Лекция 10. Трансформеры в NLP.

Глубинное обучение

Антон Кленицкий

Recap

Encoder-Decoder architecture

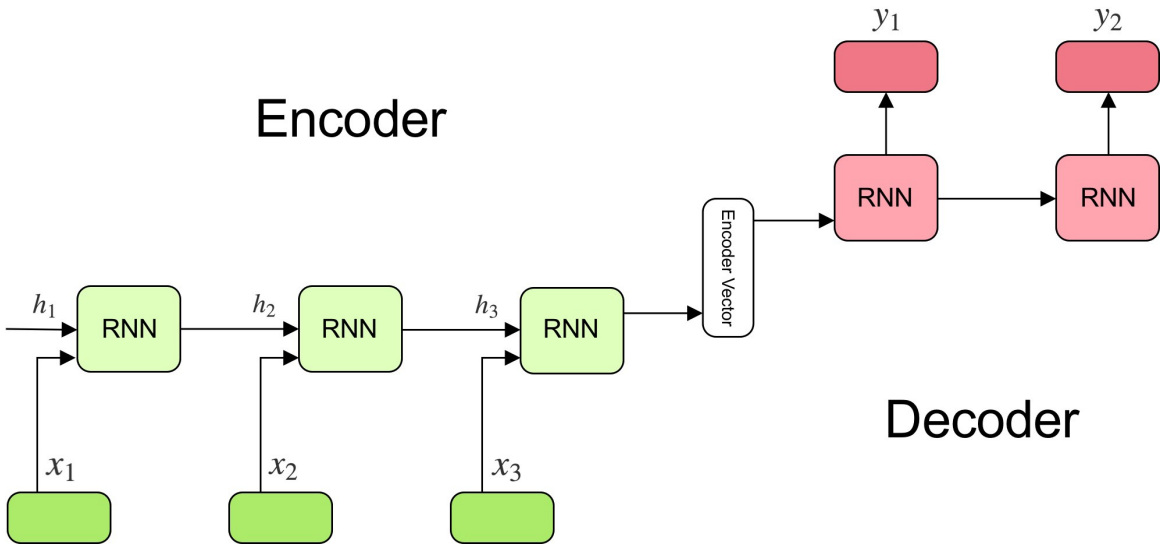


Image credit

Encoder сворачивает входную последовательность в вектор контекста c (как правило, последнее скрытое состояние h_T):

Decoder предсказывает следующий элемент на основе скрытого состояния s_t, предыдущего элемента y_{t-1} и вектора контекста c

Attention

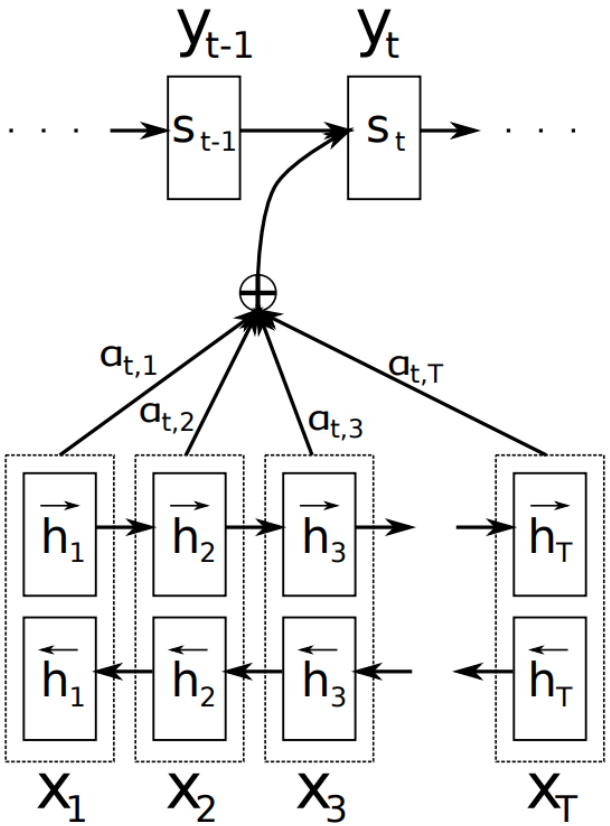


Image credit

Скрытое состояние декодера

$$s_t = f(y_{t-1}, s_{t-1}, c_t)$$

Вектор контекста

$$c_t = \sum_{i=1}^{n} \alpha_{ti} h_i$$

Веса

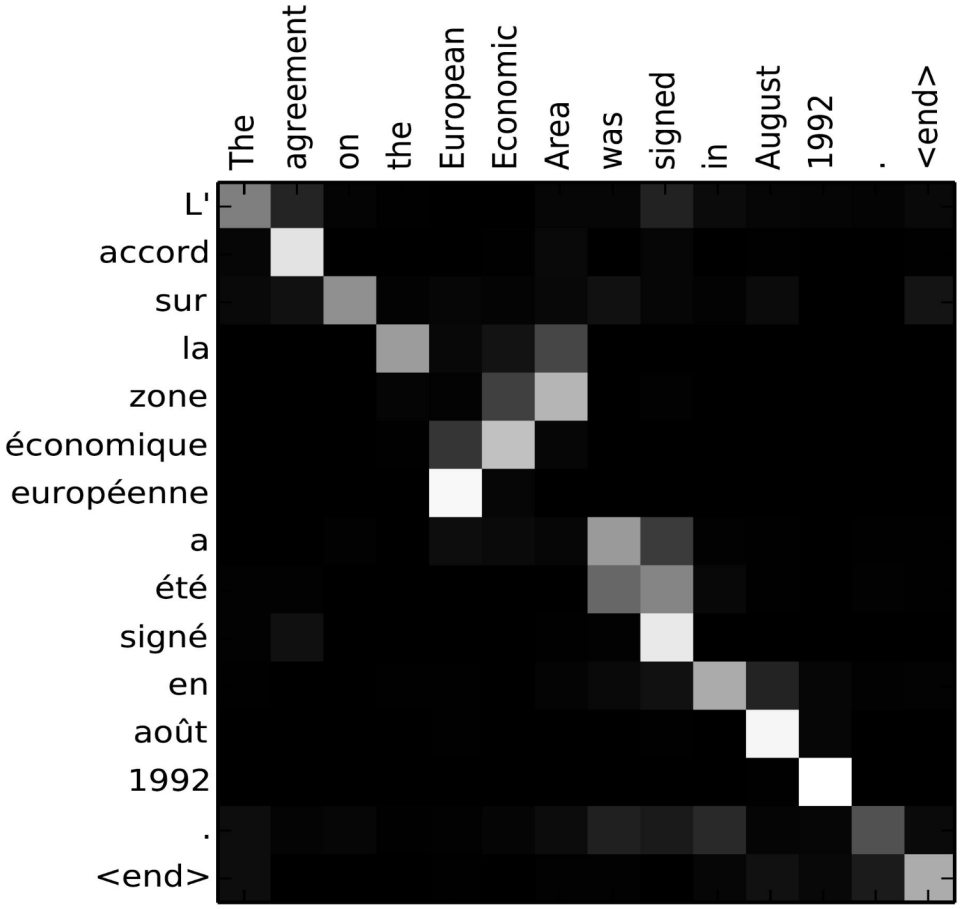
$$\alpha_{ti} = \text{softmax}(\text{score}(s_{t-1}, h_i))$$

Dot-product attention

$$\text{score}(s_{t-1}, h_i) = s_{t-1}^T h_i$$

Интерпретация attention

Матрица весов α_{ti} определяет alignment (выравнивание) элементов входной и выходной последовательностей

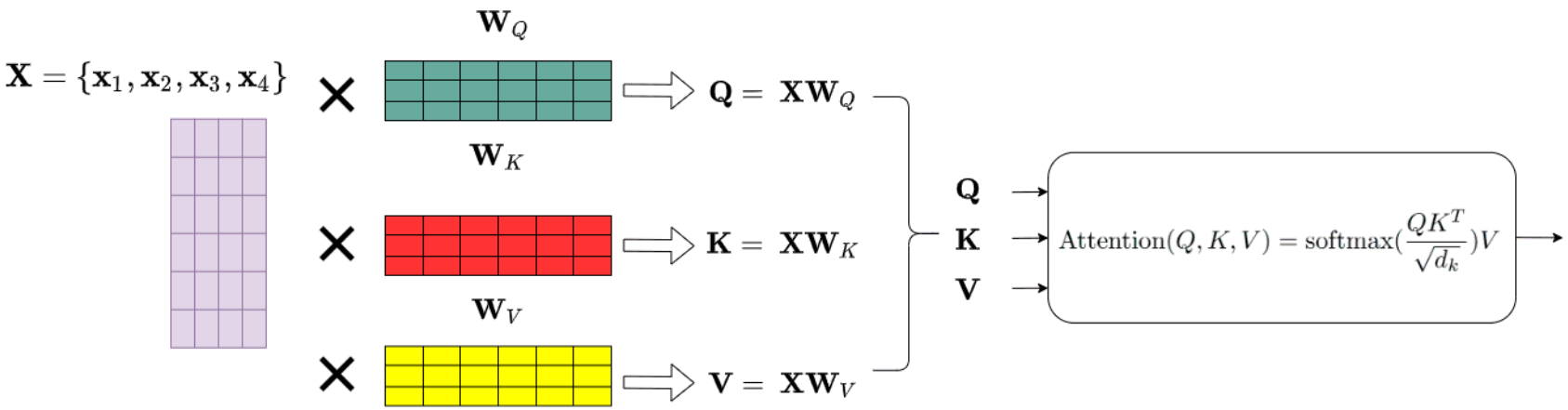


Alignment matrix (Image credit)

Self-attention

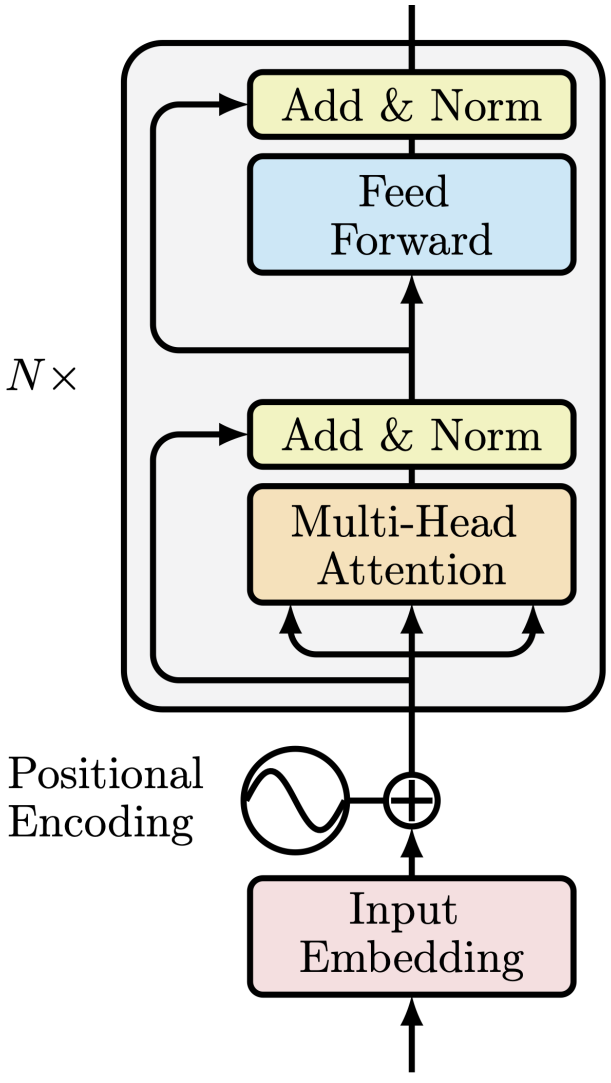
$$Q = XW^Q \qquad query$$
$$K = XW^K \qquad key$$
$$V = XW^V \qquad value$$

$$\text{Attention(Q,K,V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$



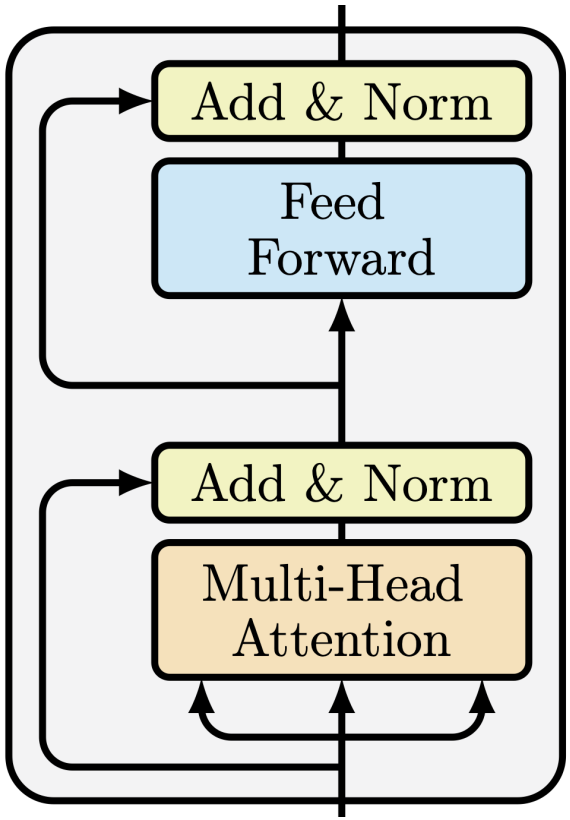
Transformer Encoder

- Embedding layer
- Positional encoding
- N Transformer blocks



Transformer block

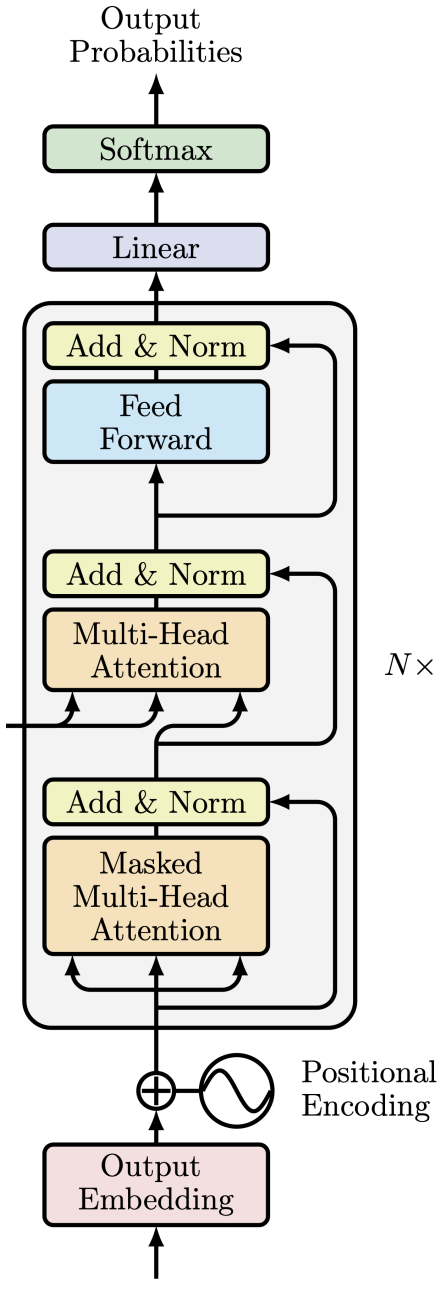
- Multi-head self-attention layer
- Pointwise feed-forward layer
- Residual connections + layer normalization



Transformer Decoder

То же, что и в Encoder, плюс:

- Masked self-attention
- Дополнительный слой с Cross-attention
- Linear + Softmax на выходе



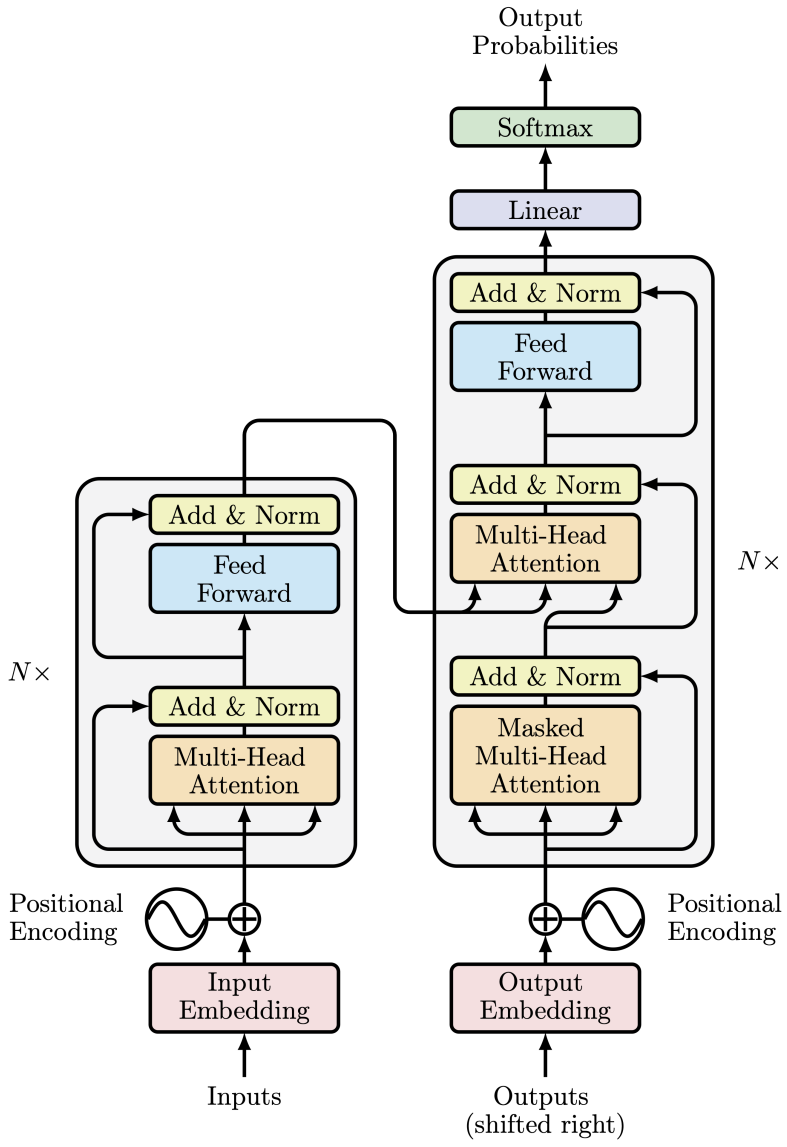
Transformer summary

Pros

- Self-attention может улавливать долгосрочные зависимости
- Параллелизуется, в отличие от RNN

Cons

- Квадратичная зависимость self-attention от длины последовательности



Pretrained Transformers

Transfer learning

Центральная идея современного NLP (и не только :)

- Pretrain - предобучаемся на общей задаче с большим количеством данных и дешевой (лучше совсем бесплатной) разметкой
- Finetune - дообучаем модель под конкретную задачу

NLP's ImageNet moment - 2018

Виды предобученных моделей

Encoder-only

- Классификация
- ELMo, BERT, RoBERTa

Decoder-only

- Text generation
- GPT family

Encoder-Decoder

- Translation, summarization
- BART, T5

ELMo

Peters M. et al. (2018) Deep contextualized word representations.

Embeddings from Language Model (ELMo)

- Контекстно зависимые представления (в отличие от word2vec)
- Представления слов - свертки поверх представлений символов
- 2-layer LSTM в прямом и обратном направлении

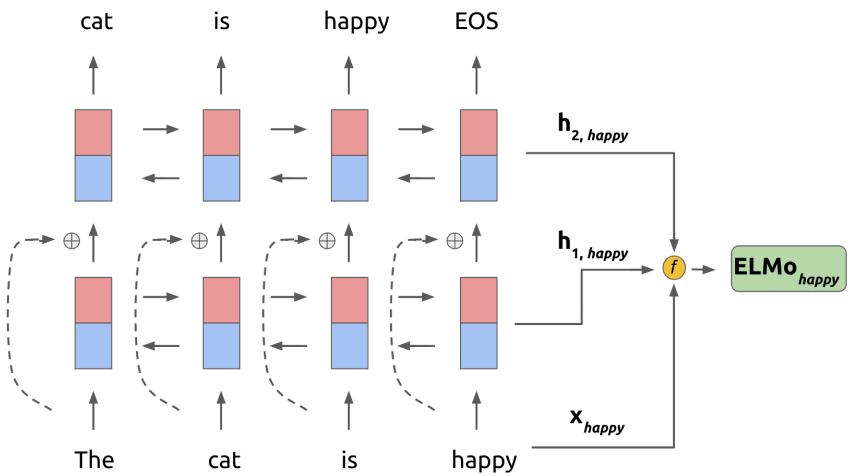


Image credit

BERT

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

Bidirectional Encoder Representations from Transformers

Transformer Encoder

- BERT base L=12, H=768, A=12, 110M параметров
- BERT large L=24, H=1024, A=16, 340M параметров

L - количество слоев

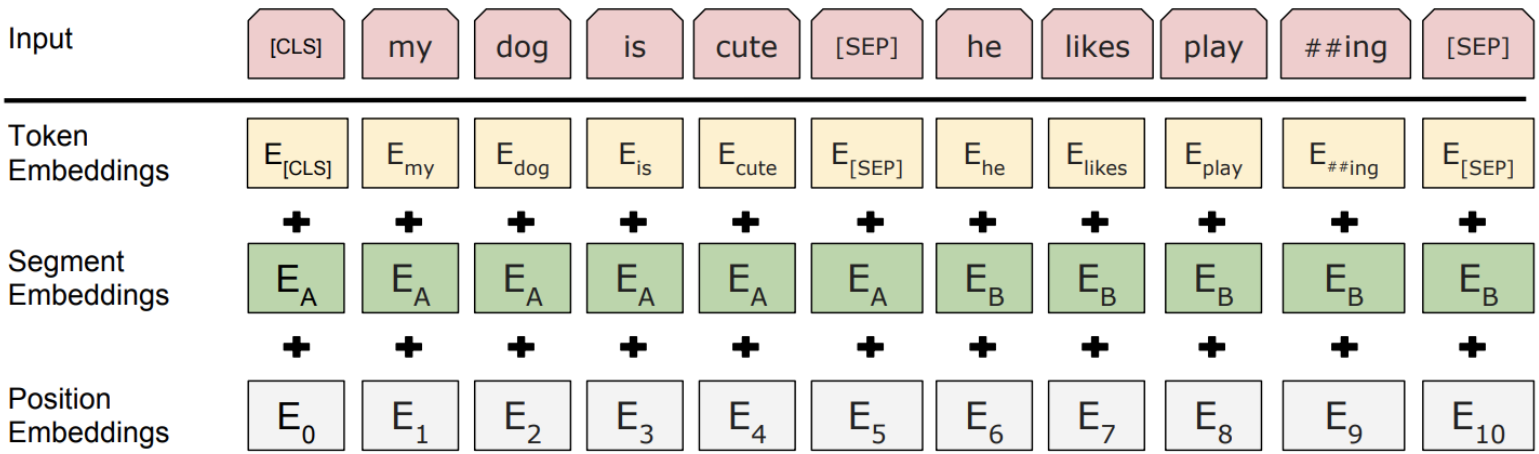
H - размерность скрытого состояния

A - количество голов внимания

BERT

Представление входных данных

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



BERT

Обучение - multitask learning

Masked Language Modeling (MLM)

- Маскируем случайные токены (например, с вероятностью 15%)
- Пытаемся их предсказать по окружающему контексту

+ Next Sentence Prediction

- Предсказываем, является ли два предложения следующими друг за другом

Masked Language Modeling (MLM)

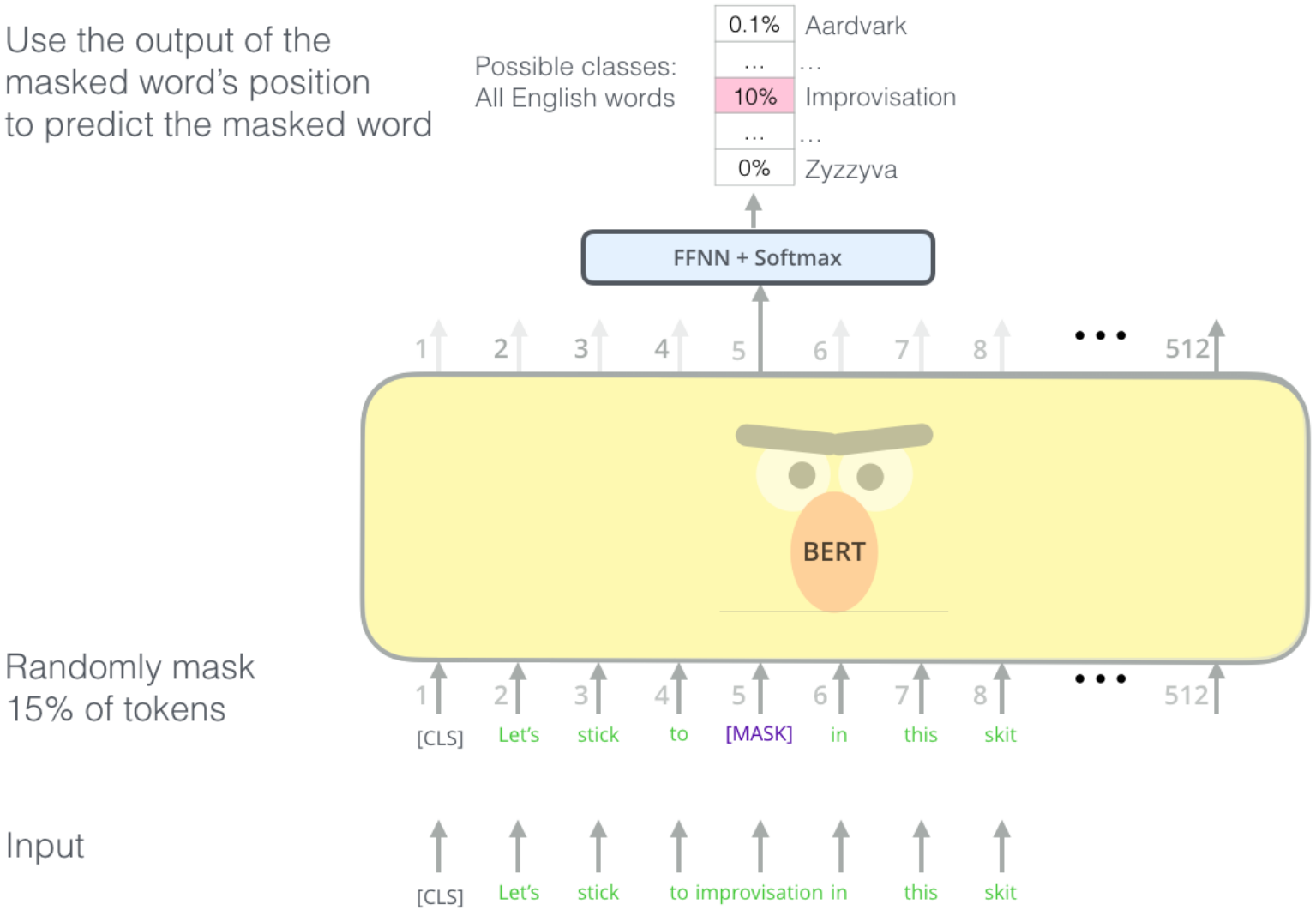
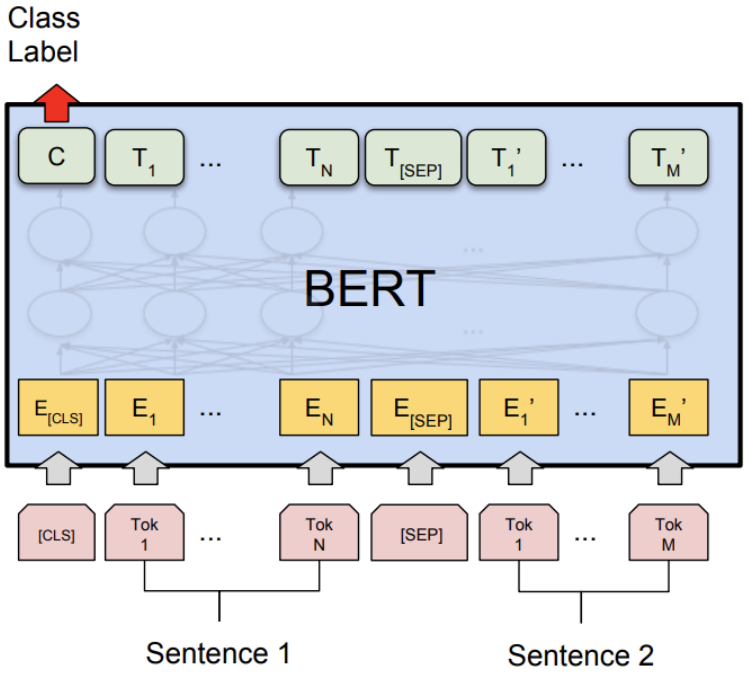
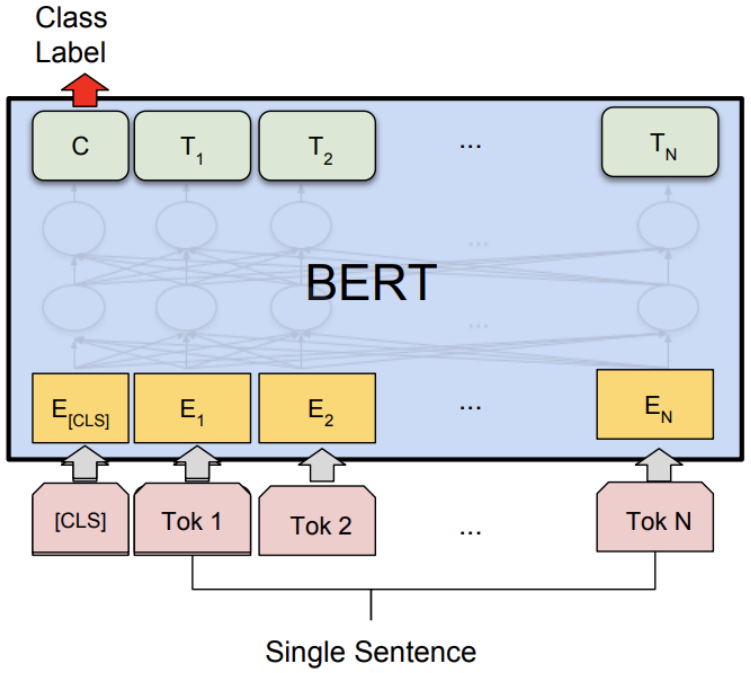


Image credit

Finetuning

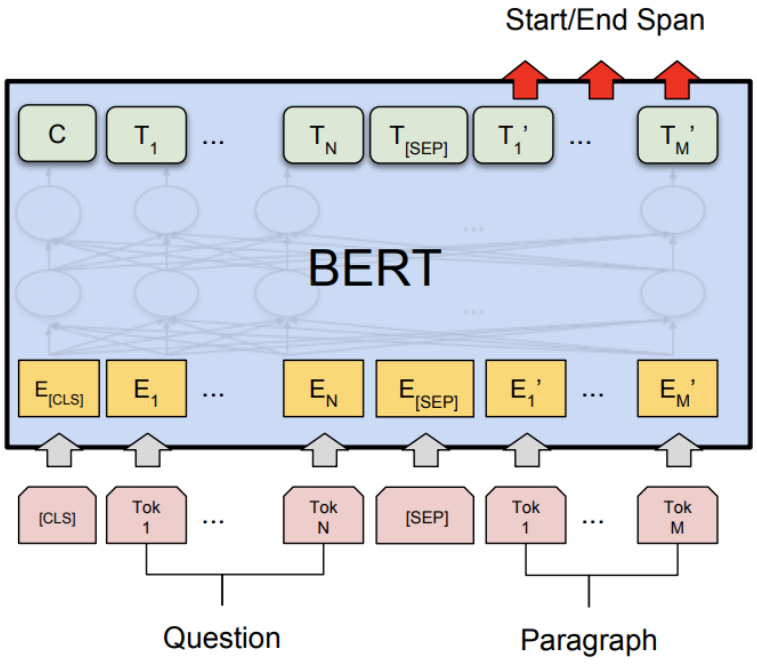


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

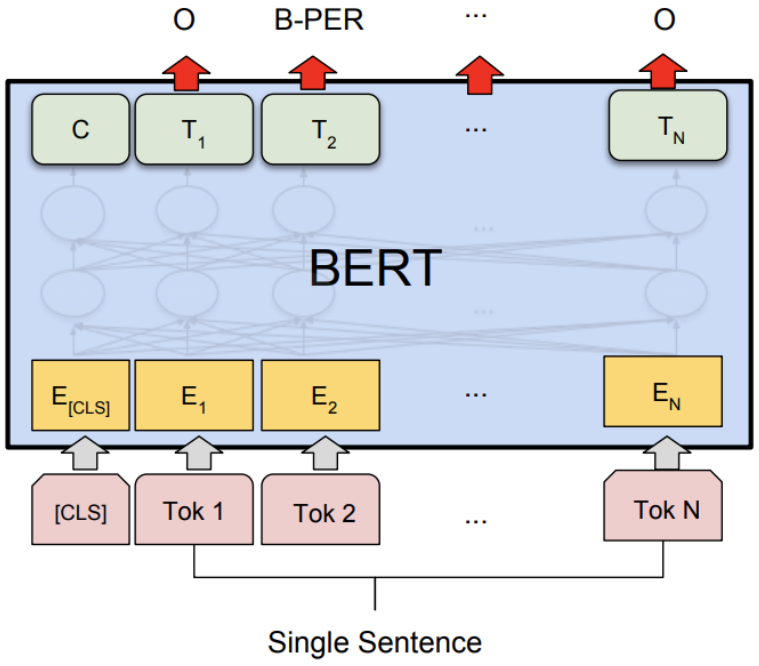


(b) Single Sentence Classification Tasks: SST-2, CoLA

Finetuning



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

RoBERTa

BERT был существенно недообучен

- Подобрали гиперпараметры
- Дольше обучали на большем количестве данных
- Обучали с большим батчем
- Динамическое маскирование - генерация масок «на лету», а не заранее
- Выкинули Next Sentence Prediction

DistilBERT

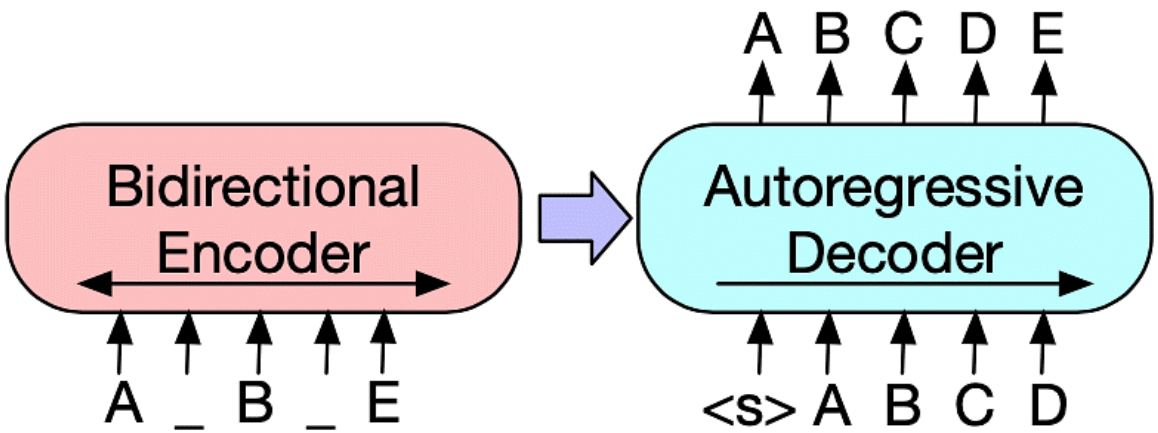
Knowledge distillation - обучение маленькой модели (ученика) воспроизводить поведение большой модели (учителя).

- DistilBERT - 6 слоев вместо 12
- Инициализация слоями обученного BERT'а
- обычный MLM лосс
- + лосс, измеряющий сходство выходов учителя и ученика: $L_{ce} = \sum_i t_i \log s_i$
- + лосс, измеряющий похожесть скрытых состояний учителя и ученика (косинусное расстояние)

BART (Bidirectional and Auto-Regressive Transformers)

Lewis M. et al. (2019) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

- Transformer Encoder-Decoder
- Denoising seq2seq autoencoder, обучается на восстановлении зашумленных данных
- Text Infilling task



BART (Bidirectional and Auto-Regressive Transformers)

Пробовали разные pretrain tasks

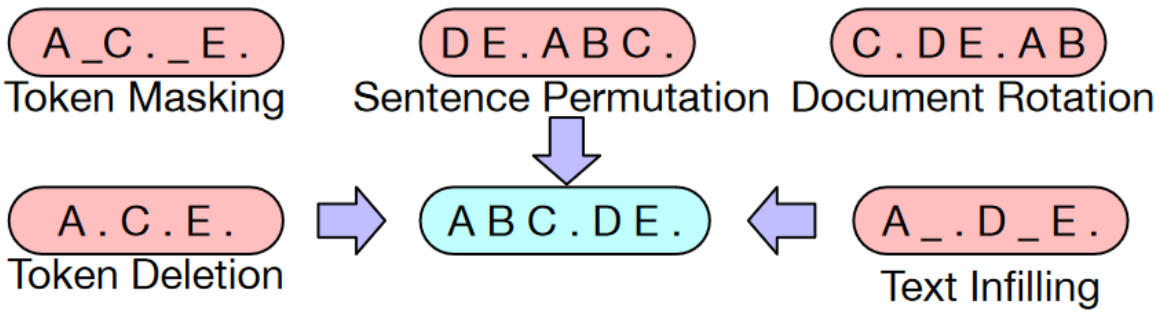
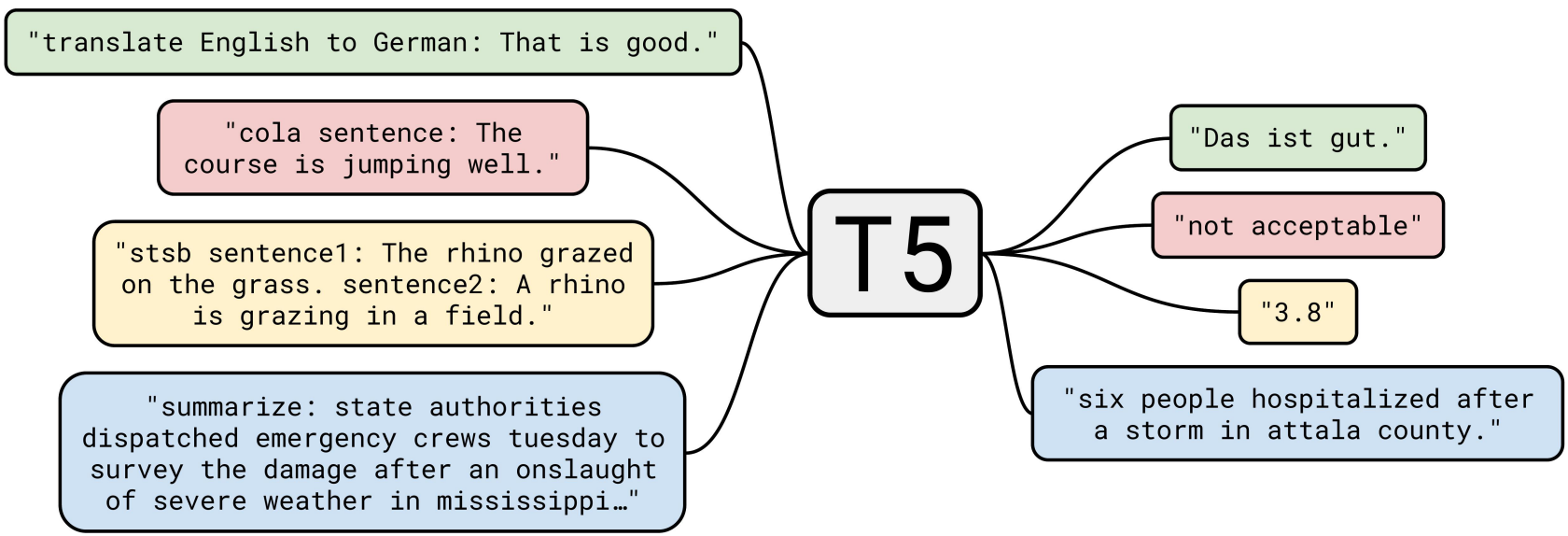


Figure 2: Transformations for noising the input that we experiment with. These transformations can be composed.

T5 (Text-to-Text Transfer Transformer)

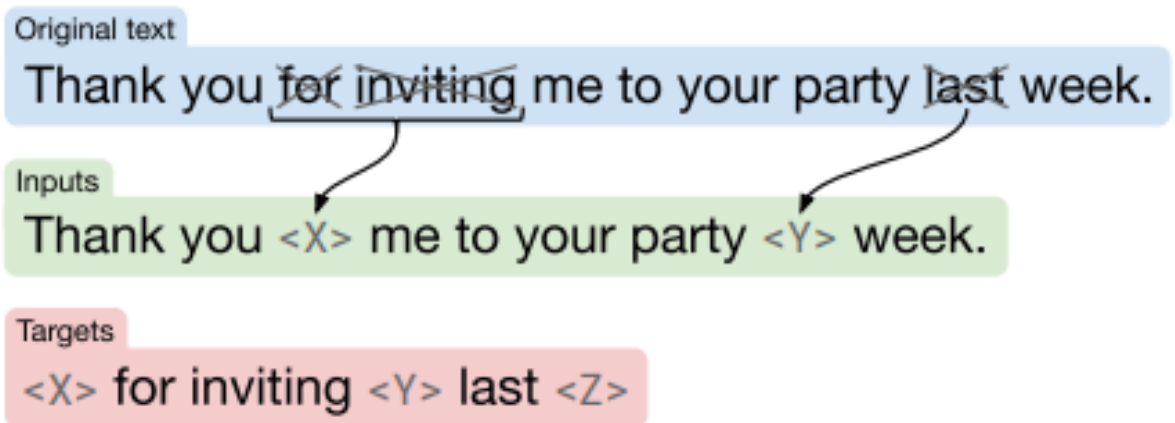
Raffel et al. (2019) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

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



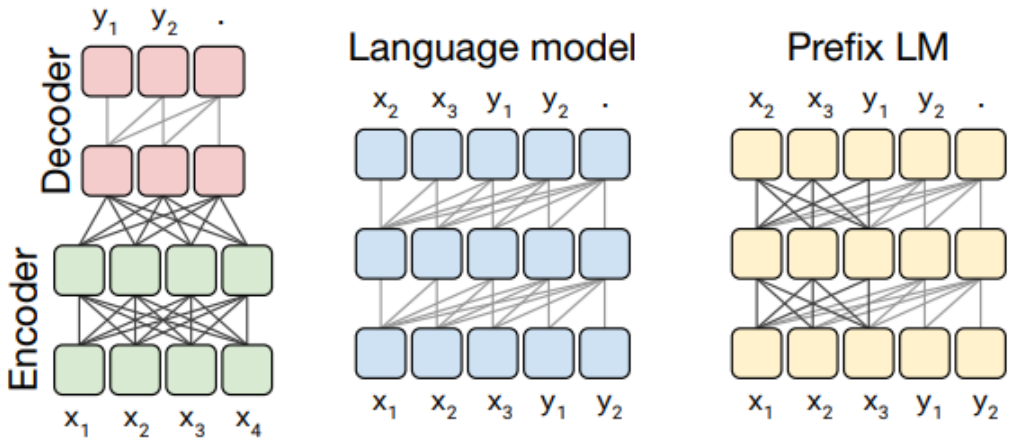
T5 (Text-to-Text Transfer Transformer)

- Encoder-Decoder архитектура
- Размер до 11B параметров
- Новый большой и хороший датасет C4 (Colossal Clean Crawled Corpus)
- Supervised pretrain tasks (translation, summarization,..)
- Unsupervised pretrain task - replace span, заполнение пропусков

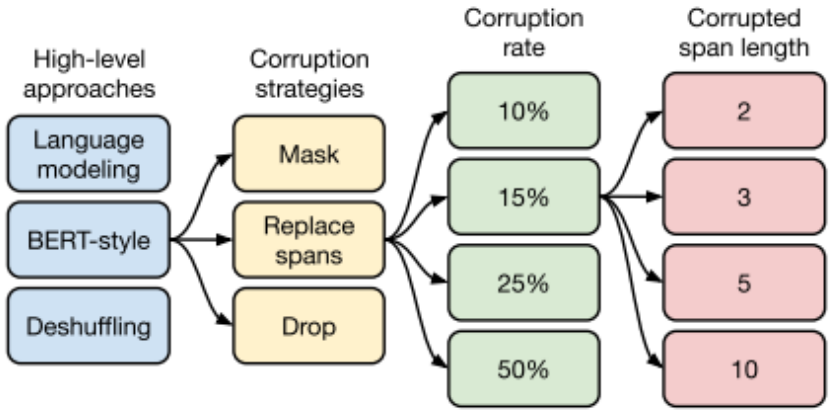


T5 (Text-to-Text Transfer Transformer)

Попробовали разные архитектуры



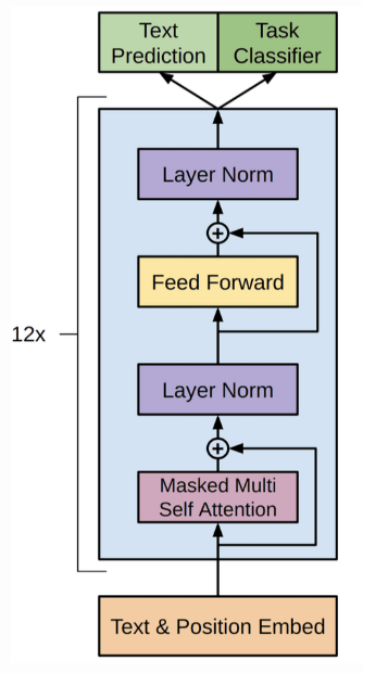
Сравнили разных варианты unsupervised pretrain tasks

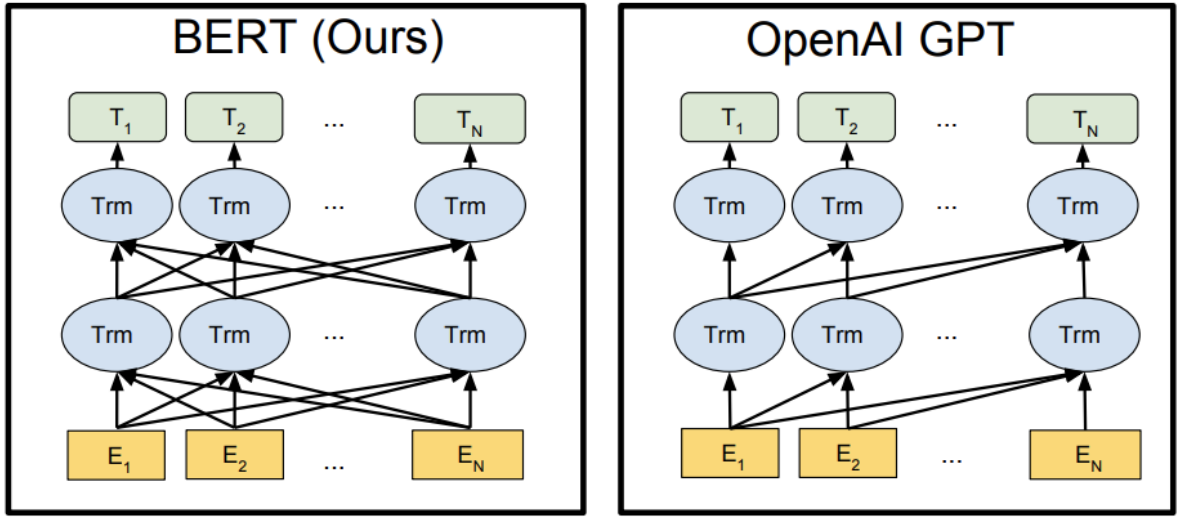


GPT

Generative Pre-trained Transformer (GPT)

- Transformer Decoder
- Просто предсказываем следующий токен





GPT family

Последовательное увеличение количества данных, размера моделей и вычислений

GPT-1

- Radford A. et al. (2018) Improving language understanding by generative pre-training.
- 117M параметров, L=12, H=768, A=12
- BooksCorpus dataset, 4.5 ГБ текста, 7K книг
- Размер контекста 512 токенов

GPT-2

- Radford A. et al. (2019) Language models are unsupervised multitask learners.
- 1.5B параметров, L=48, H=1600
- WebText dataset, 40 ГБ текста, 8M веб-страниц
- Размер контекста 1024 токенов

GPT family

GPT-3

- Brown T. et al. (2020) Language models are few-shot learners.
- 175B параметров, L=96, H=12288, A=96
- 570ГБ текста, 300B токенов (Common Crawl, WebText, ..)
- Размер контекста 2048 токенов

GPT-4

- ???
- Размер контекста 8192 и 32768 токенов
- Мультимодальная модель (работает также с изображениями)

Zero-shot learning

- Использование модели совсем без дообучения!
- Язык - универсальный интерфейс
- Форматируем вход (prompt) так, чтобы была понятна задача
- 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.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

Few-shot learning

Без дообучения, но показываем модели несколько примеров того, как решать задачу

Few-shot

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

```
1  Translate English to French:
2  sea otter => loutre de mer
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese =>
```

task description

examples

prompt

Few-shot learning

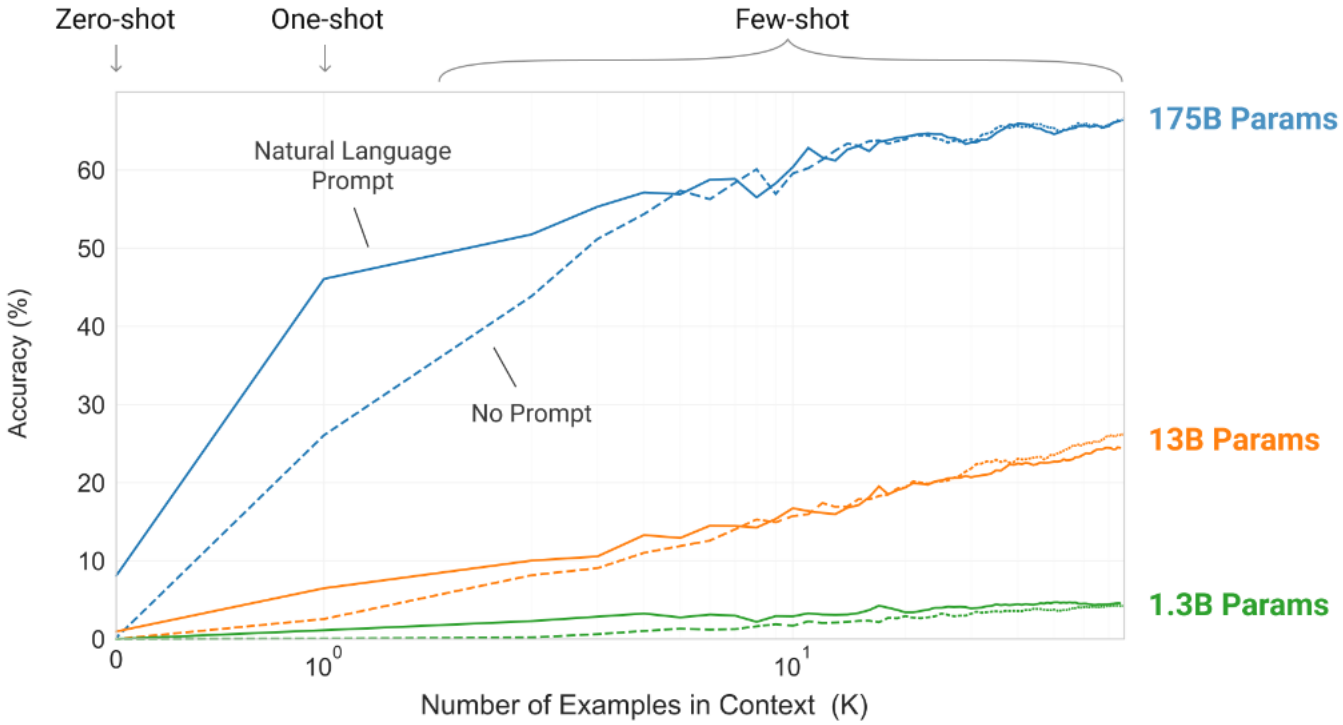


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

The Bitter Lesson

Richard Sutton

> *The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation.*

`http://www.incompleteideas.net/IncIdeas/BitterLesson.html`

ChatGPT

GPT-3 + Alignment of Language Model

Ouyang L. et al. (2022) Training language models to follow instructions with human feedback. [InstructGPT]

- Supervised finetuning на примерах правильного поведения
- RLHF (Reinforcement Learning from Human Feedback) - вишенка на торте

ChatGPT

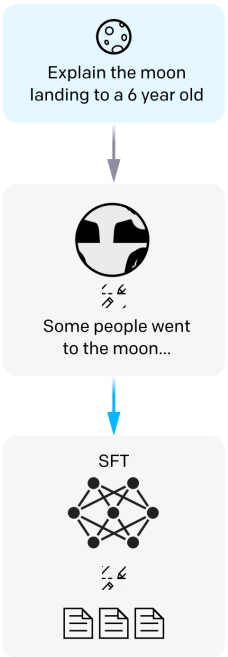
Step 1

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.



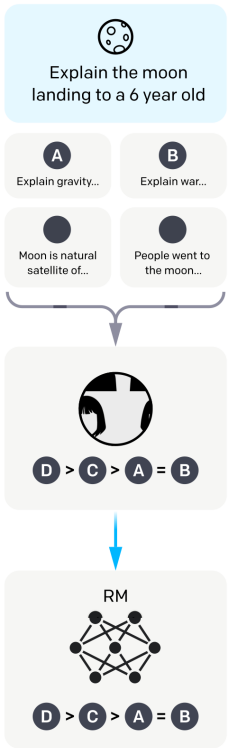
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



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.

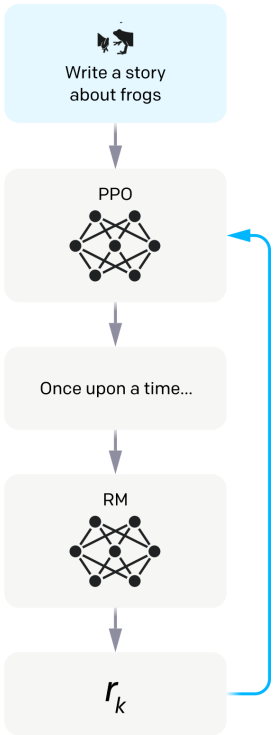
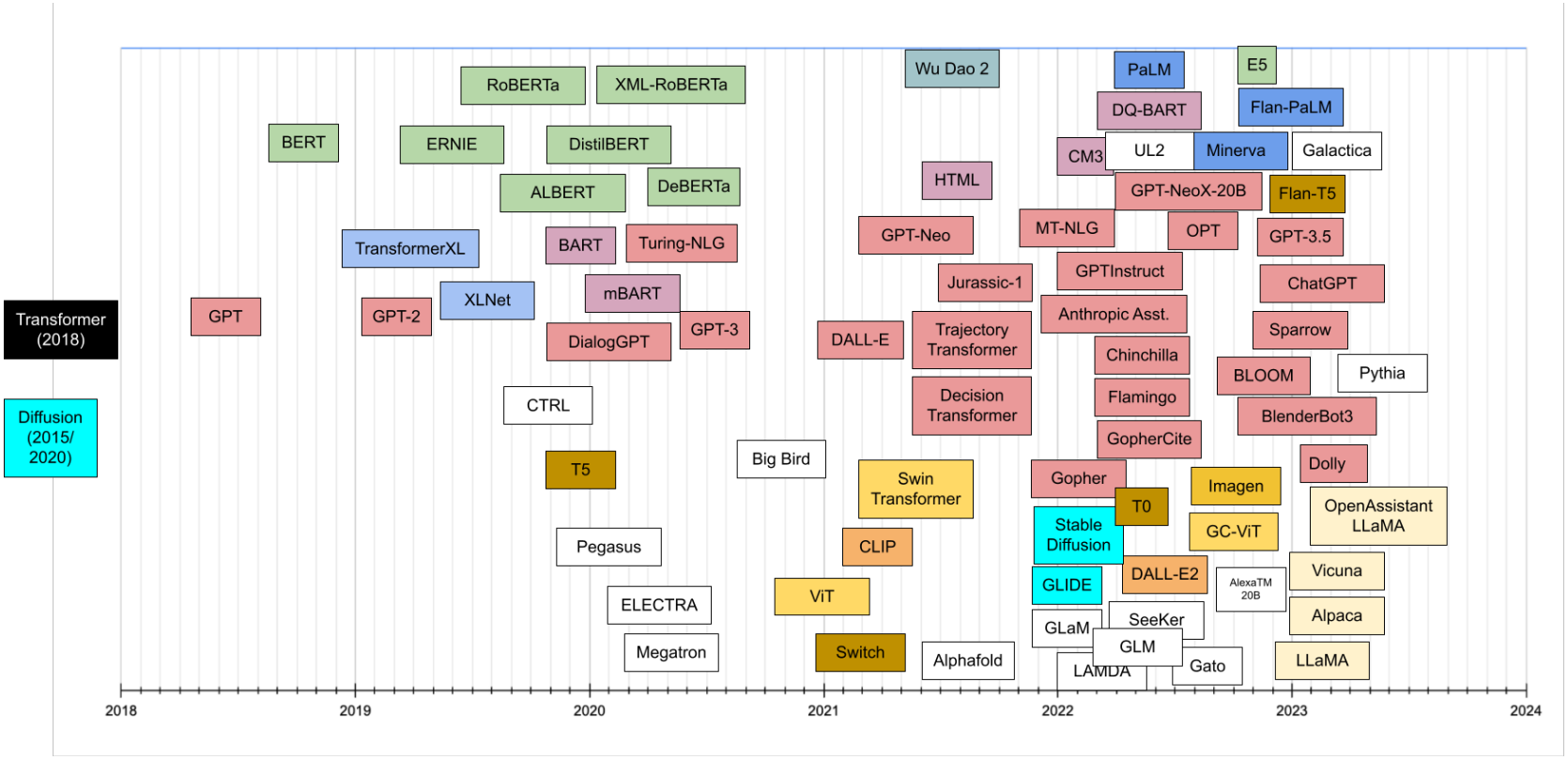


Image credit

Зоопарк трансформеров



Transformer models: an introduction and catalog

Зоопарк трансформеров: большой обзор моделей от BERT до Alpaca

Large Language Models

Large Language Models - stochastic parrots or AGI?

Berglund L. et al. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A"

Bender E. et al. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

В следующий раз

- Vision Transformers
- Self-supervised learning
- Contrastive learning
- CLIP