noise



(i) Gaussian blur



(j) Sobel filtering

Лучше всего crop + color distortion

NT-Xent (Normalized Temperature-scaled Cross Entropy) Loss

$$L_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

$$\text{sim}(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\| \|z_j\|}$$

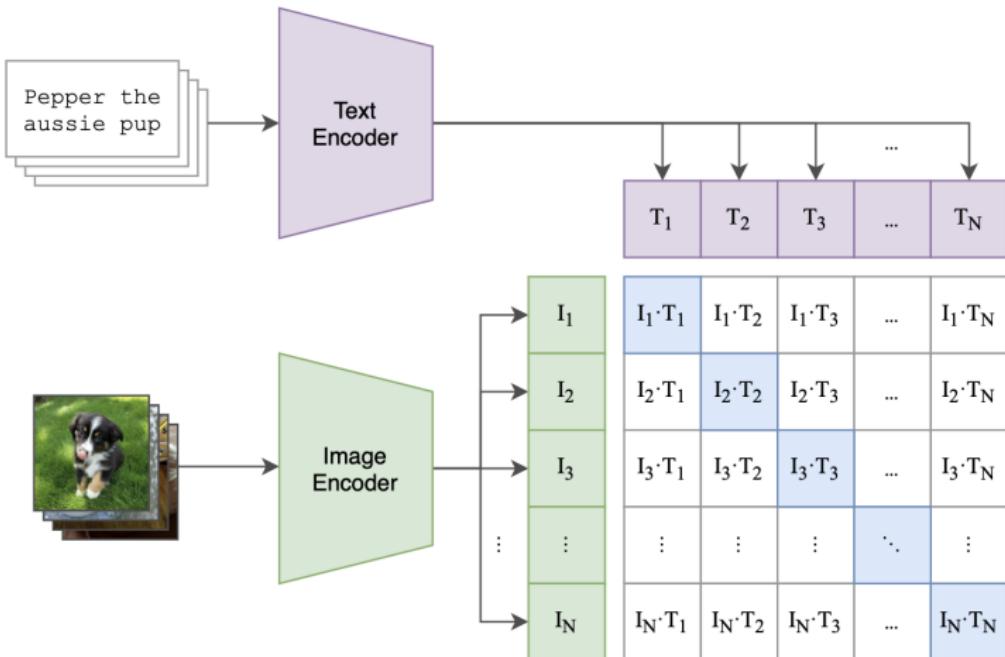
- N примеров в батче
- $2N$ примеров после аугментаций
- Для каждой позитивной пары есть $2(N - 1)$ негативных пар

- Не все аугментации одинаково полезны
- Композиция нескольких аугментаций помогает
- Нужны более сильные аугментации по сравнению с supervised learning
- Projection head важна для хороших результатов
- Нужны больший размер батча и более длительное обучение по сравнению с supervised learning

Radford A. et al. (2021) Learning transferable visual models from natural language supervision.

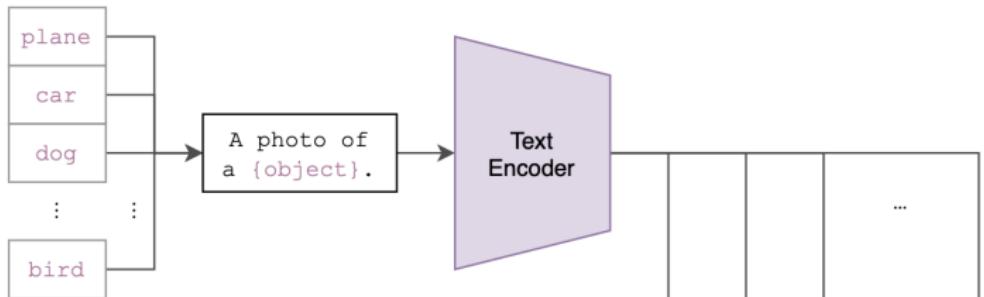
- Мультимодальная модель - текст + изображения
- Преобразует изображения и текст в общее пространство эмбеддингов, в котором их можно сравнивать между собой
- Обучена на 400M пар (картинка, подпись) из интернета

(1) Contrastive pre-training

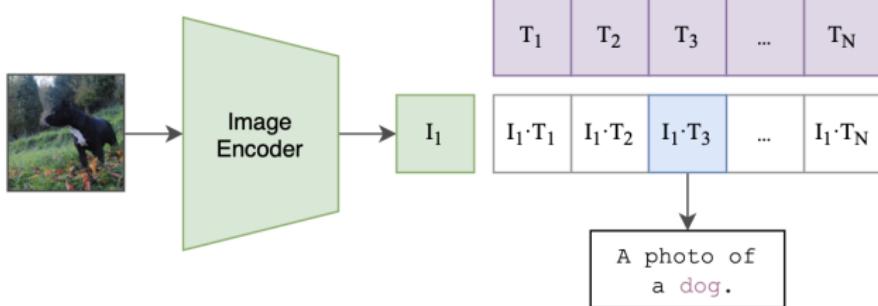


CLIP

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]        - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t              - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

CLIP

Food101

guacamole (90.1%) Ranked 1 out of 101 labels



- a photo of **guacamole**, a type of food.
- a photo of **ceviche**, a type of food.
- a photo of **edamame**, a type of food.
- a photo of **tuna tartare**, a type of food.
- a photo of **hummus**, a type of food.

Youtube-BB

airplane, person (89.0%) Ranked 1 out of 23 labels



- a photo of a **airplane**.
- a photo of a **bird**.
- a photo of a **bear**.
- a photo of a **giraffe**.
- a photo of a **car**.

SUN397

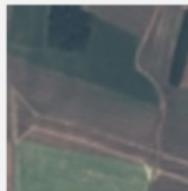
television studio (90.2%) Ranked 1 out of 397 labels



- a photo of a **television studio**.
- a photo of a **podium Indoor**.
- a photo of a **conference room**.
- a photo of a **lecture room**.
- a photo of a **control room**.

EuroSAT

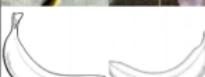
annual crop land (46.5%) Ranked 4 out of 10 labels



- a centered satellite photo of **permanent crop land**.
- a centered satellite photo of **pasture land**.
- a centered satellite photo of **highway or road**.
- a centered satellite photo of **annual crop land**.
- a centered satellite photo of **brushland or shrubland**.

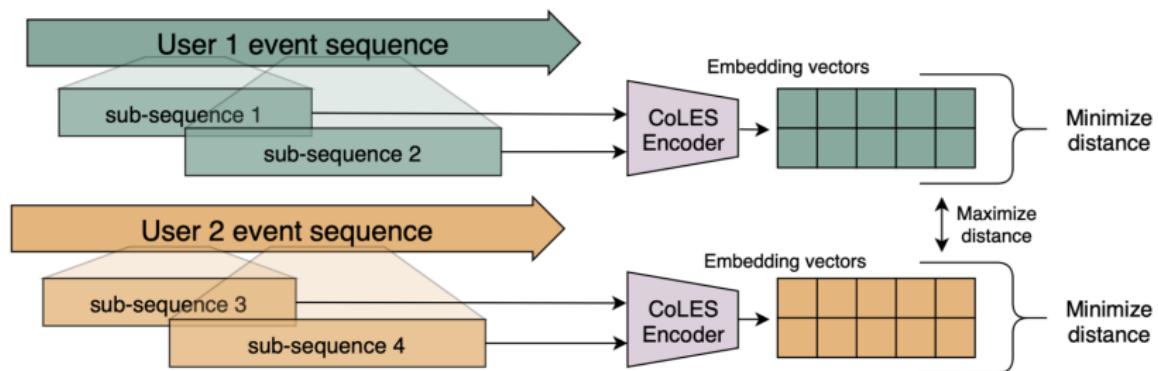
CLIP

Более устойчивая модель, так как не подгонялась под конкретные классы в конкретном датасете

	Dataset Examples						ImageNet ResNet101	Zero-Shot CLIP	Δ Score	
ImageNet										76.2 76.2 0%
ImageNetV2										64.3 70.1 +5.8%
ImageNet-R										37.7 88.9 +51.2%
ObjectNet										32.6 72.3 +39.7%
ImageNet Sketch										25.2 60.2 +35.0%
ImageNet-A										2.7 77.1 +74.4%

Можно использовать contrastive learning не только для изображений и текстов

Babaev, D. et al. (2022) Coles: Contrastive learning for event sequences with self-supervision.



В следующий раз

Гостевая лекция

- Large Language Models (LLM)