

Decoder-based language models: GPT

Marco Kuhlmann

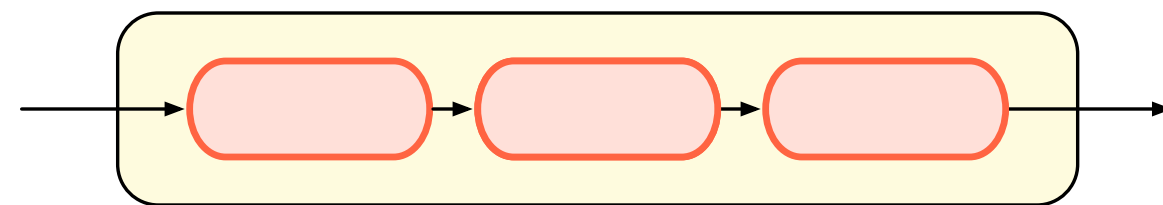
Department of Computer and Information Science

Pre-training and fine-tuning

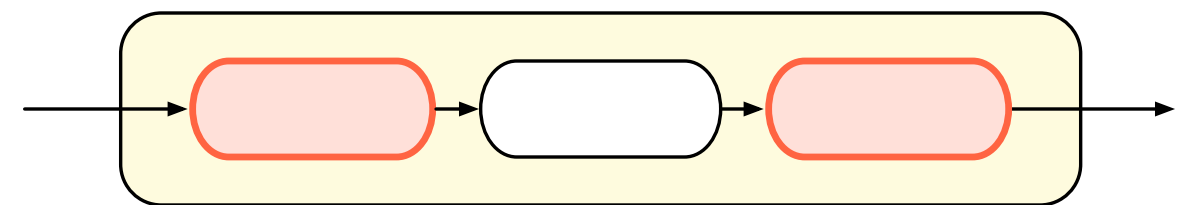
- **Transfer learning** aims to re-use knowledge gained while solving one problem when solving the next problem.

reduce the need for training data

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

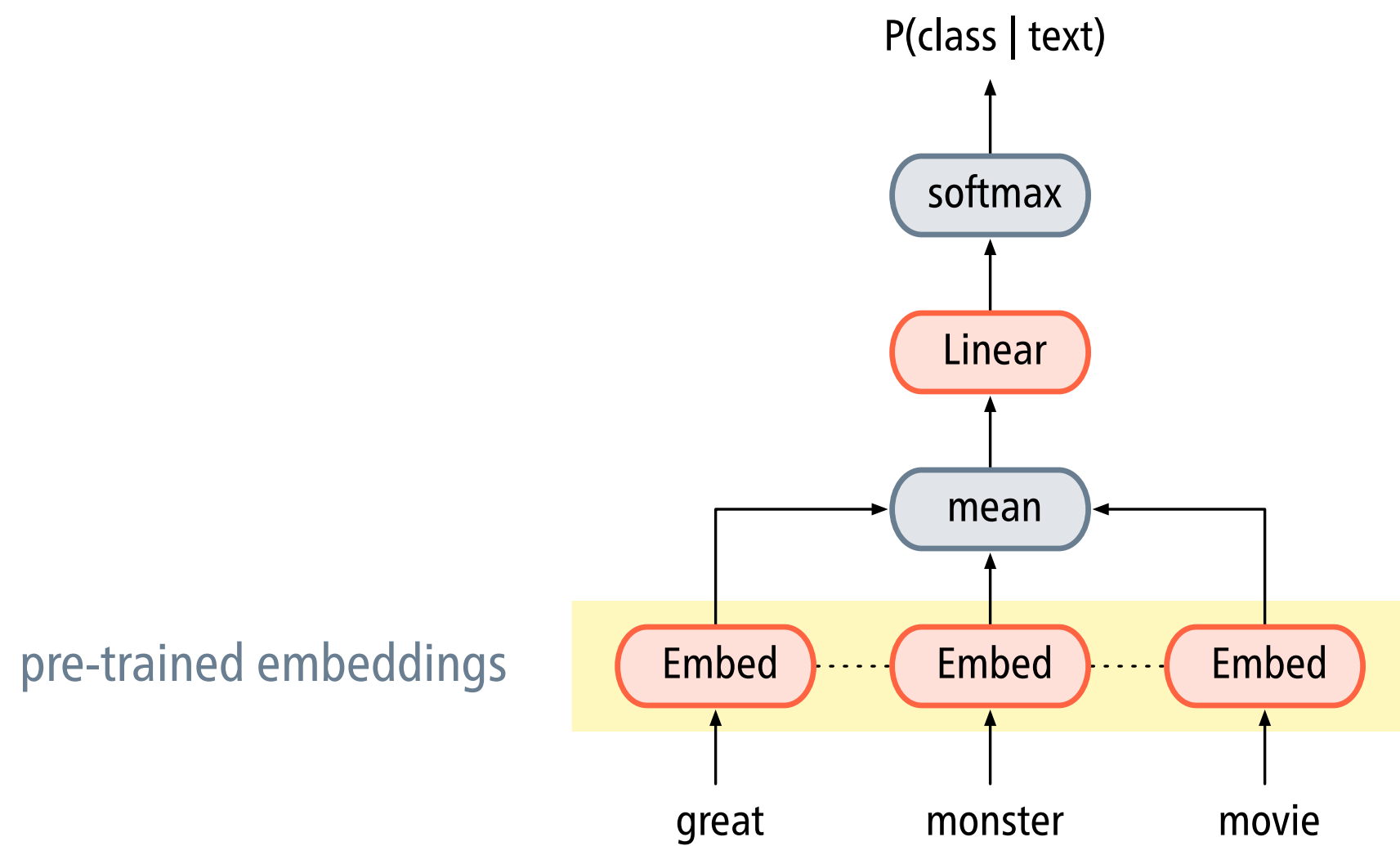


Model trained on task A



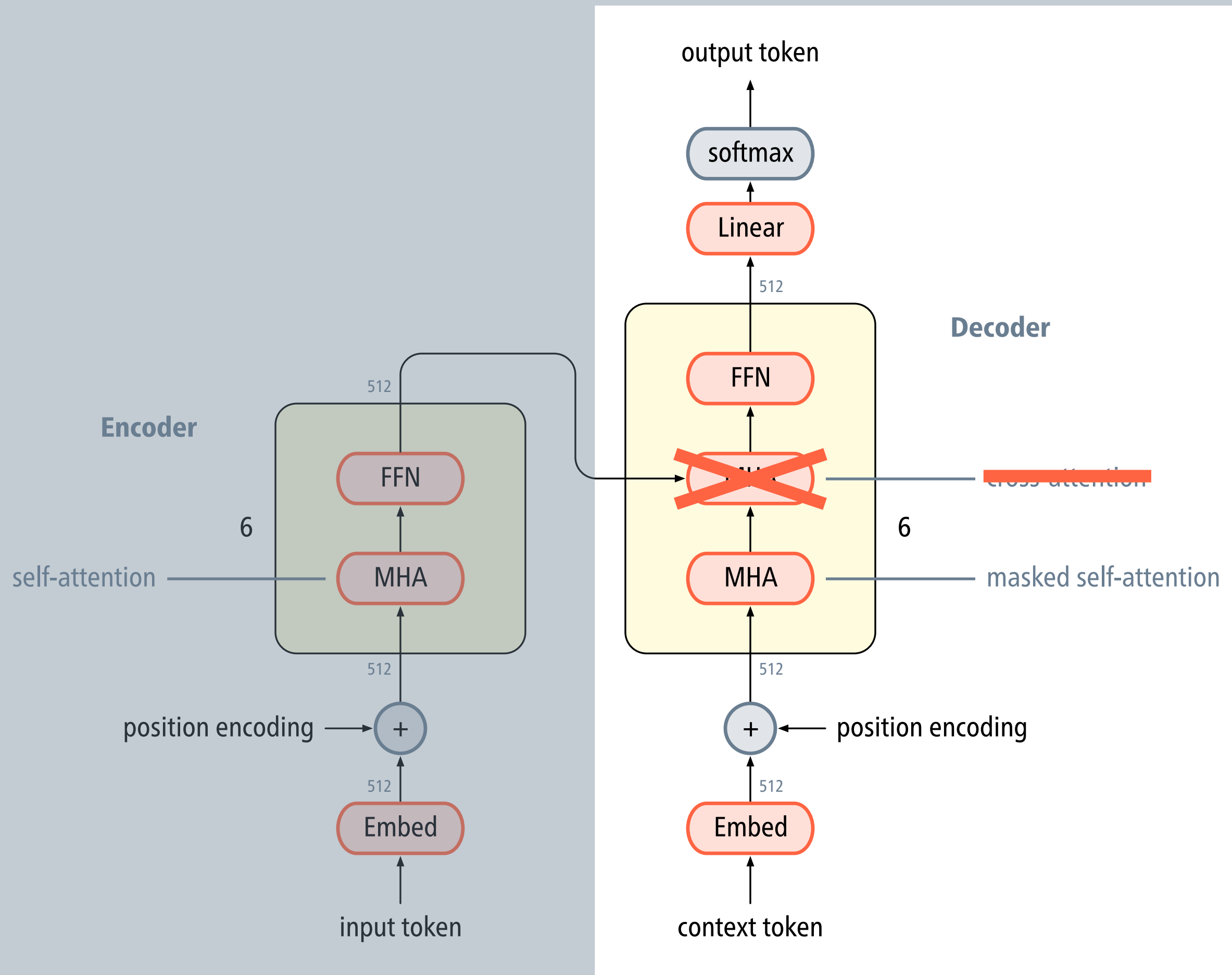
Model to be trained on task B

Pre-training and fine-tuning

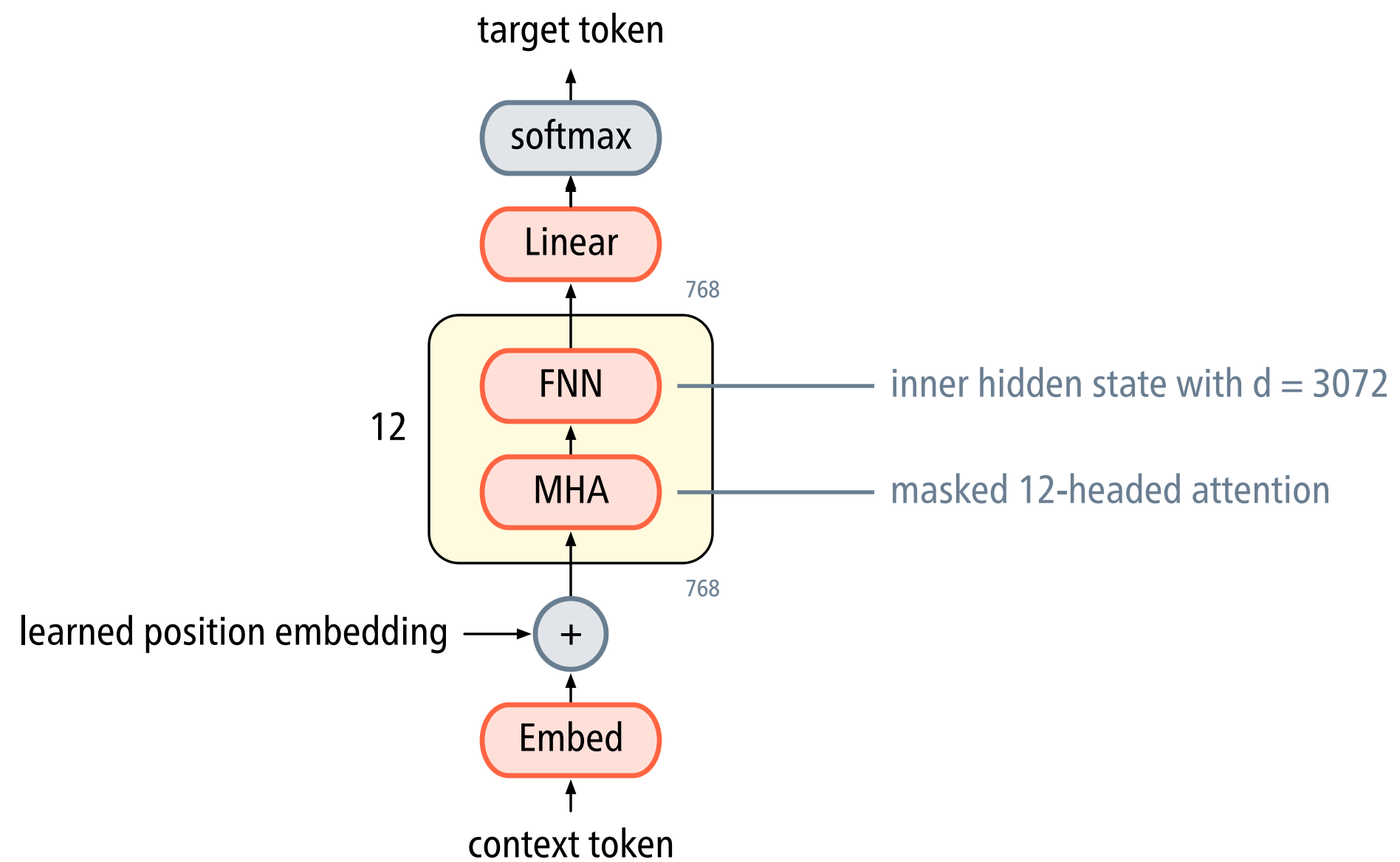


Generative pre-training, discriminative fine-tuning

- Idea: Combine generative pre-training (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.



GPT model architecture



Model statistics (largest models)

	GPT-1	GPT-2	GPT-3	GPT-4
Number of dimensions	768	1,600	12,288	?
Number of layers	12	48	96	120
Trainable parameters	0.117 B	1.542 B	175 B	1,800 B
Training data size (tokens)	800 M	(40 GB text)	499 B	13,000 B

[Radford et al. \(2018\)](#), [Radford et al. \(2019\)](#), [Brown et al. \(2020\)](#)

GPT as a language 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. [...]

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

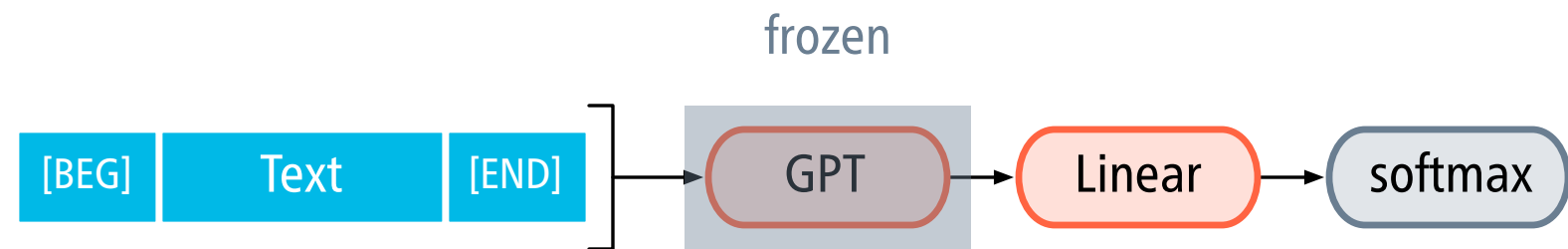
- For fine-tuning on classification, the final transformer block's output is fed into an added linear layer followed by a softmax.
- The transformer is frozen; the only extra parameters introduced in fine-tuning are those of the added linear layer.

plus embeddings for delimiter tokens

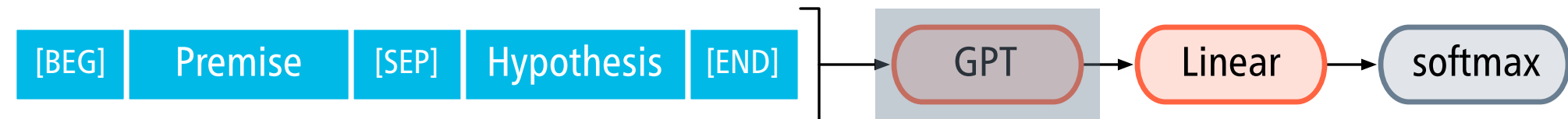
- Including language modelling as an auxiliary training objective to the fine-tuning improves generalisation and convergence.

overall loss = classification loss + language modelling loss

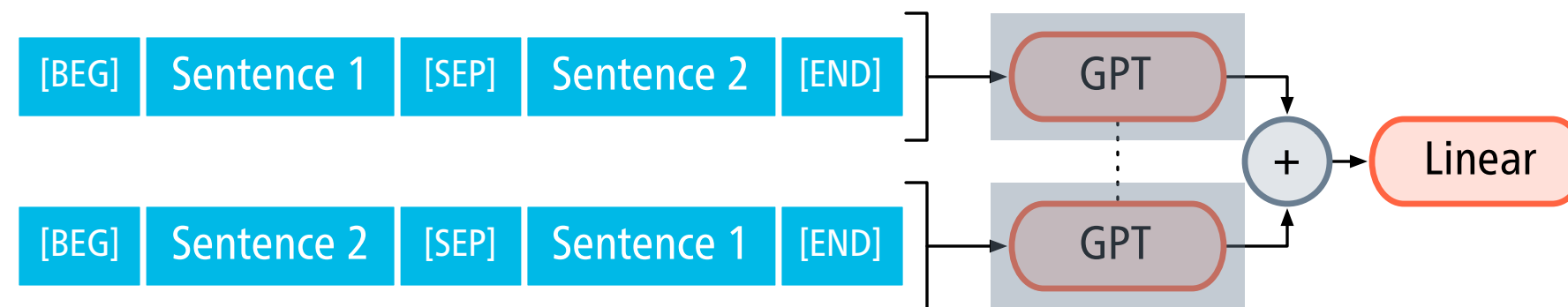
Classification



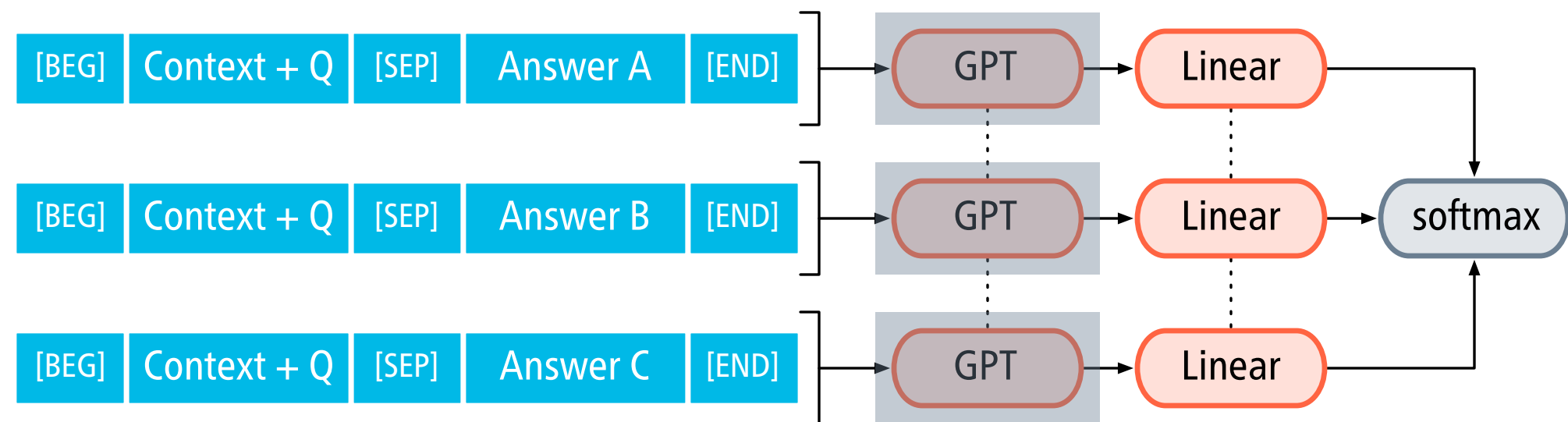
Natural language inference



Sentence similarity



Question answering



Large language models are zero-shot learners

- Many downstream tasks are directly or indirectly “demonstrated” in the text used for training language models.

translation: dictionaries, product manuals written in several languages

- Language models may learn these tasks as a by-product of learning to predict the demonstrating text.
- Large language models can solve downstream tasks without any task-specific fine-tuning; they are **zero-shot** learners.

Prompting for natural language processing

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 giraffe => girafe en peluche

cheese => **fromage**

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