Lecture

Knowledge-based Systems

Part 5 – (Pre-)Training Language Models

Dr. Mohsen Mesgar

Universität Duisburg-Essen

Exam

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- The exam date is **01.08.2022 16:00 -18:00.**
- Die globale **Anmeldephase** läuft vom **02.05.2022** bis **13.05.2022**
- **Where?** I'll update you

Recall …

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **What is (artificial) intelligence?** The ability to acquire and apply knowledge and skills to achieve complex goals.

- **Symbolic**: Knowledge is encoded by symbols that refer to the knowledge.
- **Connectionist**: Knowledge is **embedded** in parameters of a model.
  - **Pretrained language models (LMs)**
    - Unidirectional
    - Bidirectional
  - **LMs as knowledge bases**
    - factual knowledge
    - linguistic knowledge
    - word sense knowledge

Example

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)\ldots p(w_n|w_1,w_2,..,w_{n-1})$$
$$= \prod_{i=1}^{n} p(w_i|w_1,\ldots,w_{i-1})$$

S = Where are we going

Previous words (Context)

Word being predicted

P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

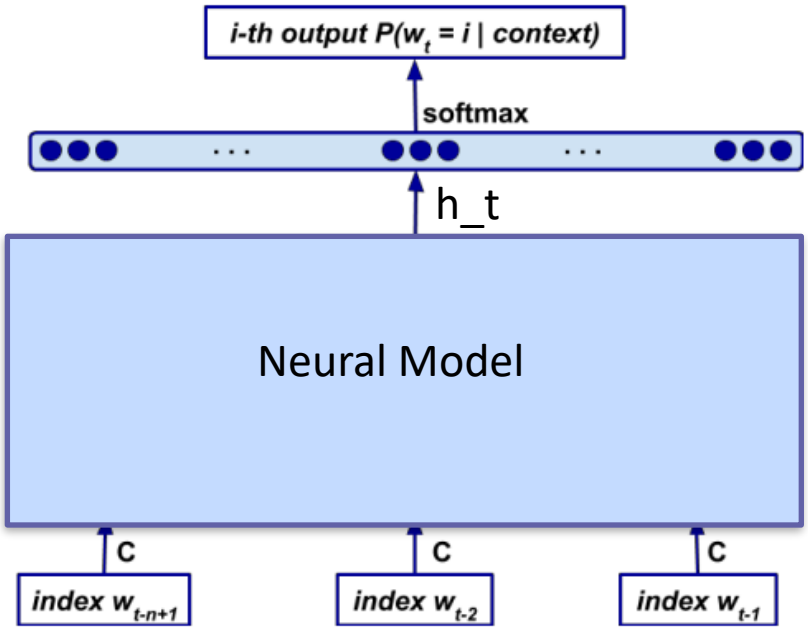
https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

How to get the probability?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- There are different ways to define the probability function
  - p(w_t|w_(t_1),…w_1)
- State-of-the-art LMs use deep neural models and softmax to estimate the probability



Any other open questions?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

In this lecture, you learn about …

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **How do neural language models learn knowledge?**
  - Autoregressive methods
  - Autoencoding methods

In this lecture, you learn about …

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **How do neural language models learn knowledge?**
  - **Autoregressive methods**
  - Autoencoding methods

Autoregressive

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Train LMs on the classic language modeling objective:
  - guess the next token having read all the previous ones.
- Some example Pre-trained LMs using autoregressive method
  - ELMo (2018)
  - GPT (2018)
  - GPT-2 (2019)
  - Transformer-XL (2019)
  - XLNet (2020)
  - GPT-3 (2020)
  - PaLM (2022)
  - OPT (May 2022)

Today

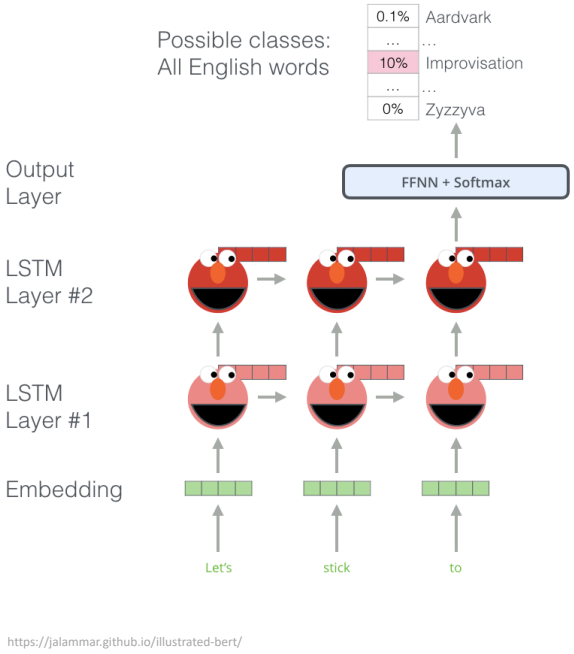
ELMo

ELMo

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Stands for Embeddings from Language Model (Peters, et al, 2018)
- Idea: predict the next word in a sequence of word



https://jalammar.github.io/illustrated-bert/

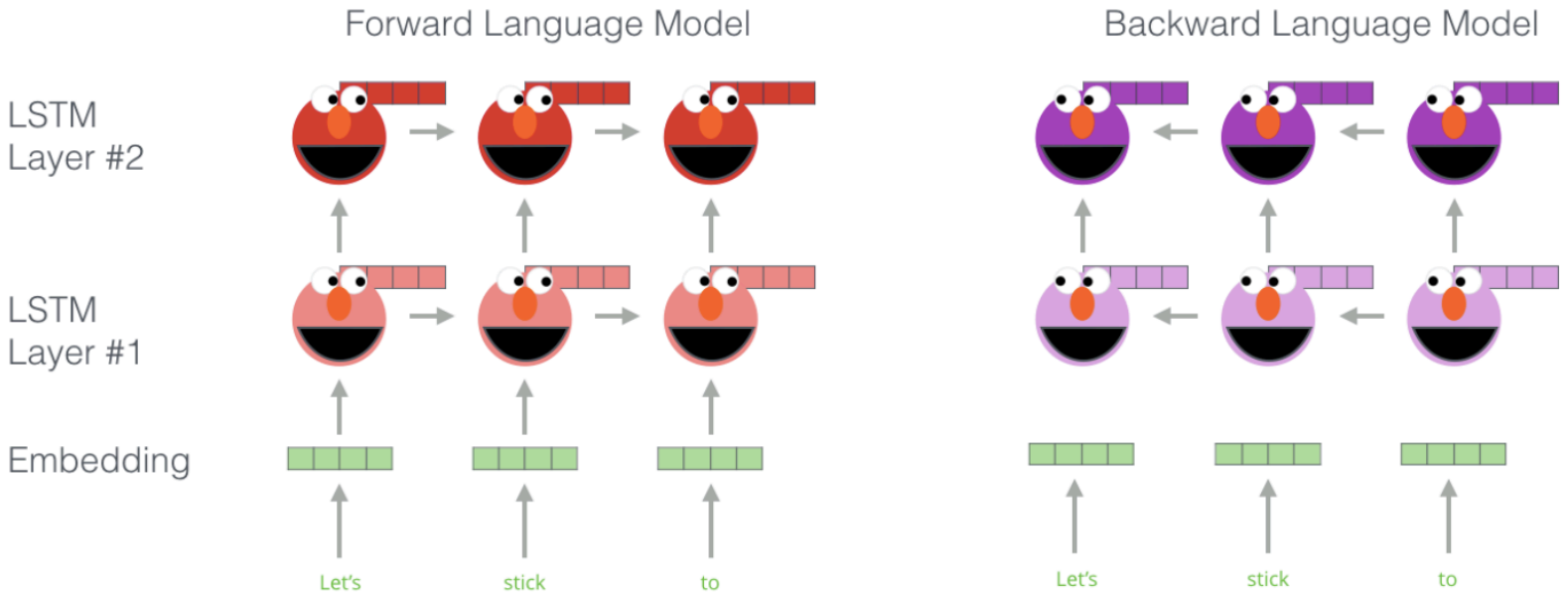
- Data: 1B Word Benchmark

ELMo

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Idea: predict the target word using its left and right context



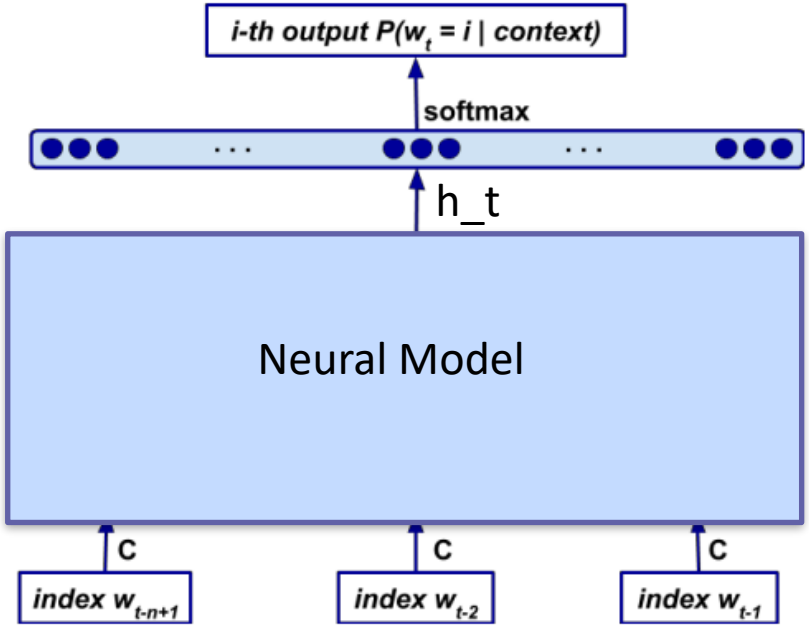
https://jalammar.github.io/illustrated-bert/

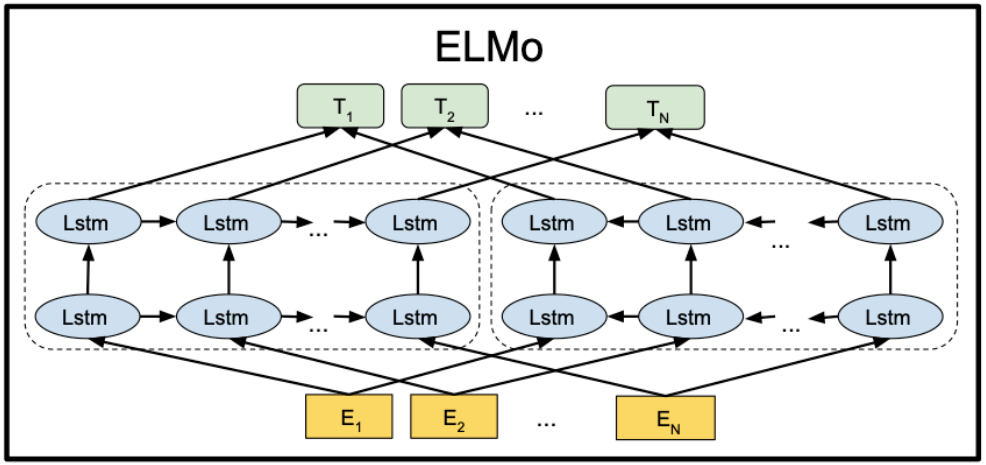
ELMo

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

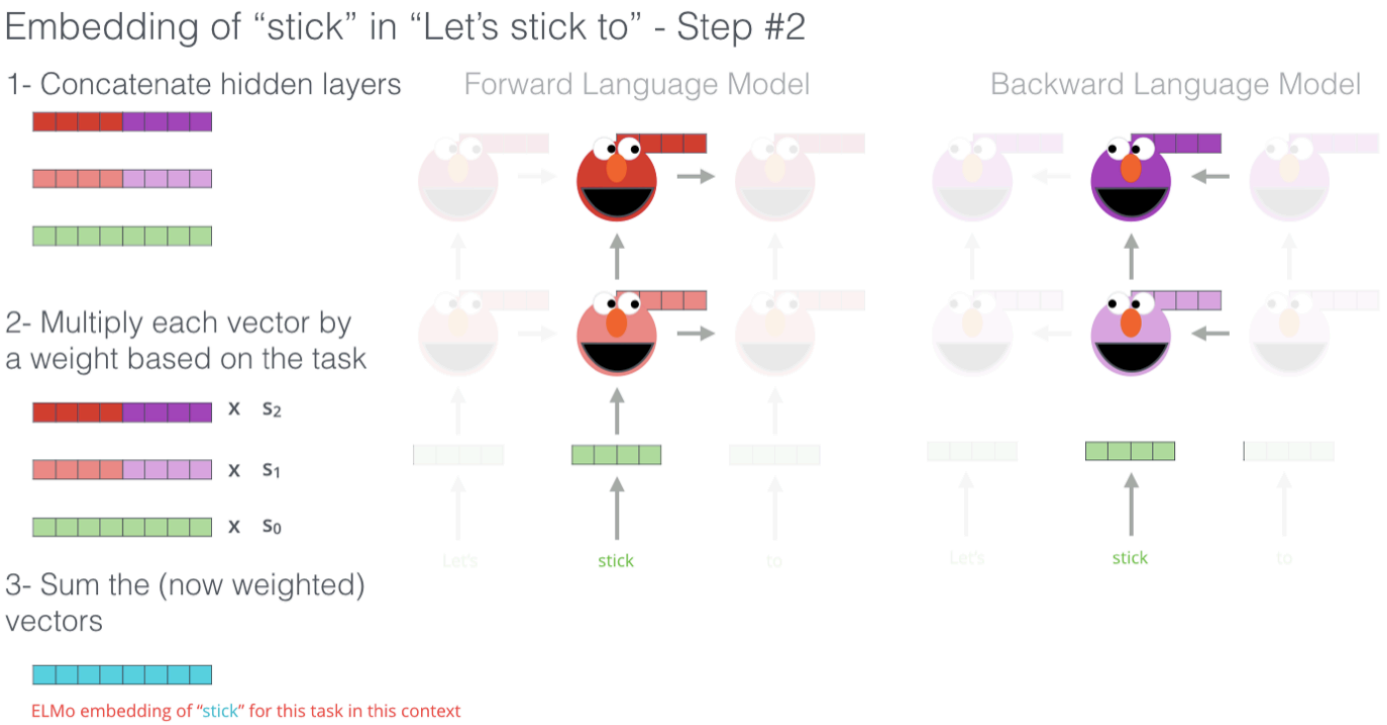
- Idea: predict the target word using its left and right context





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

- Each word can be represented by a vector



https://jalammar.github.io/illustrated-bert/

ELMo

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Pre-trained ELMo Models

Model	Link (Weights/Options File)	# Parameters (Millions)	LSTM Hidden Size/Output size
Small	weights | options	13.6	1024/128
Medium	weights | options	28.0	2048/256
Original	weights | options	93.6	4096/512
Original (5.5B)	weights | options	93.6	4096/512

https://allenai.org/allennlp/software/elmo

- All models except for the 5.5B model were trained on the **1 Billion Word Benchmark,** approximately **800M tokens of news crawl data**.
- The ELMo 5.5B model was trained on a dataset of **5.5B tokens consisting of Wikipedia (1.9B) and all of the monolingual news crawl data** from WMT 2008-2012 (3.6B).
- The 5.5B model has slightly higher performance than the original ELMo model, so they recommend it as a default model.

Today

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

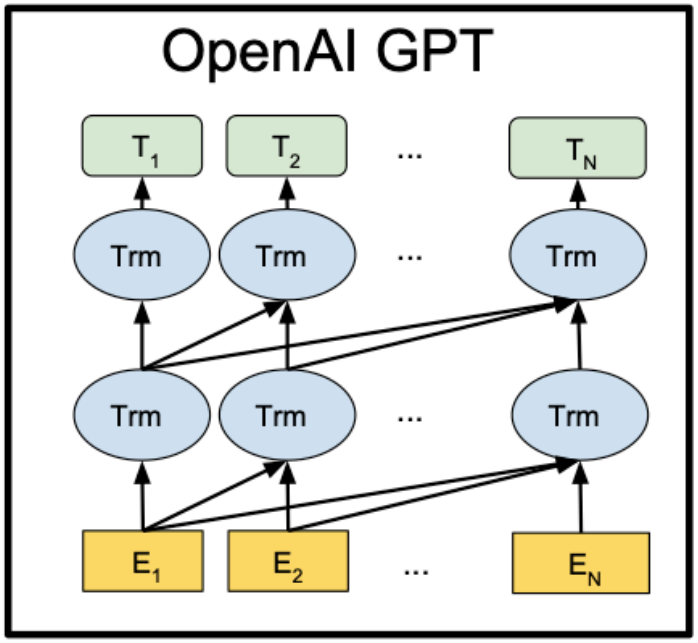
GPT

GPT (a.k.a OpenAI GPT)

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Stands for **Generative Pre-training Transformer**



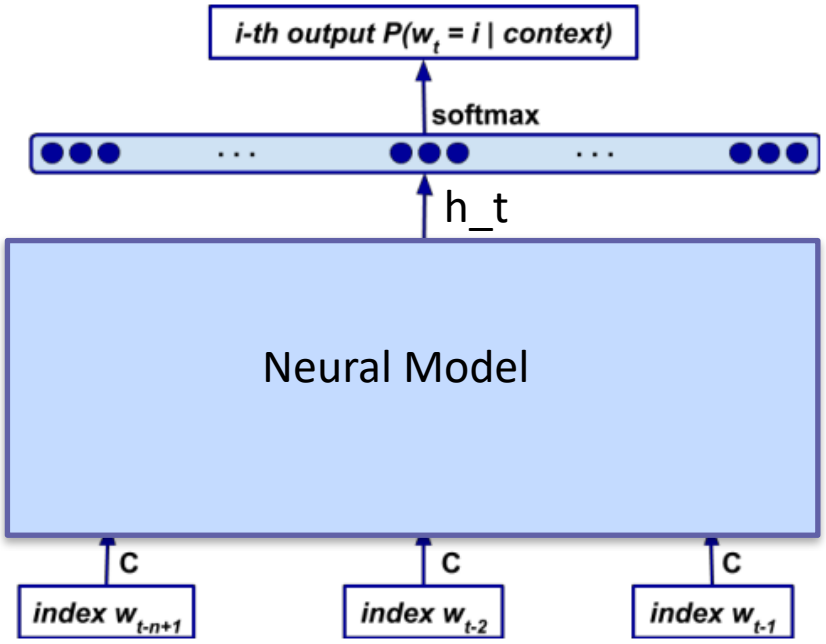
https://arxiv.org/pdf/1810.04805.pdf

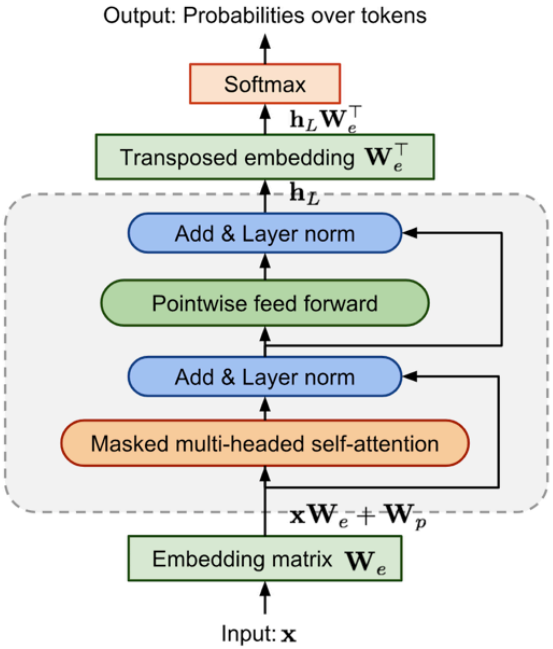
GPT (a.k.a OpenAI GPT)

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- It uses transformers instead of LSTMs





Transformer Block
Repeat x L=12

$$\mathbf{h}_\ell = \text{transformer\_block}(\mathbf{h}_{\ell-1})$$
$$\ell = 1, \ldots, L$$

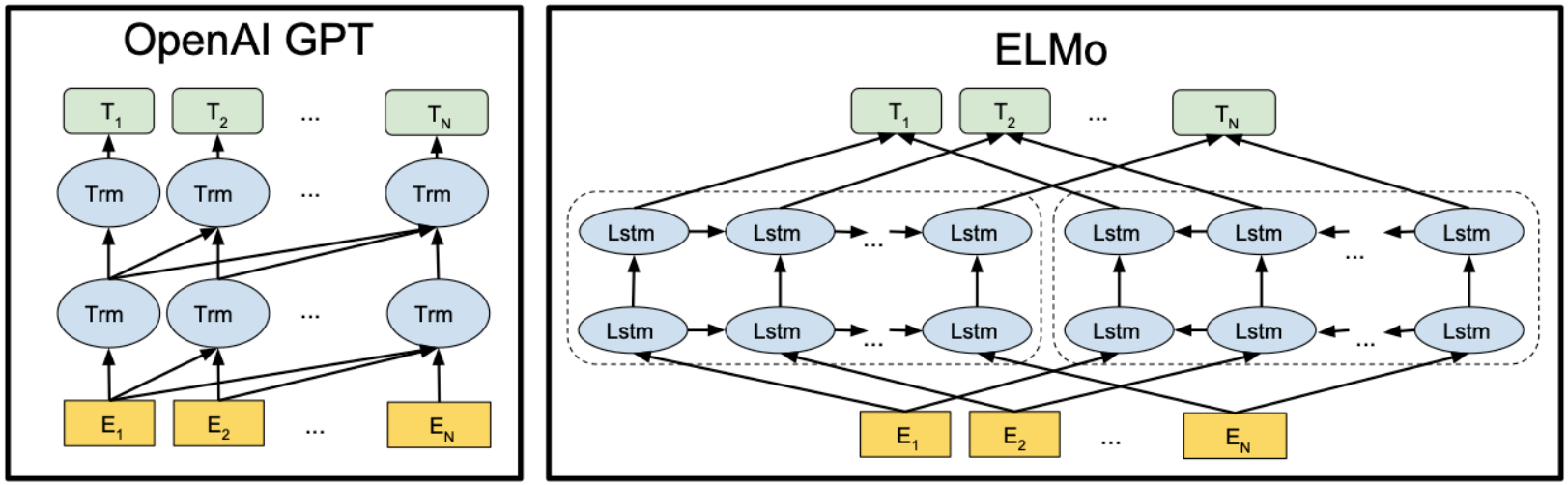
https://lilianweng.github.io/posts/2019-01-31-lm/#cove

GPT (a.k.a OpenAI GPT)

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Comparing with ELMo: The model architectures are different. ELMo uses a shallow concatenation of independently trained left-to-right and right-to-left multi-layer LSTMs, while GPT is a multi-layer transformer.



GPT (a.k.a OpenAI GPT)

UNIVERSITÄT DUISBURG ESSEN

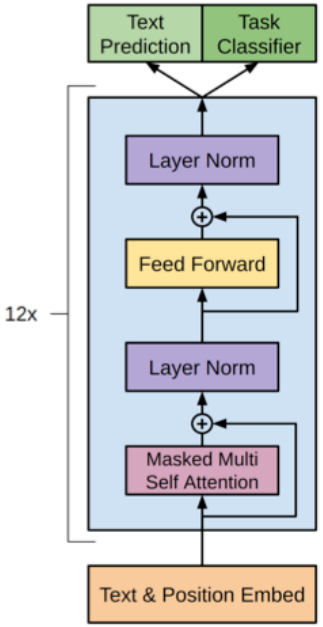
Offen im Denken

- Another novelty of GPT is that when the pre-training is finished, GPT is fine-tuned for downstream task.
- To do so, GPT uses a specific objection for the task and a set of examples of that task.
- This step is known as fine-tuning.

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y | x^1, \ldots, x^m)$$

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$



Language model objective helps in GPT in
(a) improving generalization of the supervised model,
(b) accelerating convergence.

GPT (a.k.a OpenAI GPT)

- The pre-training is conducted on the BooksCorpus dataset
  - over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance
  - Books contain long stretches of contiguous text, which allows the generative model to learn to condition on long-range information
  - 1B Word Benchmark is approximately the same size but is shuffled at a sentence level - destroying long-range structure
- Then fine-tuning can be conducted for any task, e.g., Question-answering, and text classification
  - Idea: instead of changing the objective, just change the input to GPT

GPT (a.k.a OpenAI GPT)

UNIVERSITÄT DUISBURG ESSEN

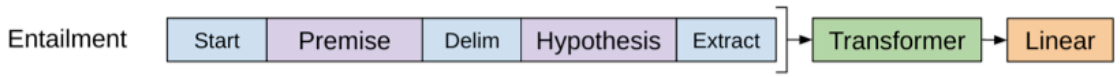
Offen im Denken

- Some examined tasks:
  - NLI: Natural language inference (NLI), also known as recognizing textual entailment, is the task of determining whether a "**hypothesis**" is true (**entailment**), false (**contradiction**), or undetermined (**neutral**) given a "**premise**".

P[a]	A senior is waiting at the window of a restaurant that serves sandwiches.	Relationship
H[b]	A person waits to be served his food.	Entailment
	A man is looking to order a grilled cheese sandwich.	Neutral
	A man is waiting in line for the bus.	Contradiction

[a]P, Premise.
[b]H, Hypothesis.

  - The task remains challenging due to the presence of a wide variety lexical entailment, coreference, and lexical and syntactic ambiguity.



GPT (a.k.a OpenAI GPT)

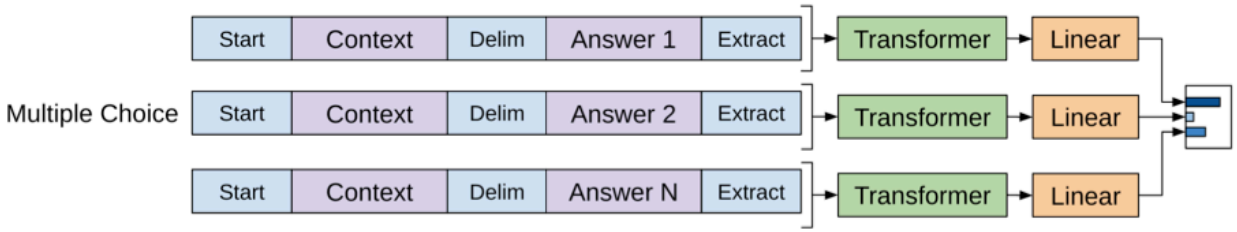
UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Some examined tasks:
  - QA: This task requires aspects of single and multi-sentence reasoning.

Story Cloze Test

Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She found her phone in the car.	She didn't want her phone anymore.



GPT (a.k.a OpenAI GPT)

UNIVERSITÄT DUISBURG ESSEN

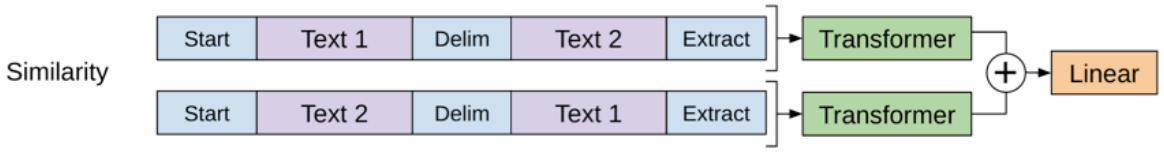
Offen im Denken

- Some examined tasks:
  - Sentence Similarity (or paraphrase detection): involve predicting whether two sentences are semantically equivalent or not.
  - The challenges lie in recognizing rephrasing of concepts, understanding negation, and handling syntactic ambiguity.
- Dataset Preview

Subset: qqp

Split: train

question1 (string)	question2 (string)	label (class label)	idx (int)
How is the life of a math student? Could you describe your own experiences?	Which level of prepration is enough for the exam jlpt5?	0 (not_duplicate)	0
How do I control my horny emotions?	How do you control your horniness?	1 (duplicate)	1
What causes stool color to change to yellow?	What can cause stool to come out as little balls?	0 (not_duplicate)	2
What can one do after MBBS?	What do i do after my MBBS ?	1 (duplicate)	3
Where can I find a power outlet for my laptop at Melbourne Airport?	Would a second airport in Sydney, Australia be needed if a high-speed rail link was created between Melbourne and Sydney?	0 (not_duplicate)	4
How not to feel guilty since I am Muslim and I'm conscious we won't have sex together?	I don't beleive I am bulimic, but I force throw up atleast once a day after I eat something and feel guilty. Should I tell somebody, and if so who?	0 (not_duplicate)	5



GPT (a.k.a OpenAI GPT)

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Some examined tasks:
  - Sentence Classification: assign a sentence to a category
  - Sentiment Classification: (https://huggingface.co/datasets)

Dataset Preview

Subset: sst2 | Split: train

sentence (string)	label (class label)
will find little of interest in this film , which is often preachy and poorly acted	0 (negative)
by far the worst movie of the year	0 (negative)
sit through ,	0 (negative)
more than another `` best man '' clone by weaving a theme throughout this funny film	1 (positive)
it 's about issues most adults have to face in marriage and i think that 's what i liked about it -- the real issues tucked between the silly and crude storyline	1 (positive)
heroes	1 (positive)
oblivious to the existence of this film	0 (negative)
sharply	1 (positive)

GPT is really large

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- It has 117M parameters
- Much larger than ELMo

Model	Link (Weights/Options File)	# Parameters (Millions)	LSTM Hidden Size/Output size
Small	weights | options	13.6	1024/128
Medium	weights | options	28.0	2048/256
Original	weights | options	93.6	4096/512
Original (5.5B)	weights | options	93.6	4096/512

Today

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

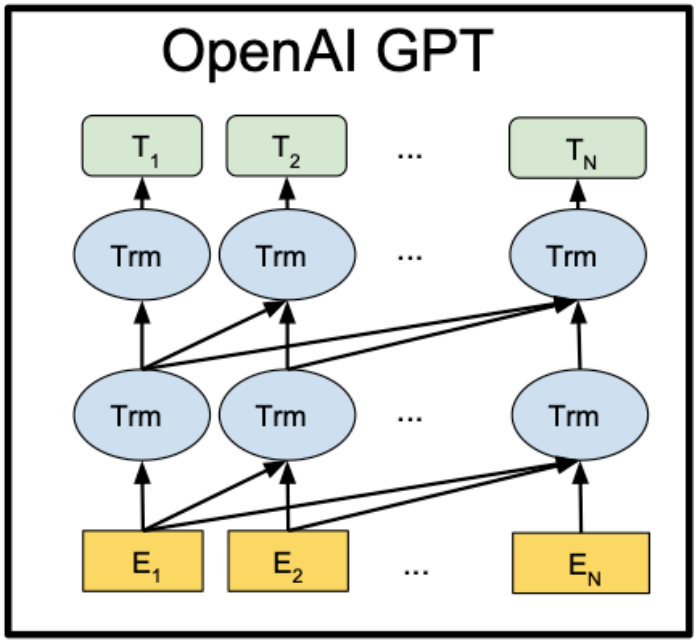
GPT-2

GPT-2

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Similar to GPT, GPT-2 uses transformers
-



GPT-2

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Trained for the language model objective, i.e., predicting the next token given previous tokens
- However GPT-2 uses much larger set of parameters than GPT
  - GPT-2 has 1.5 billion parameters. It is 10 times more than GPT-1 (117M parameters).

GPT-2

- **Data:** WebText
  - Aiming at building a dataset as large and diverse as possible in order to collect natural language demonstrations of tasks in as varied of domains and contexts as possible.
  - Scrape the web and collect 40GB of text data from over 8 million documents.
  - WebText is much larger than the dataset used to train OpenAI GPT.

GPT-2

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Add some description about the task to the input to the model
  - This modification is known as **task conditioning**, where the model is expected to produce **different output** for **same input** for **different tasks**
- Task descriptions provide examples or natural language instructions to the model to perform a task.
- The main idea is that natural language provides a flexible way to specify tasks, inputs, and outputs all as a sequence of symbols.

GPT-2

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Task description for a translation training example can be written as the sequence
  - (translate to german, *english text*, *german text*)
  - (translate to german, *I love NLP, ich liebe NLP*)
- The model is supposed to understand that it is a translation task and give German counterpart of English sentence.

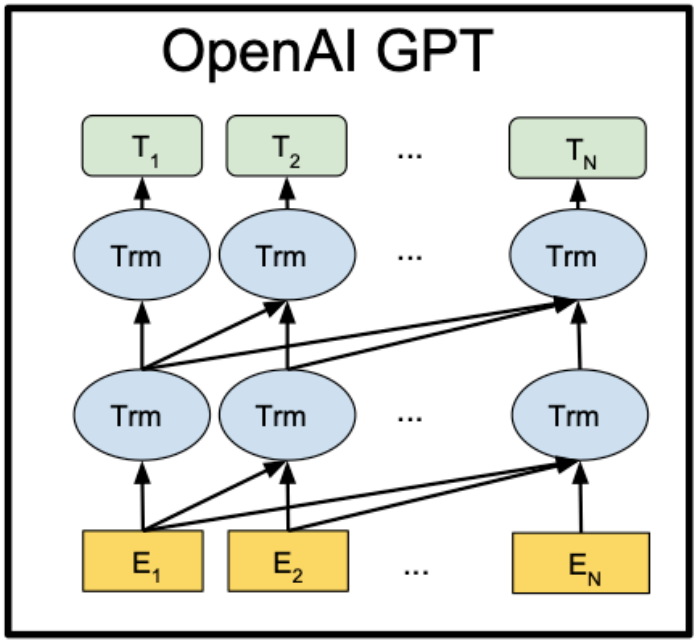
- Task description for a reading comprehension training example can be written as
  - (answer the question, document, question, answer)

GPT-2

- Task conditioning forms the basis for **zero-shot task transfer**.
- **Zero shot task transfer** refers to the setting in which the model is presented with **few to no examples**, to make it understand the task.
- The term zero shot comes from the fact that no gradient updates are performed.
- The model is supposed to understand the task based on the examples and instruction.
- **Zero shot learning** is a special case of zero shot task transfer where no examples are provided at all and the model understands the task based on the given instruction
- GPT-2 is able to do zero-shot task transfer.

GPT-3

GPT-3



GPT-3

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **Size:** GPT-3 model with **175 billion parameters**
  - 100 times more parameters than GPT-2
- **Data:** GPT-3 was trained on a mix of five different corpora:
  - Common Crawl, WebText2, Books1, Books2 and Wikipedia
  - Each having certain weight: High quality datasets were sampled more often, and model was trained more often on those datasets.

- **In-Context Learning:** With in-context learning, the text given to the model is a written description (optional) plus some examples. The last example is left unfinished for the model to complete.
- In-context learning is flexible. We can use this scheme to describe many possible tasks, from translating between languages to improving grammar to coming up with joke punch-lines

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

GPT-3

- GPT-3 is good in few-shot and zero-shot transfer
- **Few-shot, one-shot and zero-shot setting**:
  - few, one and zero-shot settings are specialized cases of zero-shot task transfer.
  - In few-shot setting, the model is provided with task description and as many examples as fit into the context window of model.
  - In one-shot setting the model is provided exactly one example
  - In zero-shot setting no example is provided.
  - With increase in capacity of model, few, one and zero-shot capability of model also improves.

GPT-3

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- GPT-3 is good in few-shot and zero-shot transfer
- **GPT-3 is not publicly available**

Today

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

OPT

OPT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- The architecture is Transformers
- Size: ranging from 125M to 175B parameters.
- OPT-175B is comparable to GPT-3
- **Data:**
  - Pile (**CommonCrawl**, **DM Mathematics, Project Gutenberg, OpenSubtitle, OpenWebText2, Wikipedia)**,
  - **PushShift.io** Reddit,
  - **BookCorpus, Stories, an up- dated version of CCNews, containing news stories crawled through September 28, 2021,**

- All corpora were previously collected or filtered to contain predominantly English text, but a small amount of non-English data is still present within the corpus via CommonCrawl.

.

In this lecture, you learn about …

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **How do neural language models learn knowledge?**
  - Autoregressive methods
  - **Autoencoding methods**

Autoencoding

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Corrupt tokens of a sentence
- Try to reconstruct the original sentence
- Example language models that are trained by auto encoding
  - BERT (2018)
  - RoBERTa (2019)
  - XLM (2019)
  - DistillBERT (2020)

Today

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

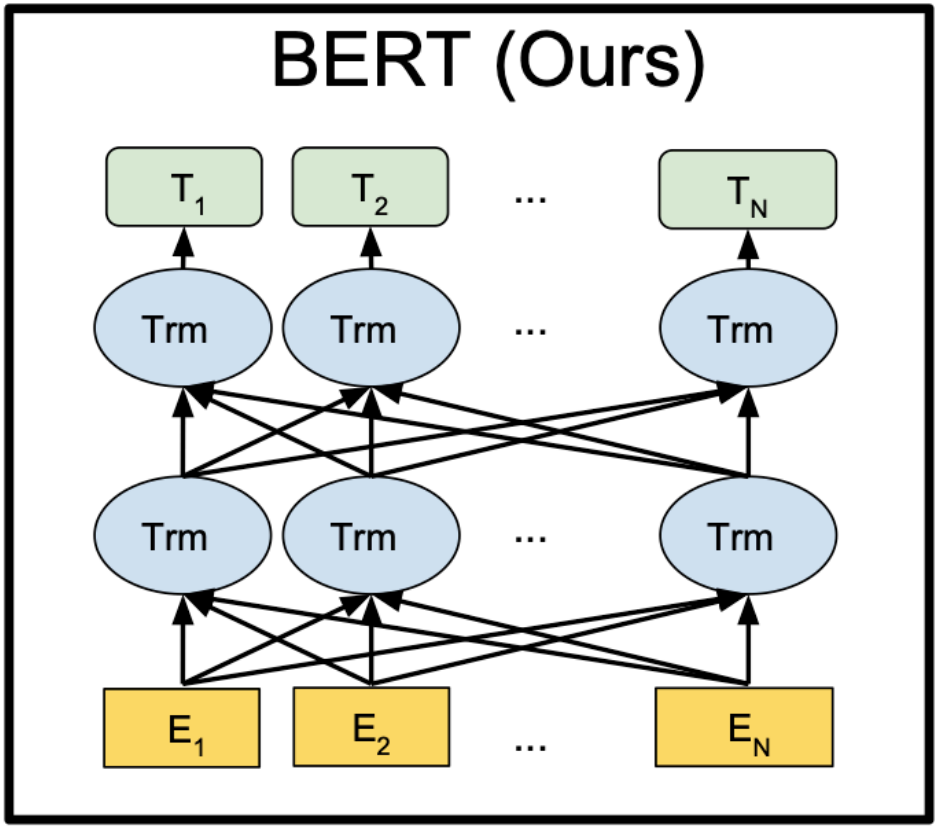
BERT

BERT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

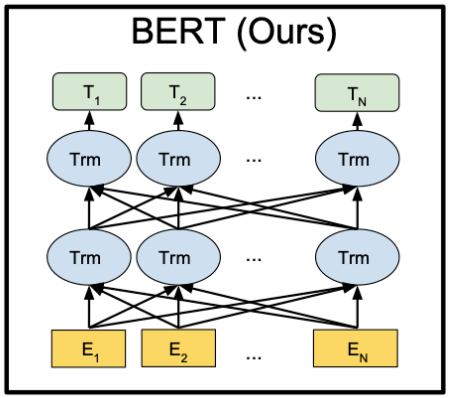
- Stands for Bidirectional Encoder Representations from Transformers

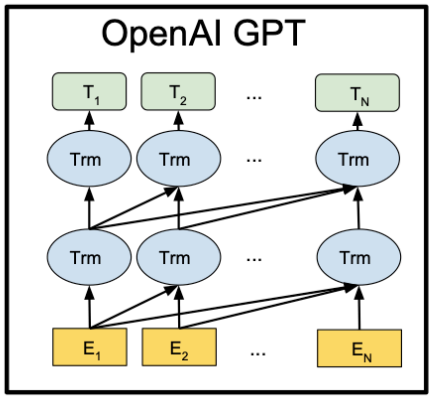


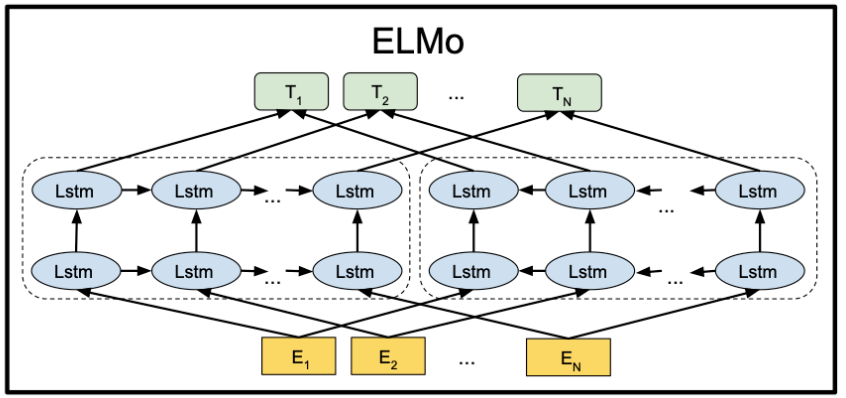
BERT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken





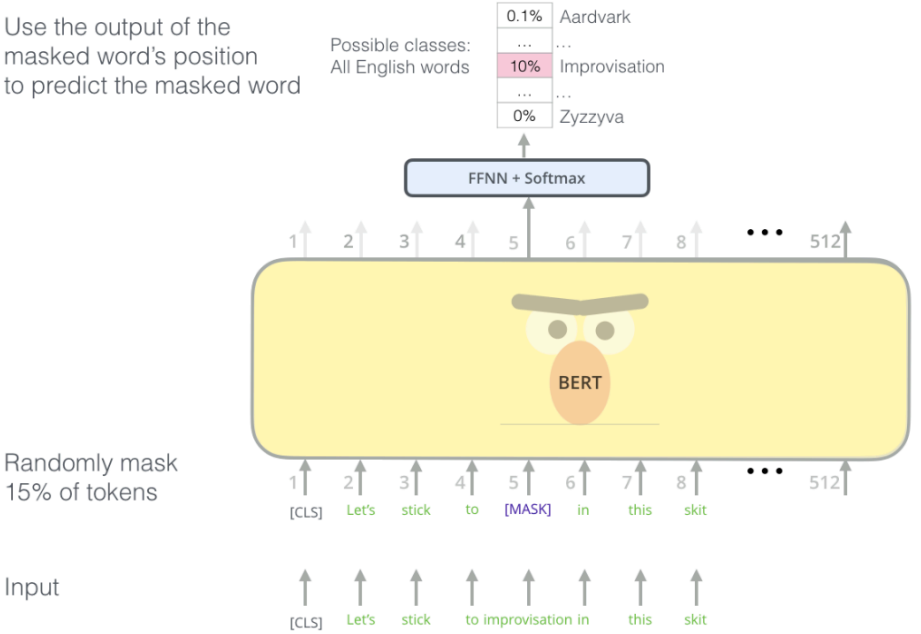


BERT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- BERT relies on Transformers as well
- BERT is bidirectional
- BERT is a **Masked Language Model**
-



BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

https://jalammar.github.io/illustrated-bert/

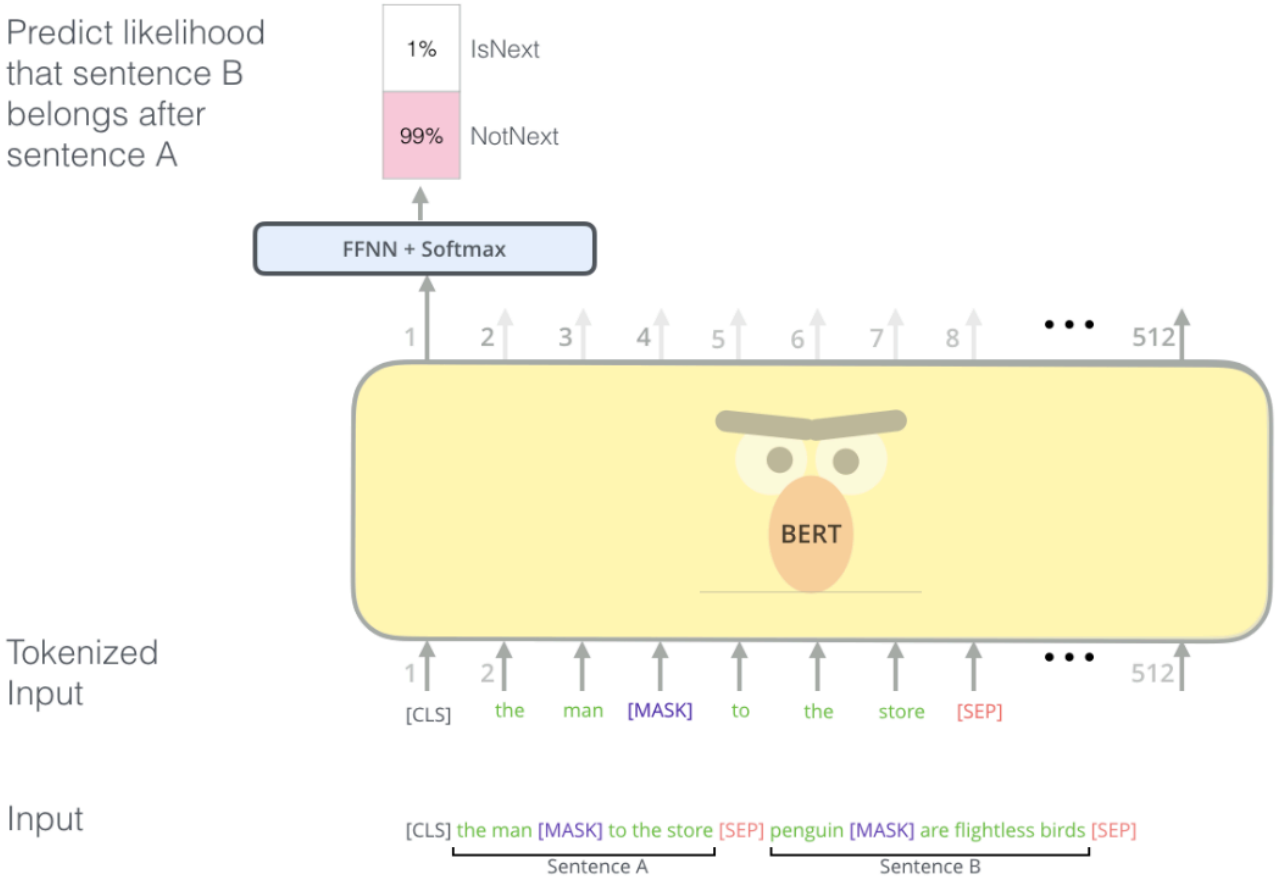
BERT

- Corrupts the inputs by using random masking,
- during pretraining, a given percentage of tokens (usually 15%) is masked by:
  - a special mask token with probability 0.8
  - a random token different from the one masked with probability 0.1
  - the same token with probability 0.1
- The model is trained for two objectives:
  - It must predict the original sentence
  - Inputs are two sentences A and B (with a separation token in between). With probability 50%, the sentences are consecutive in the corpus, in the remaining 50% they are not related. The model must predict if the sentences are consecutive or not.

BERT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



The second task BERT is pre-trained on is a two-sentence classification task. The tokenization is oversimplified in this graphic as BERT actually uses WordPieces as tokens rather than words --- so some words are broken down into smaller chunks.

BERT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- BERT has two versions:
  - $\text{BERT}_{\text{BASE}}$: 110 million parameters
  - $\text{BERT}_{\text{LARGE}}$: 340 million parameters

- Data: 16GB unlabeled text data extracted from
  - BooksCorpus with 800M words
  - English Wikipedia with 2,500M words.
- BERT was trained on over 100 languages, it wasn't optimized for multi-lingual models — most of the vocabulary isn't shared between languages and therefore the shared knowledge is limited.

Today

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

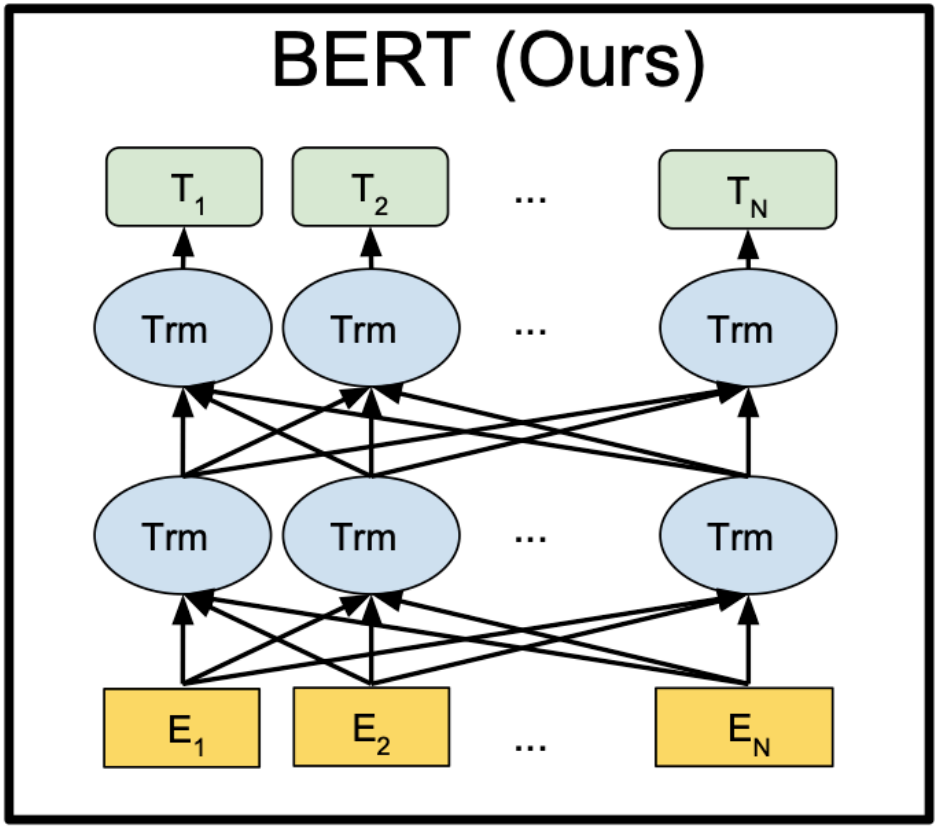
RoBERTa

RoBERTa

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Stands for Robustly Optimized BERT Pre-training Approach



RoBERTa

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Same as BERT with more intuitive pre-training tricks:
  - dynamic masking: tokens are masked differently at each learning step, whereas BERT does it once and for all
  - No NSP (next sentence prediction) objective instead of putting just two sentences together, put a chunk of contiguous texts together to reach 512 tokens (so the sentences are in an order than may span several documents)

RoBERTa

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Use 160GB text data
  - Bookcoupus, CC- NEWS, OpenWebText, and Stories
  - 10 times more than what BERT is trained on)

- Size:
  - $\text{RoBERTa}_{\text{BASE}}$ has123 million parameters
  - $\text{RoBERTa}_{\text{LARGE}}$ has 354 million parameters
  - $\text{BERT}_{\text{BASE}}$: 110 million parameters and $\text{BERT}_{\text{LARGE}}$: 340 million parameters

Today

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

XLM

XLM

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Stands for Cross-lingual Language Model Pre-training
- it upgrades the BERT architecture in two manners
  - Each training sample consists of the same text in two languages, whereas in BERT each sample is built from a single language. As in BERT, the goal of the model is to predict the masked tokens, however, with the new architecture, the model can use the context from one language to predict tokens in the other, as different words are masked words in each language (they are chosen randomly).
  - The model also receives the language ID and the order of the tokens in each language, separately. The new metadata helps the model learn the relationship between related tokens in different languages.

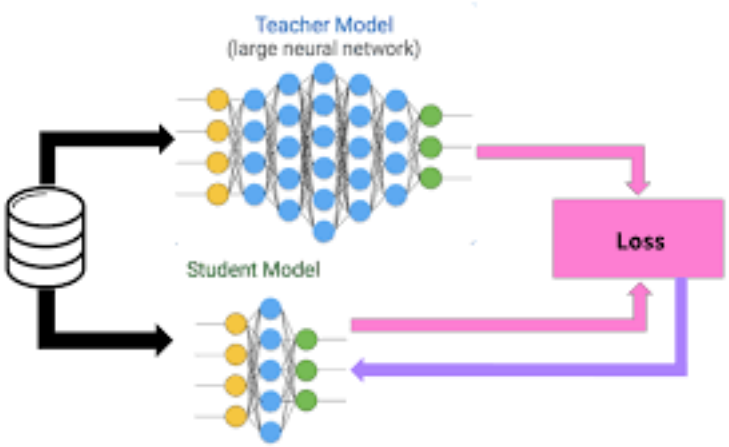
DistilBERT

DistillBERT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- A BERT-like architecture
- **Distillation**: a technique you can use to compress a large model, called the **teacher**, into a smaller model, called the **student**.
- *Knowledge distillation* (a.k.a *teacher-student learning*) is a compression technique in which a small model is trained to reproduce the behavior of a larger model (or an ensemble of models).

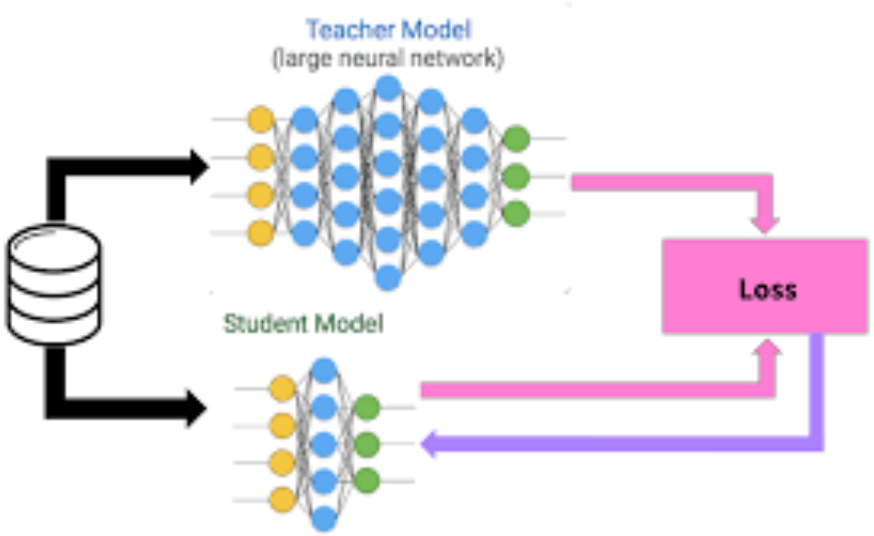


DistilBERT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- In the **teacher-student training**, we train a student network to mimic the **full output distribution** of the teacher network (its knowledge).



DistilBERT

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Using the teacher signal, we are able to train a **smaller language model,** we call **DistilBERT,** from the **supervision of BERT**
  - **the teacher is BERT**
  - **the student is a small version of BERT**
- DistilBERT, has about half the total number of parameters of BERT base, runs 60% faster while preserving, and retains 95% of BERT's performances
- DistillBERT also used the training tricks used by RoBERTa
  - dynamic masking and removed the next sentence prediction objective.
-

Summary

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **How do neural language models learn knowledge?**
  - Autoregressive methods
  - Autoencoding methods

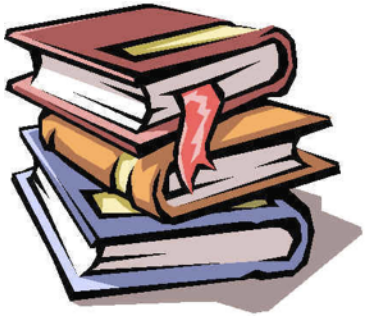
Readings

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Mandatory

- https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- https://aclanthology.org/N19-1423.pdf
- Slides



Thanks

World Knowledge

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- We observed that symbolic KB can give us factual knowledge about world
- **Google RE:** place_of_death, date_of_birth, education_degree, place_of_birth (https://code.google.com/archive/p/ relation- extraction-corpus/)

Practice I

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Use your notebook in Google Colab (https://colab.research.google.com)
- Download the **Google RE dataset** (https://code.google.com/archive/p/relation-extraction-corpus/)
  - Focus on "**place of birth**", "**date of birth**" and "**place of death**" relations
  - **How many facts do exist for each relation?**
- Define **a template for each relation** to query a LM
- Select a LM, e.g. BERT, RoBERTA, ELMo, …
- **For how many facts does the selected LM return the correct value?**
  - **compute P@1**
  - P@k: Is the correct value among the k top outputs that the LM returns?
- Write a report in overleaf **without screen shots**

Practice II

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- How to get dataset for subject-verb agreement?
  - Go to wikipedia or any other textual corpus in NLTK
  - Extract 1000 sentences
  - Mask all verbs
    - How to automatically find which word is a verb? Use NLTK or SpaCy
- https://github.com/BeckyMarvin/LM_syneval
- For how many sentences your LM returns a verb that is in agreement with its subject? Report P@1
- Write a paragraph about this experiment in overleaf.

Thanks