Transfer Learning в NLP
Лекция 2

Masked Language Modeling, BERT и его вариации

План занятия

1. Новая задача — Masked Language Modeling
2. BERT
3. RoBERTa
4. DistilBERT
5. Выводы

Masked language modeling (MLM) — обобщение классического языкового моделирования

В masked language modeling мы учим модель предсказывать произвольное слово в тексте, а не только следующее.

Masked language modeling (MLM) — обобщение классического языкового моделирования

В masked language modeling мы учим модель предсказывать произвольное слово в тексте, а не только следующее.

Вчера мы посмотрели отличный фильм в кинотеатре.

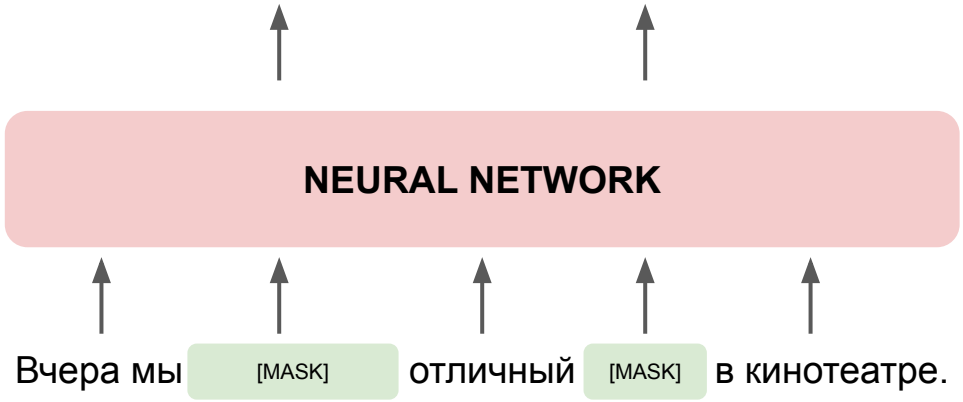
Masked language modeling (MLM) — обобщение классического языкового моделирования

В masked language modeling мы учим модель предсказывать произвольное слово в тексте, а не только следующее.

Вчера мы [MASK] отличный [MASK] в кинотеатре.

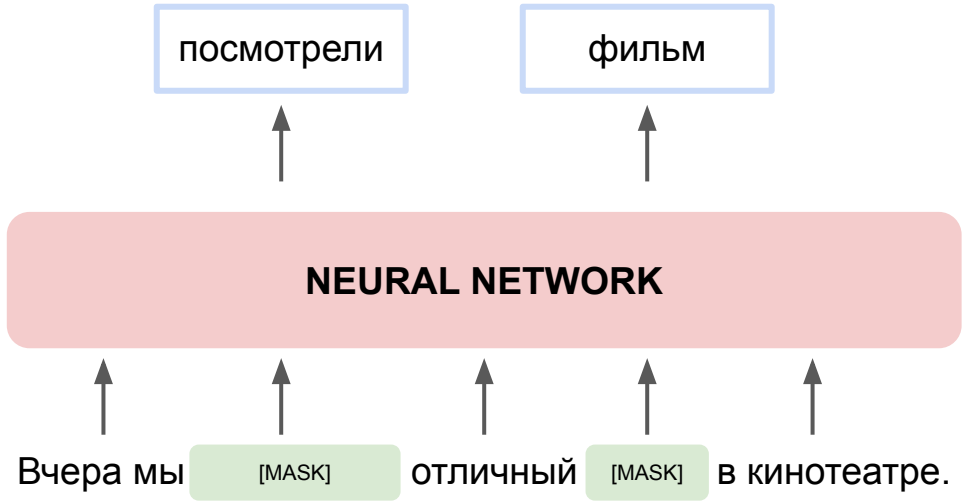
Masked language modeling (MLM) — обобщение классического языкового моделирования

В masked language modeling мы учим модель предсказывать произвольное слово в тексте, а не только следующее.



Masked language modeling (MLM) — обобщение классического языкового моделирования

В masked language modeling мы учим модель предсказывать произвольное слово в тексте, а не только следующее.

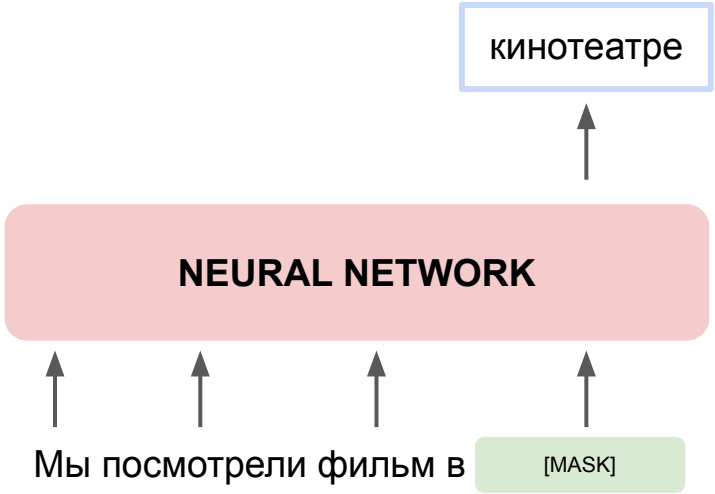


Связь обычного языкового моделирования и маскированного

Как можно представить обычное языковое моделирование через MLM?

Связь обычного языкового моделирования и маскированного

Как можно представить обычное языковое моделирование через MLM?

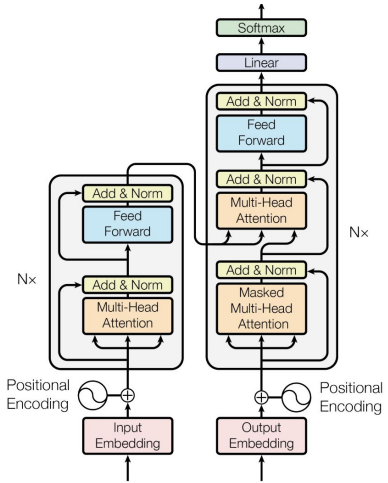


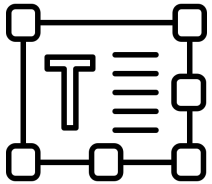
$$= \quad P_\theta(w_i \mid w_{i-1}, \ldots, w_0)$$

Маскируем последний токен в последовательности!

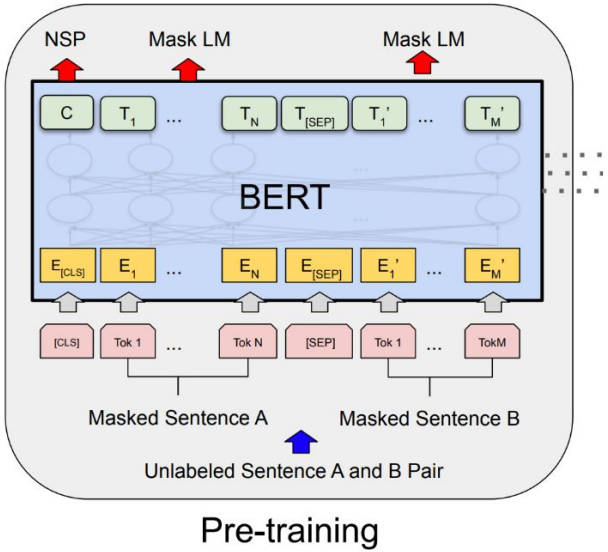
BERT — Bidirectional Encoder Representations from Transformers

Masked Language Modeling





В действительности всё сложнее



Pre-training

2 задачи на pre-training стадии:

- MLM (Masked Language Modeling)
- NSP (Next Sentence Prediction)*

https://arxiv.org/pdf/1810.04805.pdf

Как выбирать маскирование для MLM?

Детали маскирования:

- Случайным образом выбираем 15% токенов из предложения
  - 80% из них заменяем на [MASK]
  - 10% из них заменяем на случайный токен из словаря
  - 10% оставляем исходный токен (и все равно учимся предсказывать его)

Делаем это, чтобы получить хорошие представления для всех токенов, а не только [MASK].

На fine-tuning стадии модель не будет иметь [MASK]!

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI	NER	
			Fine-tune	Fine-tune	Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

Table 8: Ablation over different masking strategies.

https://arxiv.org/pdf/1810.04805.pdf

В BERT целых три вида эмбеддингов

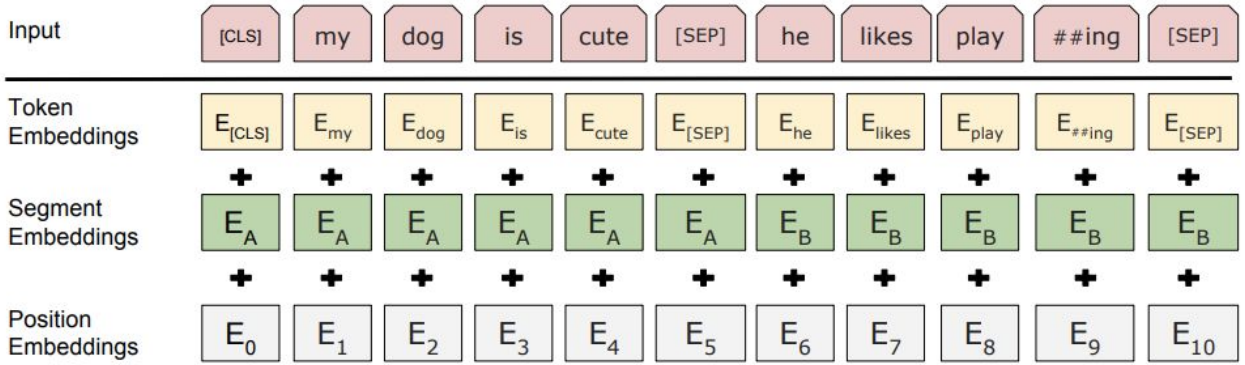
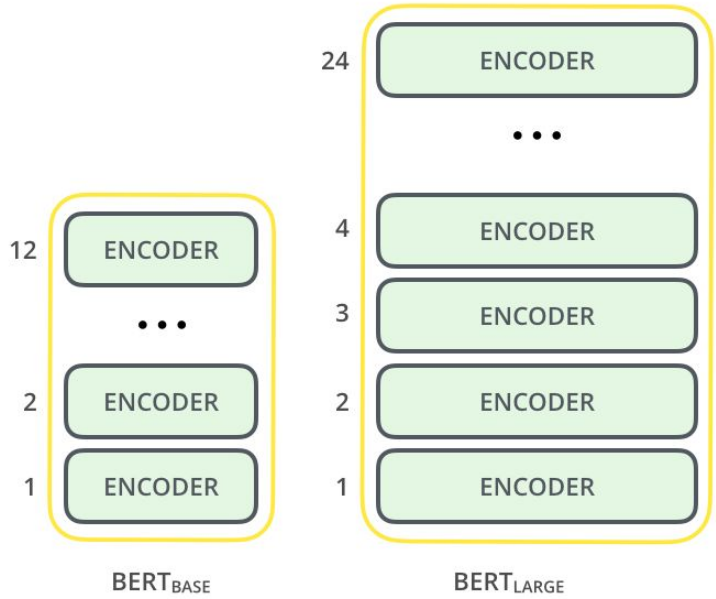


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

https://arxiv.org/pdf/1810.04805.pdf

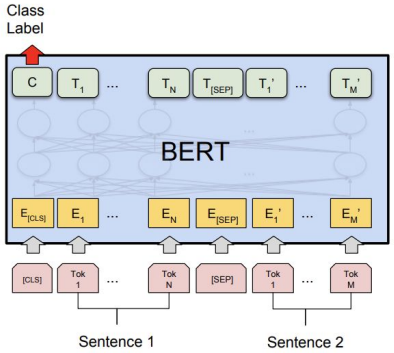
Некоторые детали

- Две конфигурации модели:
  - BERT-base: L=12, H=768, A=12, Total Parameters=110M
  - BERT-large: L=24, H=1024, A=16, Total Parameters=340M
- Данные:
  - BookCorpus (800 млн. слов)
  - English Wikipedia (2,5 млн. слов)
- Максимальная длина последовательности — 512 токенов
- “Pretrain once, finetune many times”

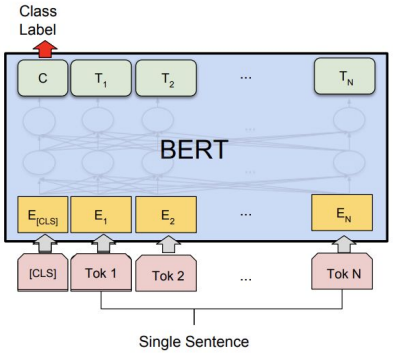


https://iq.opengenus.org/bert-base-vs-bert-large/

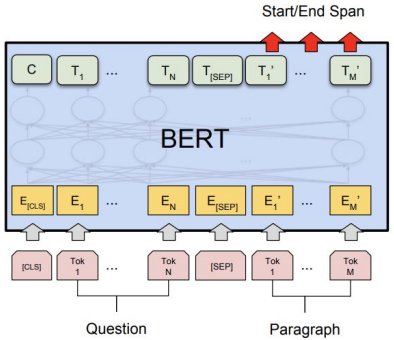
BERT — универсальный фреймворк для решения большинства NLU задач



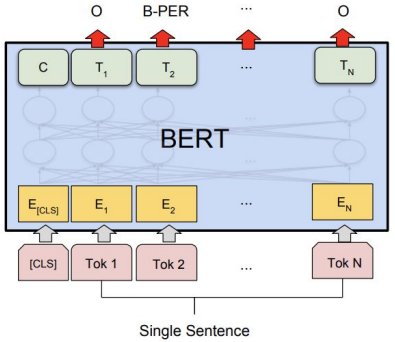
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



(c) Question Answering Tasks: SQuAD v1.1



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

https://arxiv.org/pdf/1810.04805.pdf

Результаты

Более высокое качество на GLUE по сравнению с GPT

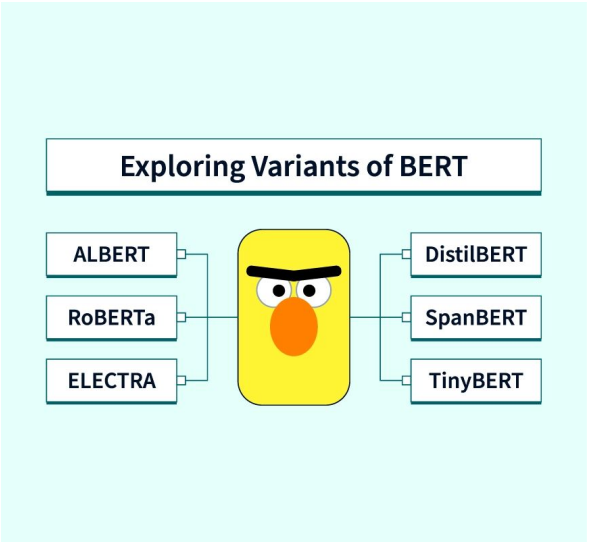
System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

https://arxiv.org/pdf/1810.04805.pdf

Что было после BERT?

Было выпущено множество дополнений и улучшений классического BERT:

1. RoBERTa
2. DeBERTa
3. DistilBERT
4. ALBERT
5. ELECTRA
6. …

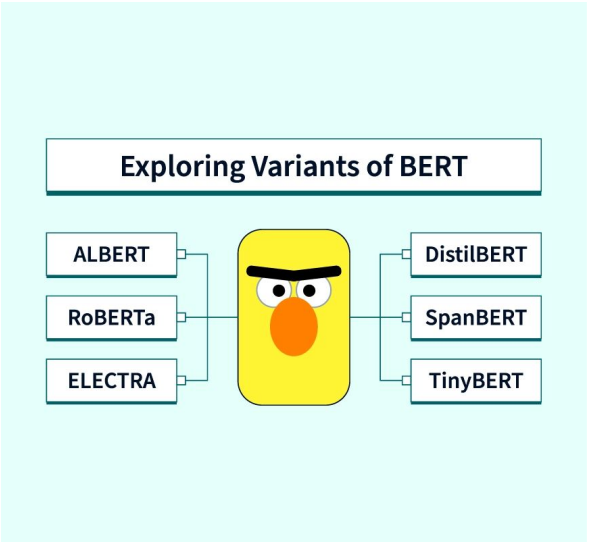


https://www.scaler.com/topics/nlp/bert-variants/

Что было после BERT?

Было выпущено множество дополнений и улучшений классического BERT:

1. **RoBERTa**
2. DeBERTa
3. **DistilBERT**
4. ALBERT
5. ELECTRA
6. …



https://www.scaler.com/topics/nlp/bert-variants/

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Обучим тот же BERT, но немного по-другому:

https://arxiv.org/pdf/1907.11692.pdf

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Обучим тот же BERT, но немного по-другому:

- Динамическое маскирование в MLM

https://arxiv.org/pdf/1907.11692.pdf

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Обучим тот же BERT, но немного по-другому:

- Динамическое маскирование в MLM
- Обучение без NSP лосса

https://arxiv.org/pdf/1907.11692.pdf

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Обучим тот же BERT, но немного по-другому:

- Динамическое маскирование в MLM
- Обучение без NSP лосса
- Больший batch size: 8 тыс. против 256 у BERT

https://arxiv.org/pdf/1907.11692.pdf

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Обучим тот же BERT, но немного по-другому:

- Динамическое маскирование в MLM
- Обучение без NSP лосса
- Больший batch size: 8 тыс. против 256 у BERT
- Больший размер датасета: 160гб против 16гб у BERT

https://arxiv.org/pdf/1907.11692.pdf

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Обучим тот же BERT, но немного по-другому:

- Динамическое маскирование в MLM
- Обучение без NSP лосса
- Больший batch size: 8 тыс. против 256 у BERT
- Больший размер датасета: 160гб против 16гб у BERT
- Более долгое обучение на pre-training

https://arxiv.org/pdf/1907.11692.pdf

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Обучим тот же BERT, но немного по-другому:

- Динамическое маскирование в MLM
- Обучение без NSP лосса
- Больший batch size: 8 тыс. против 256 у BERT
- Больший размер датасета: 160гб против 16гб у BERT
- Более долгое обучение на pre-training
- Byte-level BPE для токенизации любой последовательности без [UNK]

https://arxiv.org/pdf/1907.11692.pdf

Получили новый SOTA подход на GLUE

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-

https://arxiv.org/pdf/1907.11692.pdf

DisilBERT

У моделей из семейства BERT есть небольшая проблема — они требовательны к вычислительным ресурсам

DisilBERT

У моделей из семейства BERT есть небольшая проблема — они требовательны к вычислительным ресурсам

Решение — возьмем модель поменьше и задистиллируем в нее знания из большой модели!

DisilBERT

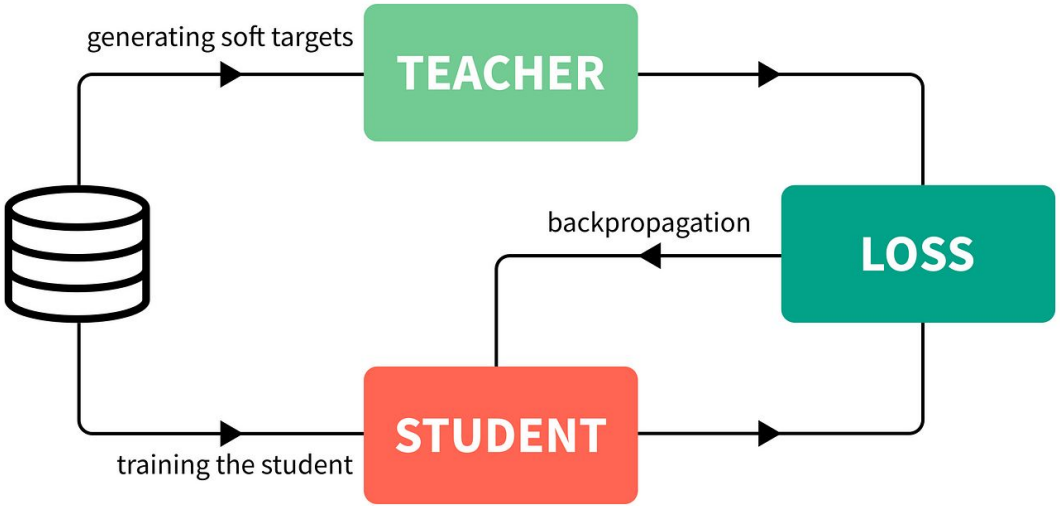
У моделей из семейства BERT есть небольшая проблема — они требовательны к вычислительным ресурсам

Решение — возьмем модель поменьше и задистиллируем в нее знания из большой модели!

Получим DistilBERT:

- на 40% меньше занимаемой памяти
- на 60% быстрее инференс модели
- 97% от качества большой модели

Knowledge distillation — просим модель-ученика повторять за моделью учителя



https://rmoklesur.medium.com/knowledge-distillation-in-deep-learning-keras-implementation-b61261c552db

Дистиллируем не только предсказания, но и скрытые представления

1. Возьмем каждый второй слов из BERT-base
2. Обучим модель предсказывать распределения большой модели, при этом приближая скрытые состояния
3. Получим 97% качества от исходной модели-учителя

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9

https://arxiv.org/pdf/1910.01108.pdf

Итоги занятия

1. Masked Language Modeling и связь с классическим языковым моделированием
2. BERT — двусторонний контекст лучше одностороннего
3. RoBERTa или как достичь лучшего качества без изменения архитектуры
4. DistilBERT: более легкий и эффективный аналог BERT, обученные с помощью подхода knowledge distillation