Pre-trained transformer models 1

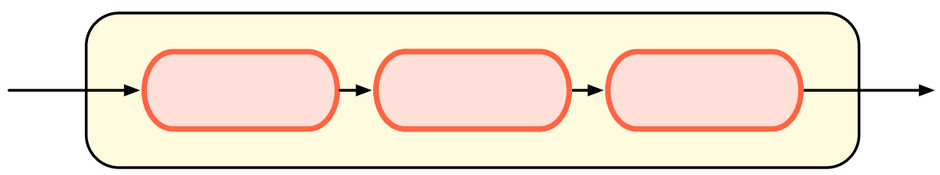
Marco Kuhlmann
Department of Computer and Information Science

Transfer learning

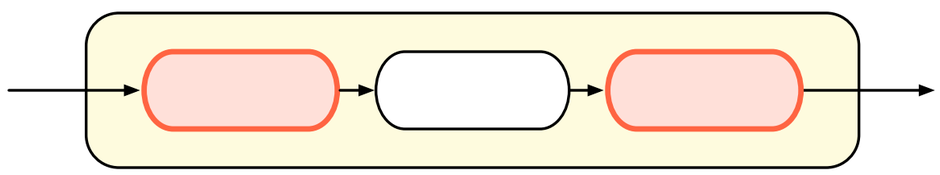
- **Transfer learning** aims to re-use knowledge gained while solving one problem when solving the next problem.

  may speed up training, reduce the need for training data

- In contemporary NLP, transfer learning is usually implemented through **pre-training and fine-tuning**.



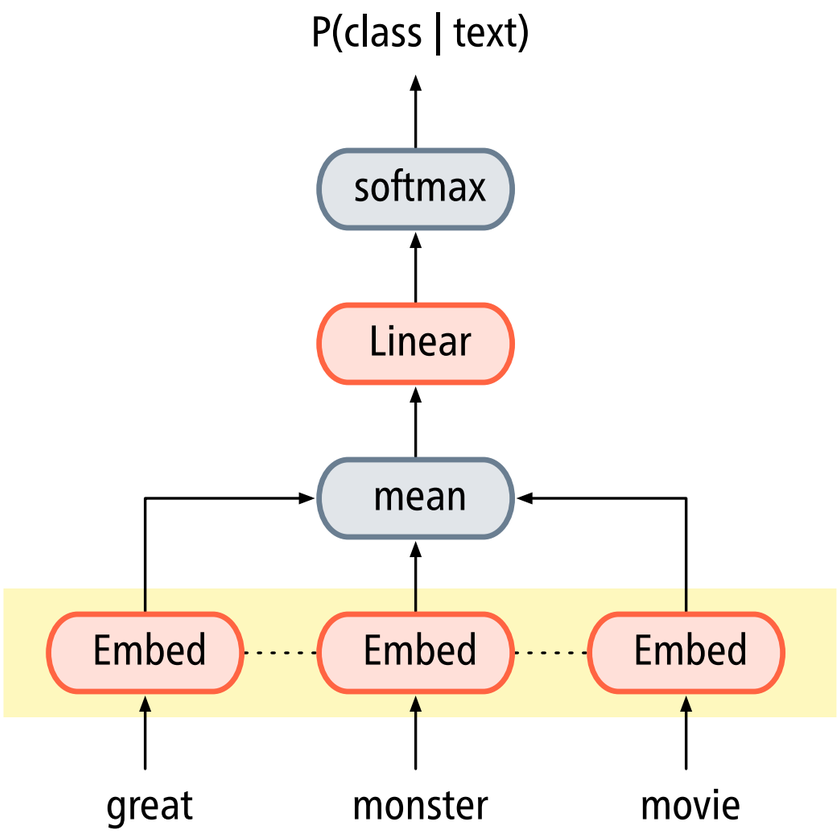
Model trained on task A



Model to be trained on task B

Pre-training and fine-tuning word embeddings

pre-trained embeddings



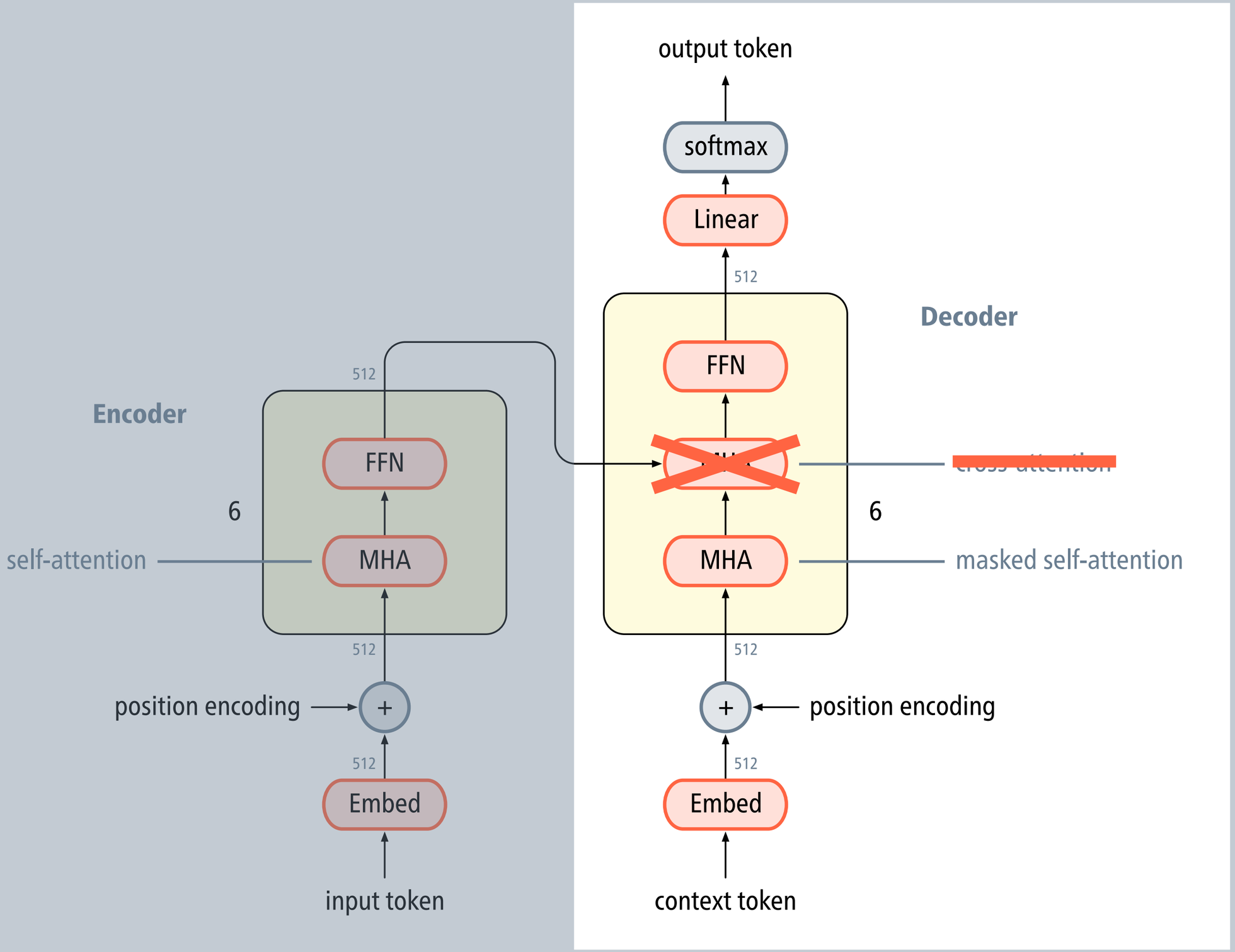
Generative pre-training, discriminative fine-tuning

- Idea: Combine generative pre-training on language modelling with discriminative fine-tuning on each task.
- Language modelling is a strong candidate for a pre-training task, as large unlabelled text corpora are abundant.

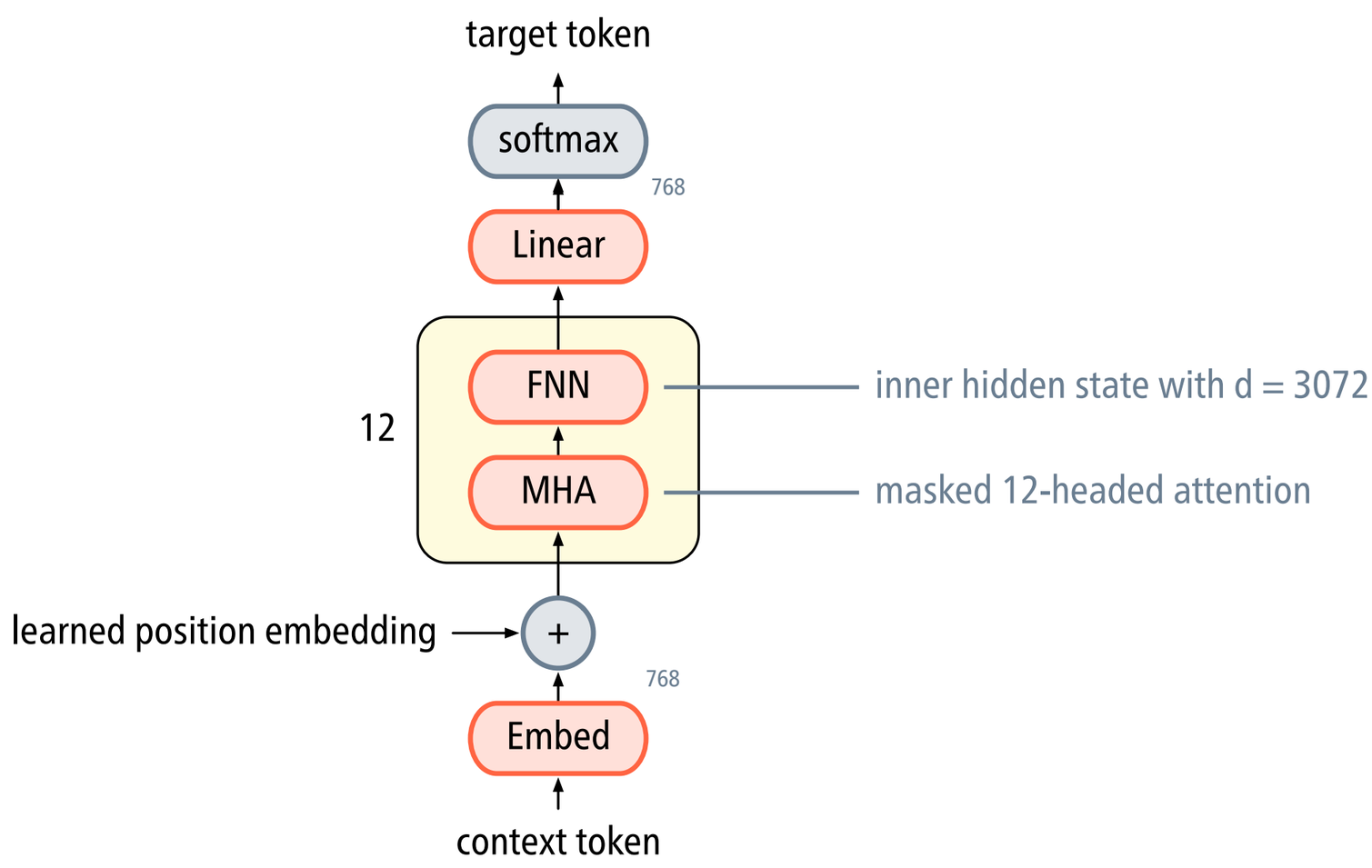
  at least for English …

- To facilitate effective transfer learning, the authors use task-specific input transformations for fine-tuning.

Radford et al. (2018)


output token
softmax
Linear
512
Decoder
FFN
cross-attention
6
MHA
masked self-attention
512
position encoding
+
512
Embed
context token
Encoder
512
FFN
6
self-attention
MHA
512
position encoding
+
512
Embed
input token

GPT model architecture



Radford et al. (2018)

Pre-training statistics (largest models)

	GPT-1	GPT-2	GPT-3
Number of dimensions	768	1,600	12,288
Number of layers	12	48	96
Trainable parameters	**0.117 B**	**1.542 B**	**175 B**
Training data size (tokens)	800 M	(40 GB text)	499 B

Radford et al. (2018), Radford et al. (2019), Brown et al. (2020)

GPT as a generative model

Model prompt (human-written)	In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.
Model completion (machine-written)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez. Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns. […]

OpenAI Blog (2019)

Fine-tuning tasks used by Radford et al. (2018)		GLUE Benchmark
Classification SST-2, CoLA	Predict the overall sentiment of a sentence: positive or negative	**Text:** It's definitely not dull. **Label:** positive
Natural language inference SNLI, MultiNLI, Question NLI, RTE	Determine the logical relation between two sentences: entailment, contradiction, neutral	**Premise:** A man inspects the uniform of a figure in some East Asian country. **Hypothesis:** The man is sleeping. **Label:** contradiction
Sentence similarity MSR Paraphrases, QQP, STS	Rate the similarity of two sentences on a scale between 0–5	**Sentence 1:** Two boys on a couch are playing video games. **Sentence 2:** Two boys are playing a video game. **Similarity score:** 4
Question answering RACE, Story Cloze	Answer multiple-choice reading comprehension questions based on a text	**Question:** The first postage stamp was made … **Candidate answers:** A. in England, B. in America, C. by Alice, D. in 1910

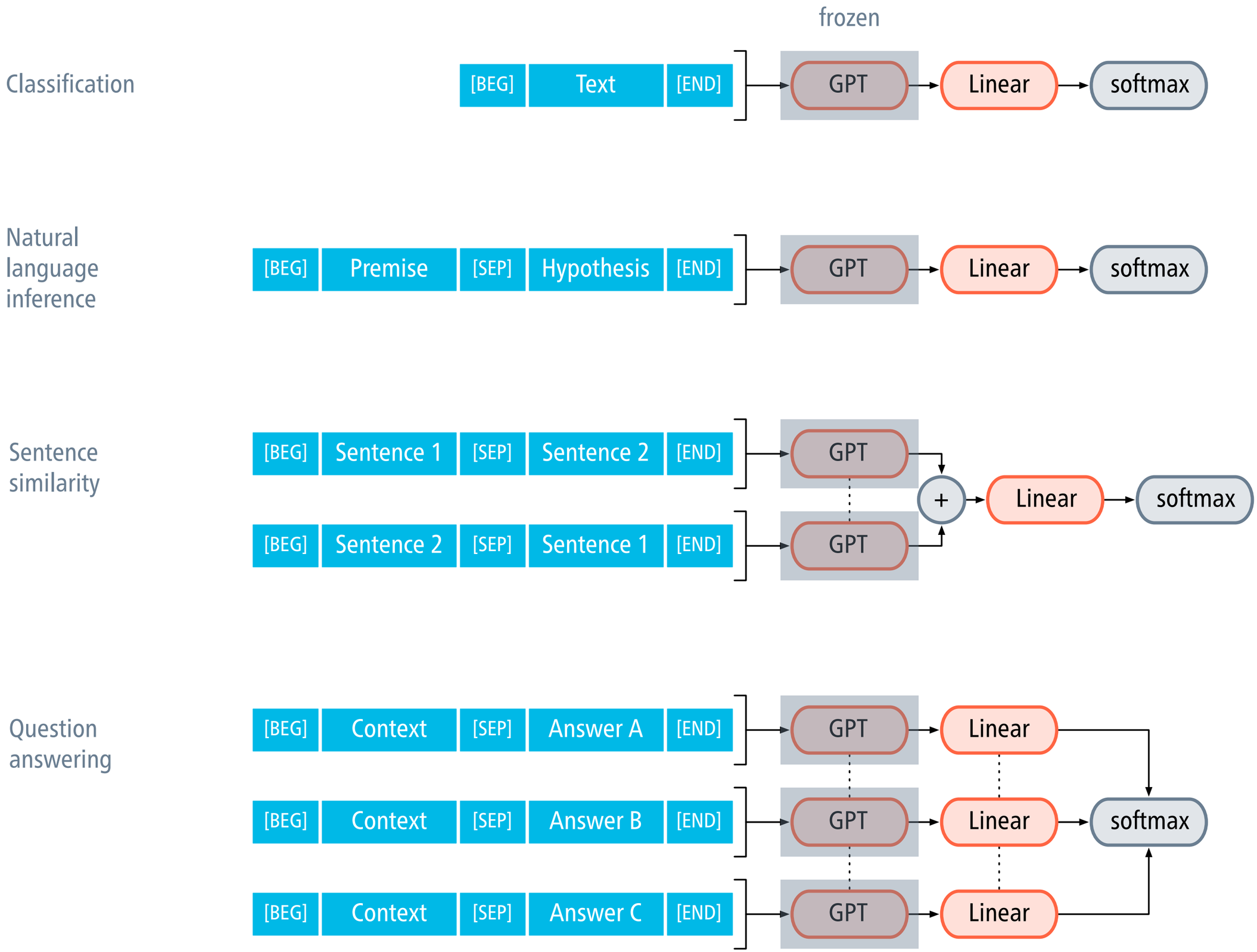
Fine-tuning on classification tasks

- For fine-tuning on classification, the final transformer block's output is fed into an added linear layer followed by a softmax.
- Thus, the only extra parameters introduced in fine-tuning are those of the added linear layer.

  plus embeddings for delimiter tokens

- Including language modelling as an auxiliary training objective to the fine-tuning improves generalisation and convergence.

  overall loss = classification loss + language modelling loss


frozen
Classification
[BEG]
Text
[END]
GPT
Linear
softmax
Natural language inference
[BEG]
Premise
[SEP]
Hypothesis
[END]
GPT
Linear
softmax
Sentence similarity
[BEG]
Sentence 1
[SEP]
Sentence 2
[END]
[BEG]
Sentence 2
[SEP]
Sentence 1
[END]
GPT
GPT
+
Linear
softmax
Question answering
[BEG]
Context
[SEP]
Answer A
[END]
[BEG]
Context
[SEP]
Answer B
[END]
[BEG]
Context
[SEP]
Answer C
[END]
GPT
GPT
GPT
Linear
Linear
Linear
softmax
Radford et al. (2018)

The zero-shot behaviour of GPT-2

- The generative model appears to learn downstream tasks as a by-product of language modelling.
- This allows us to solve downstream tasks in a **zero-shot** setup, i.e., without any task-specific fine-tuning.
- For effective zero-shot learning, we can design **prompts** that prime the pre-trained model for the task at hand.

Radford et al. (2019)

Examples of prompts

Sentiment classification

Tweet: ‘I hate it when my battery dies.’
Sentiment: Negative

Tweet: ‘My day has been great!’
Sentiment: Positive

Tweet: ‘This music video was incredible!’
Sentiment: **Positive**

Machine translation

Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese => **fromage**

black text provided by the user, red text generated by GPT-3