Deep Learning & LLMs 2

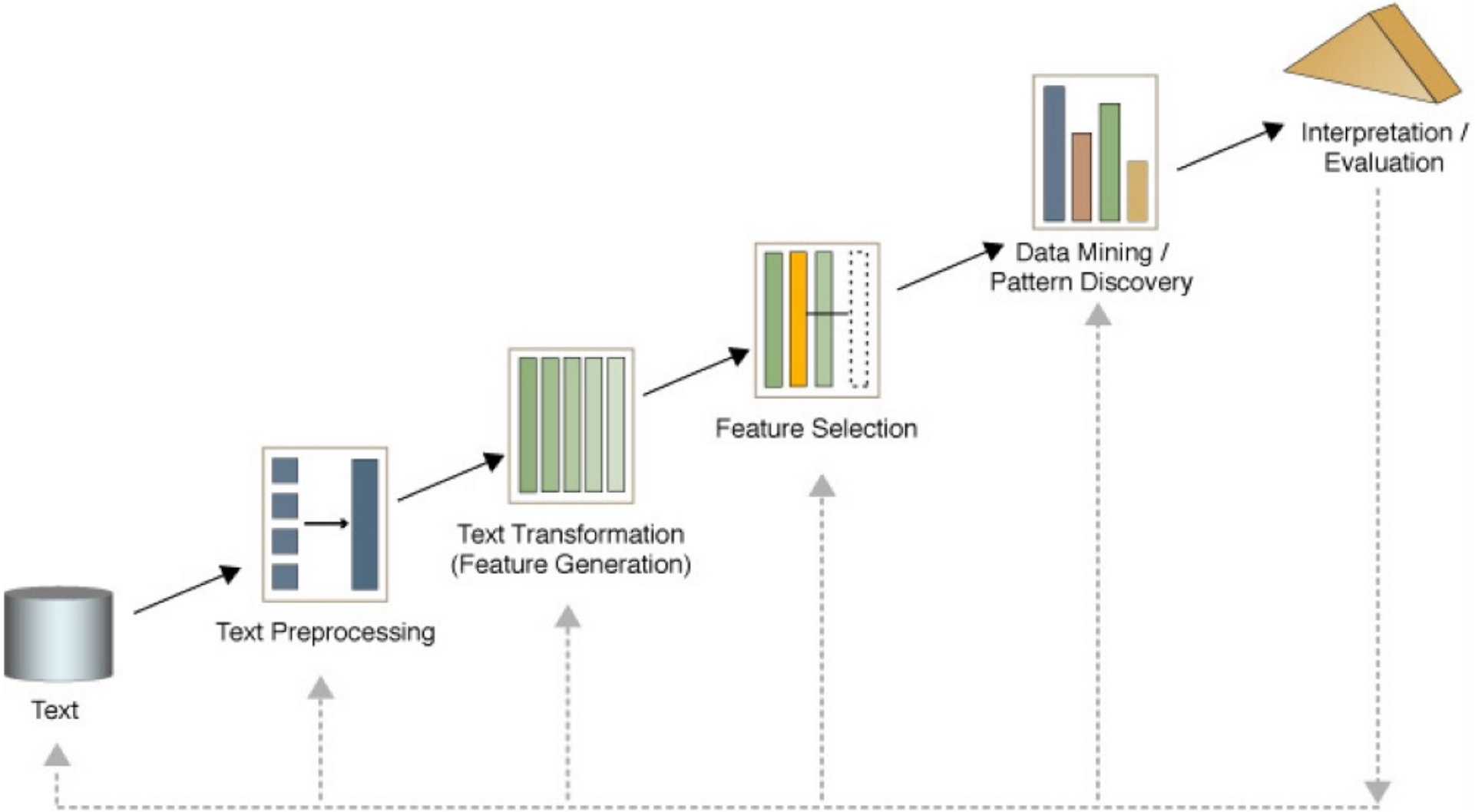
Applied Text Mining, from Foundations to Advanced

Ayoub Bagheri

This lecture

- Neural networks 2
- Convolutional neural networks
- State-of-the-art methods

Text mining process



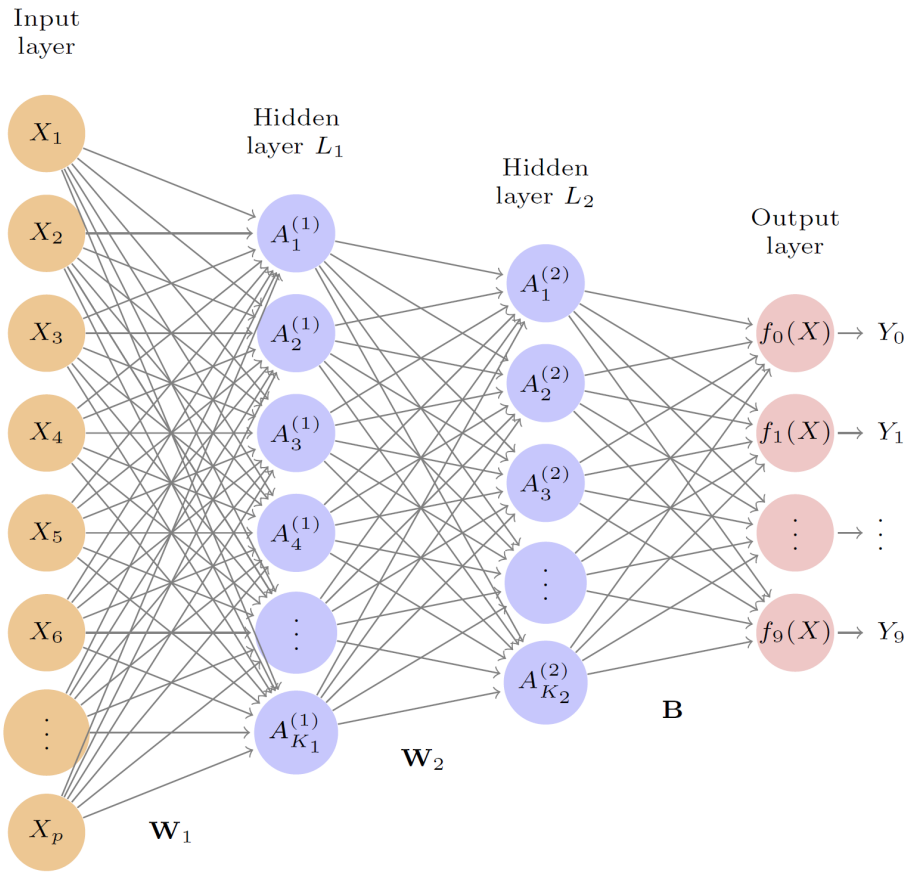
Introduction

Why should we learn this?

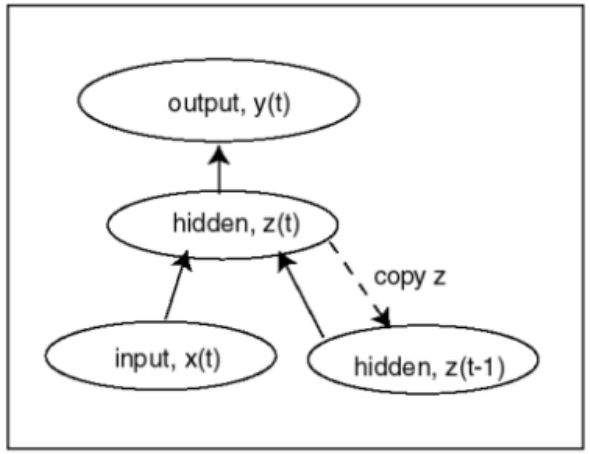
State-of-the-art performance on various tasks

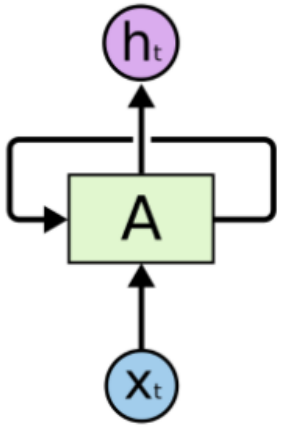
- Text prediction (your phone’s keyboard)
- Text mining
- Forecasting
- Spam filtering
- Compression (dimension reduction)
- Text generation
- Translation
- …

Feed-forward neural networks

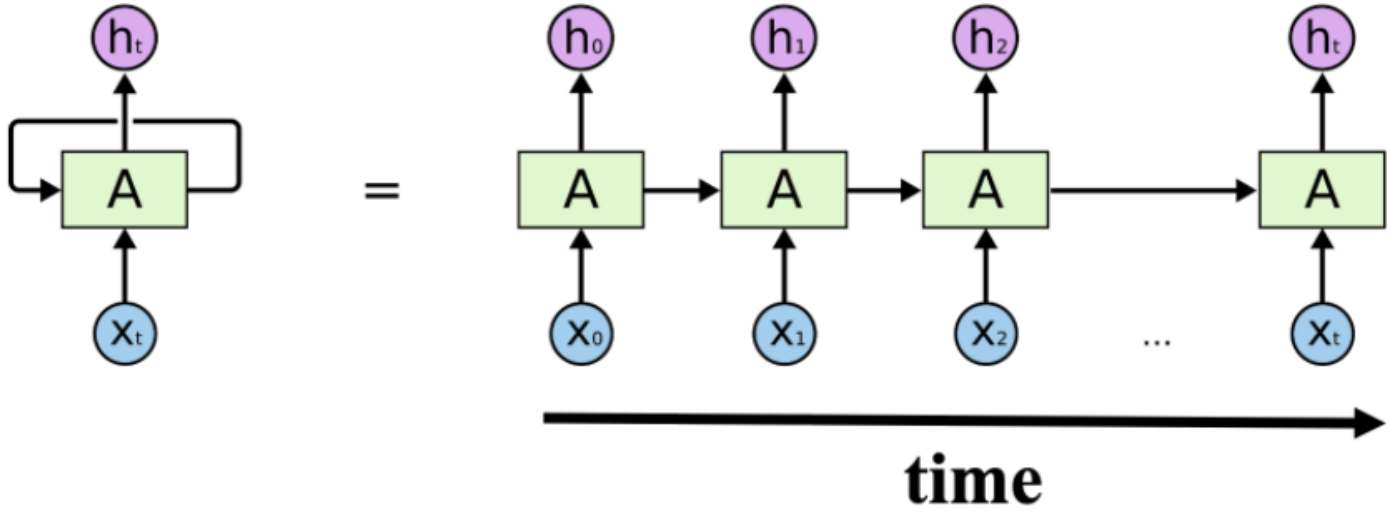


Simple recurrent network

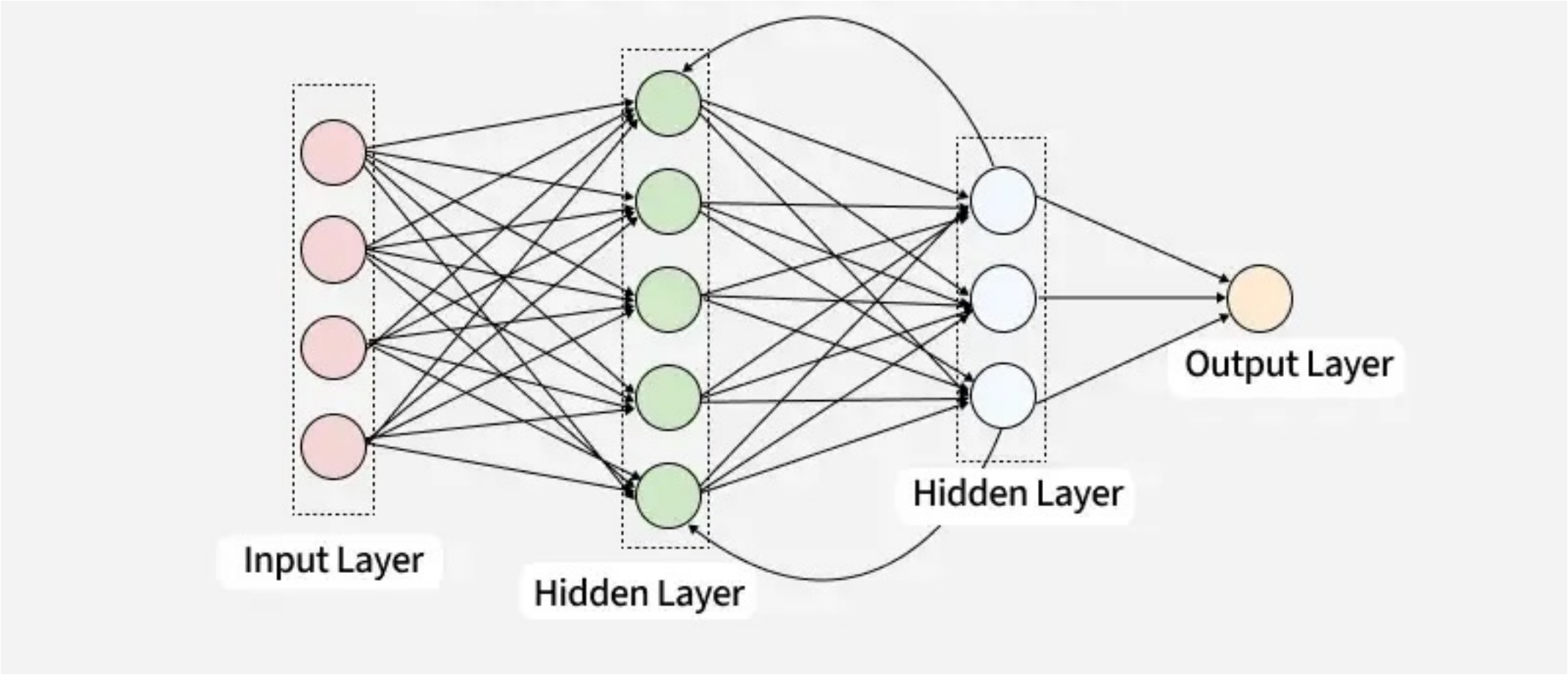




http://colah.github.io/posts/2015-08-Understanding-LSTMs/



Recurrent neural network

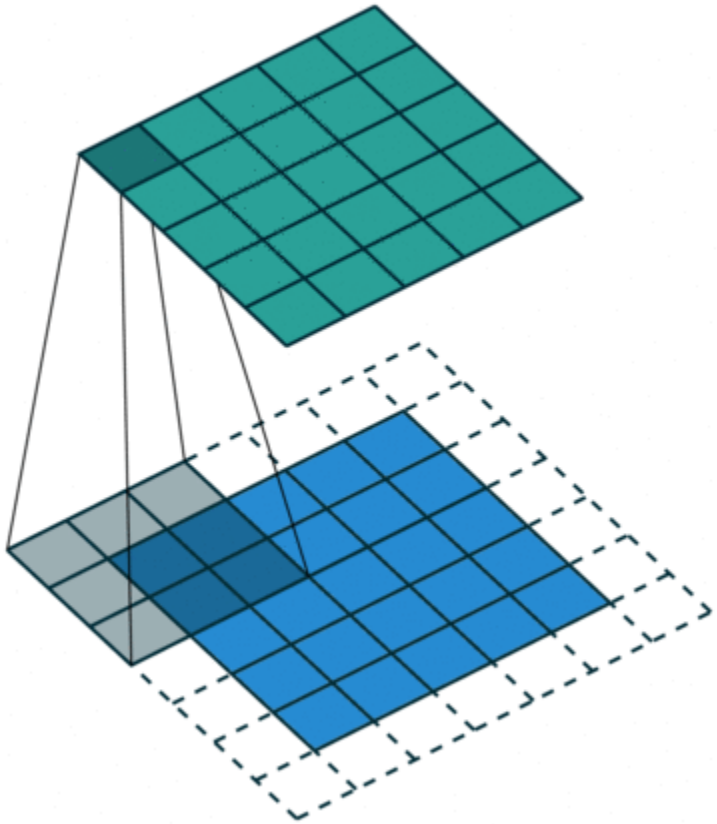


Convolutional Neural Networks

What is a convolution

- Convolution is applying a **kernel** (filter) over data (text, image, etc.)
- The kernel (filter) defines which **feature** is important in the data

What is a convolution

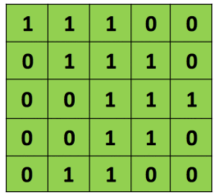


https://github.com/vdumoulin/conv_arithmetic

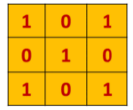
What is a convolution

1x1	1x0	1x1	0	0
0x0	1x1	1x0	1	0
0x1	0x0	1x1	1	1
0	0	1	1	0
0	1	1	0	0

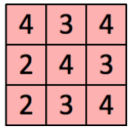
4		



5x5 input.



3x3 filter/kernel/feature detector.



3x3 convolved feature/
activation map/feature map

Convolution layers

- A convolutional neural network is a NN with one or more **convolution layers**
- The parameters / weights in a convolution layer are the elements of the filter
- The filter is **learnt** by the network!

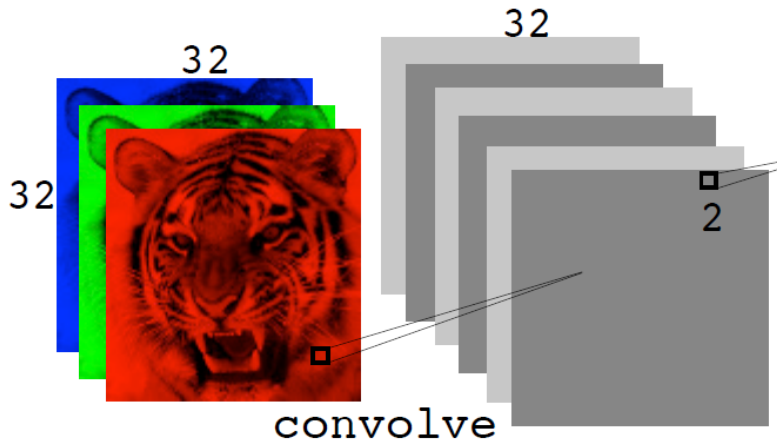


FIGURE 10.8. *Architectu*
Convolution layers are inte
size by a factor of 2 in both

Pooling layer

- Convolution layers are usually followed by a **pooling layer**
- Reduces dimensionality
- **Location invariance**: Robustness against pixel shift / small rotations
- **Max pool** most common

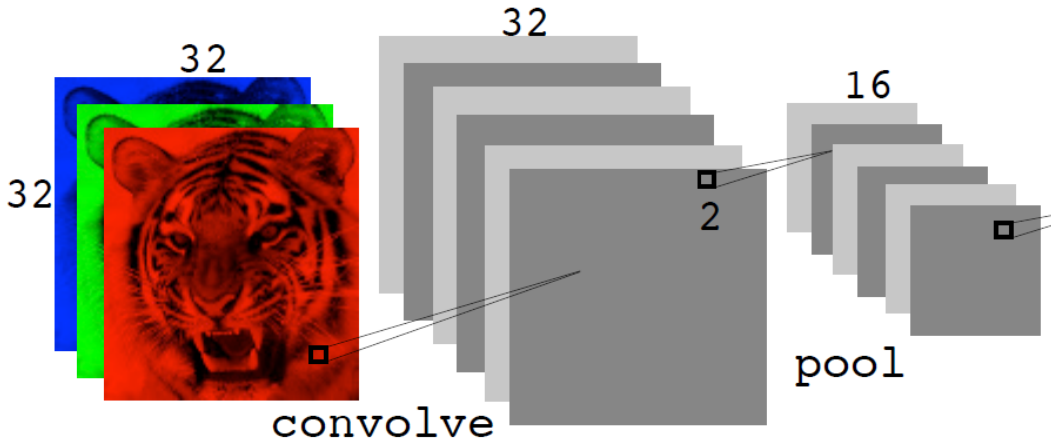


FIGURE 10.8. *Architecture of a de*
Convolution layers are interspersed u
size by a factor of 2 in both dimensio

Pooling layer

$$\text{Max pool} \begin{bmatrix} 1 & 2 & 5 & 3 \\ 3 & 0 & 1 & 2 \\ 2 & 1 & 3 & 4 \\ 1 & 1 & 2 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 3 & 5 \\ 2 & 4 \end{bmatrix}.$$

Architecture of a CNN

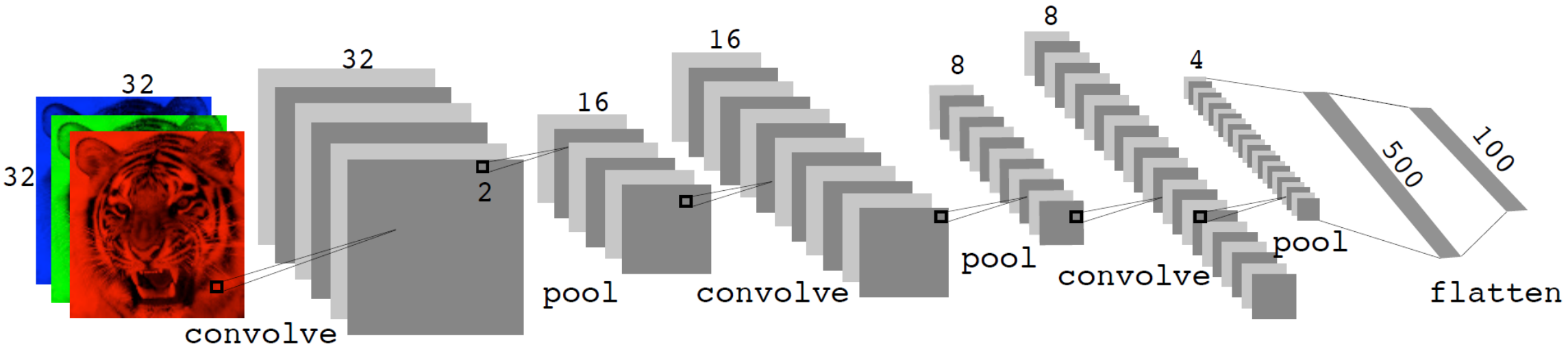
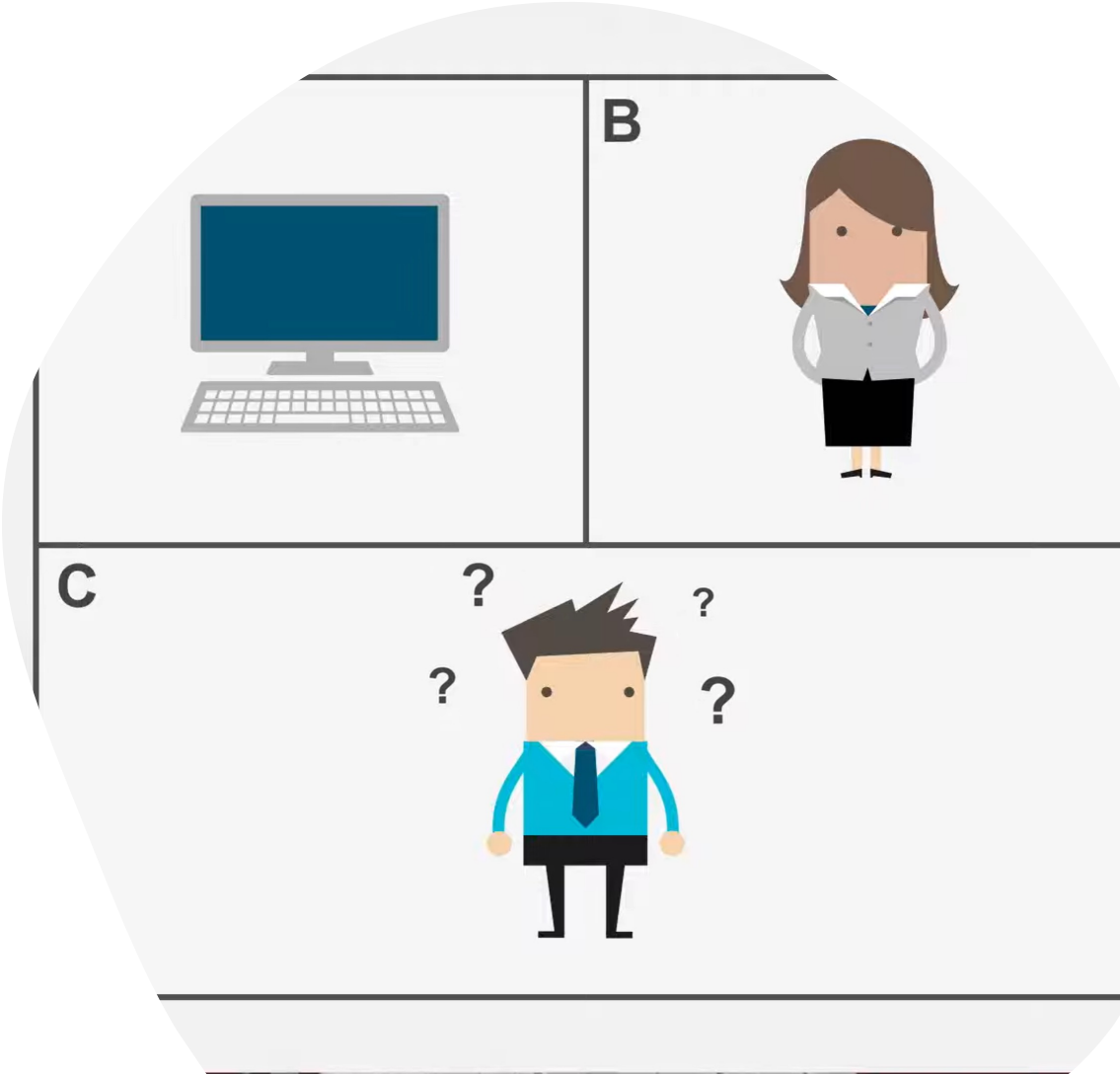


FIGURE 10.8. *Architecture of a deep CNN for the* `CIFAR100` *classification task. Convolution layers are interspersed with* 2×2 *max-pool layers, which reduce the size by a factor of 2 in both dimensions.*

Large Language Models


ONCE
UPON A TIME


B
C
?
?
?
?

Transformers!

Large Language Models

ChatGPT4o 

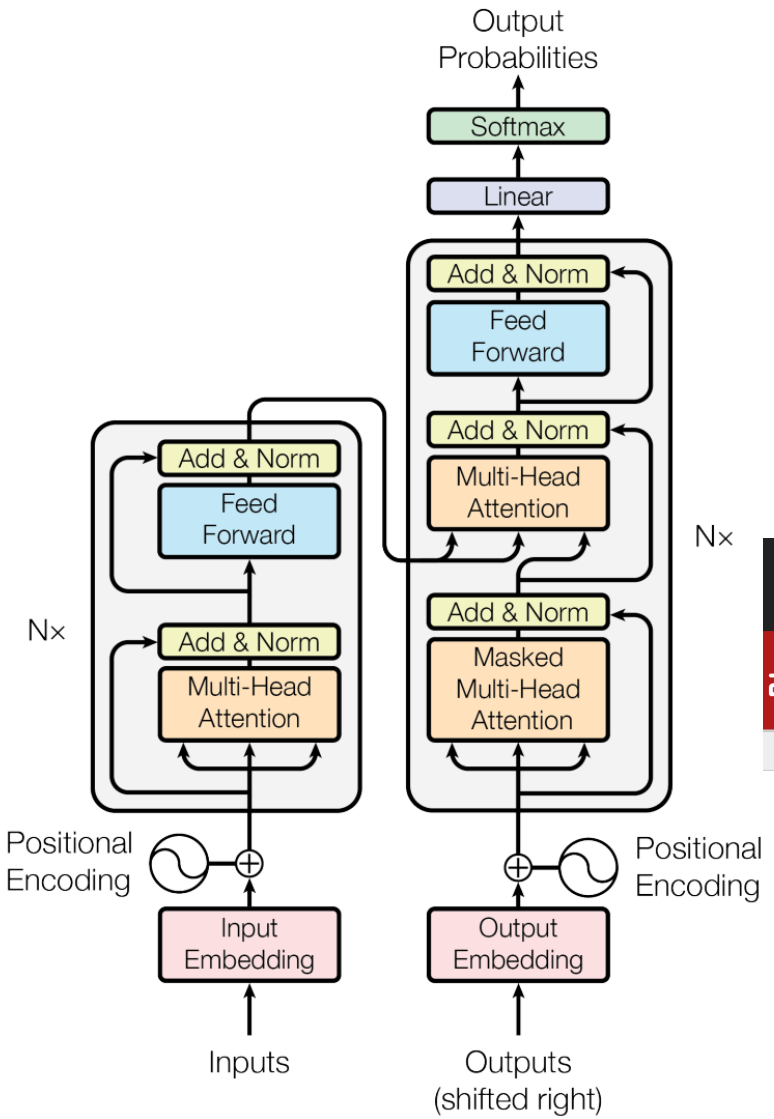
Quiz me on world capitals

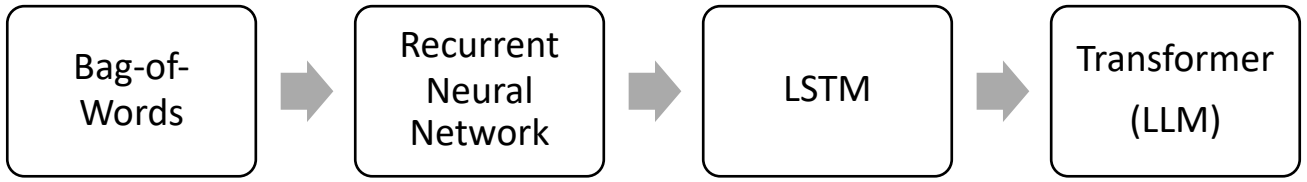
Message to comfort a friend

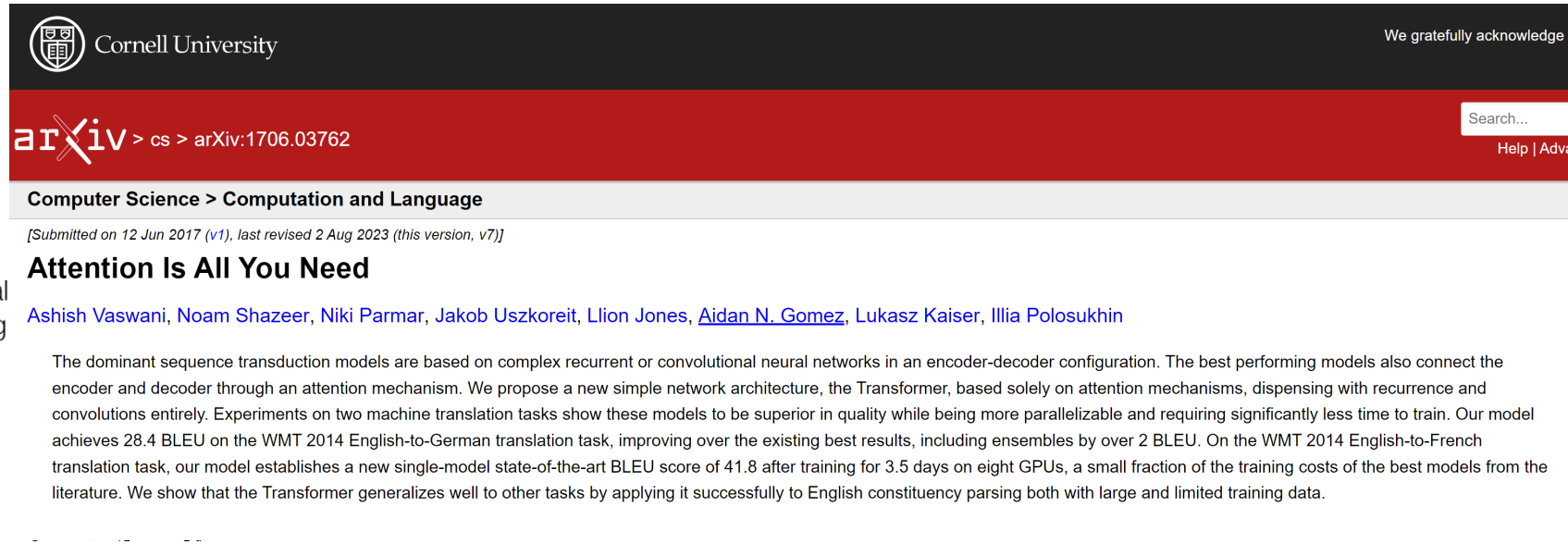
Activities to make friends in new city

Pick outfit to look good on camera

Transformers!







Cornell University

We gratefully acknowledge

arXiv > cs > arXiv:1706.03762

Search...

Help | Adva

Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

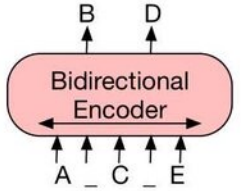
Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

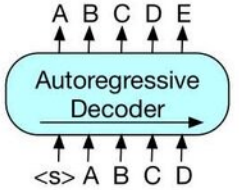
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Transformer foundation models: BERT, GPT, BART

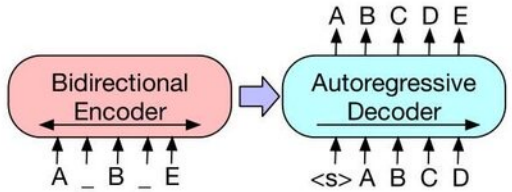
- **BERT**: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers.
  - *Masked word prediction, text representation*
- **GPT**: **G**enerative **P**re-trained **T**ransformer.
  - *Next word prediction, text generation, chat*
- **BART** = "BERT+GPT": **B**idirectional encoder and **A**uto-**R**egressive decoder **T**ransformers.
  - *Noised text reconstruction, summarization, translation, spelling correction*



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



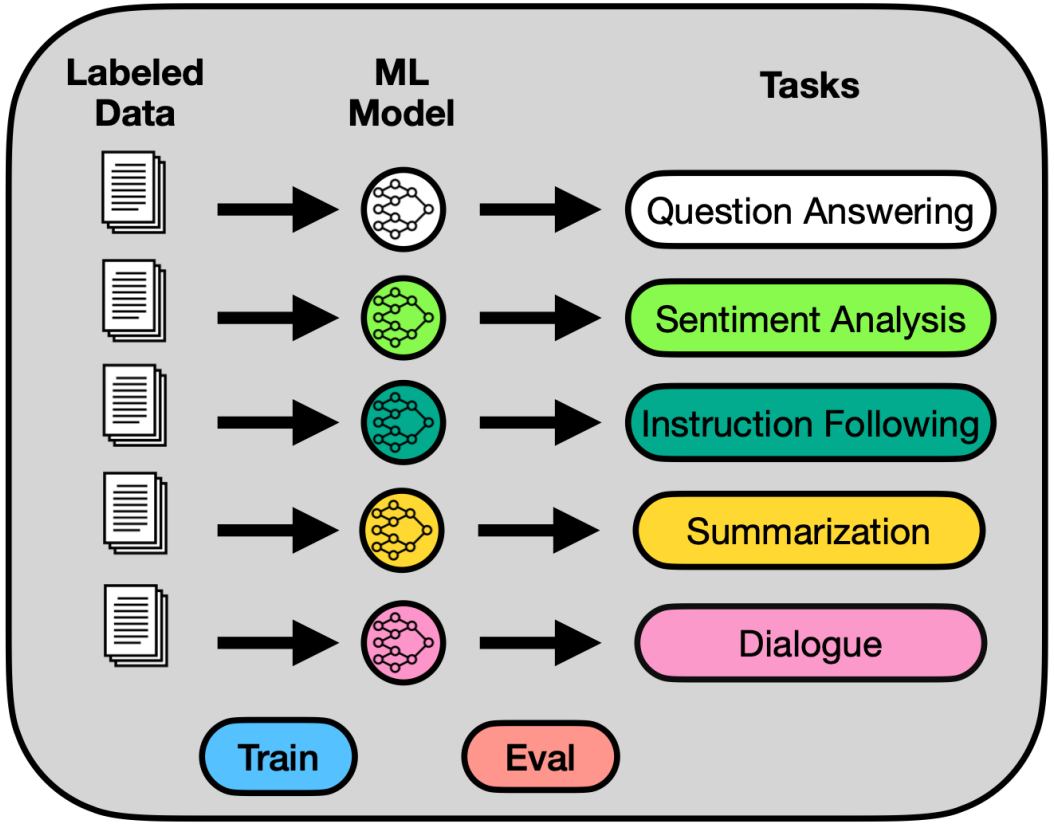
(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

"A foundation model is any model that is trained on broad data that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks"

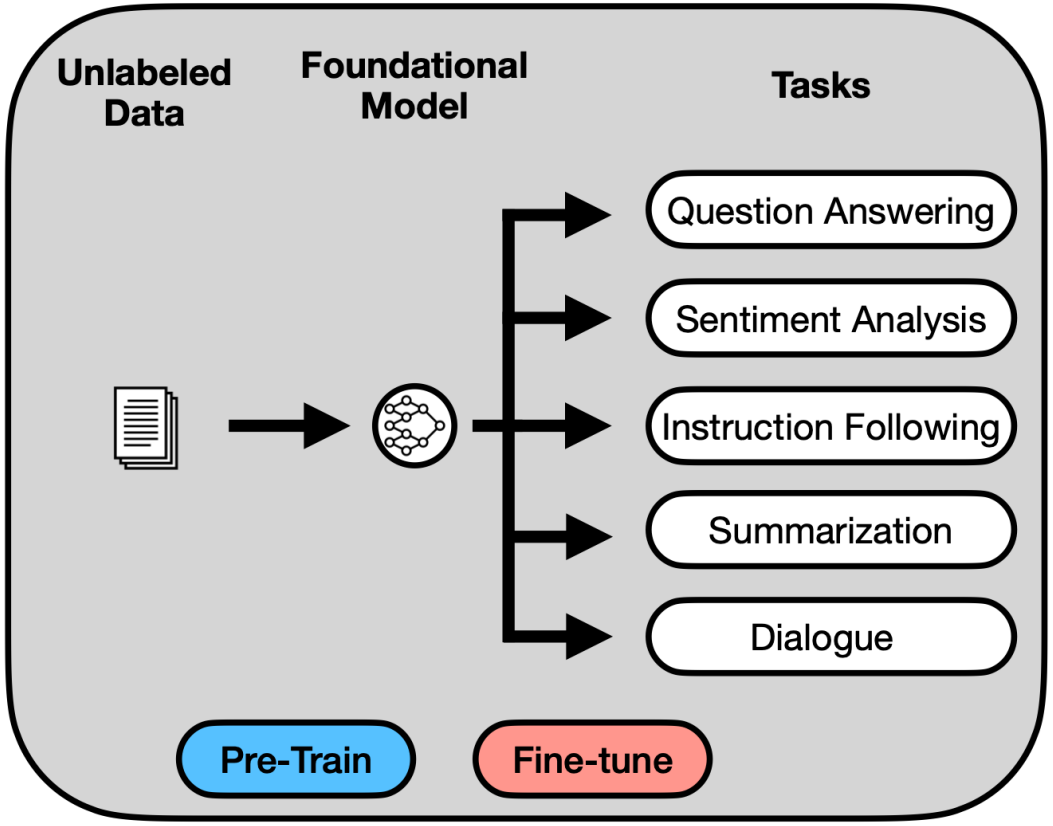
Foundational Models

On the Opportunities and Risks of Foundation Models by Bommasani et al. 2022
https://arxiv.org/pdf/2108.07258

Foundational Models



Traditional Machine Learning Models



Foundational Models

Foundational Models

> 70B params

Large Language Model

$$p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_1, y_3)\ldots = \prod_{i=1}^{n} p(y_i \mid y_1, \ldots, y_{i-1})$$

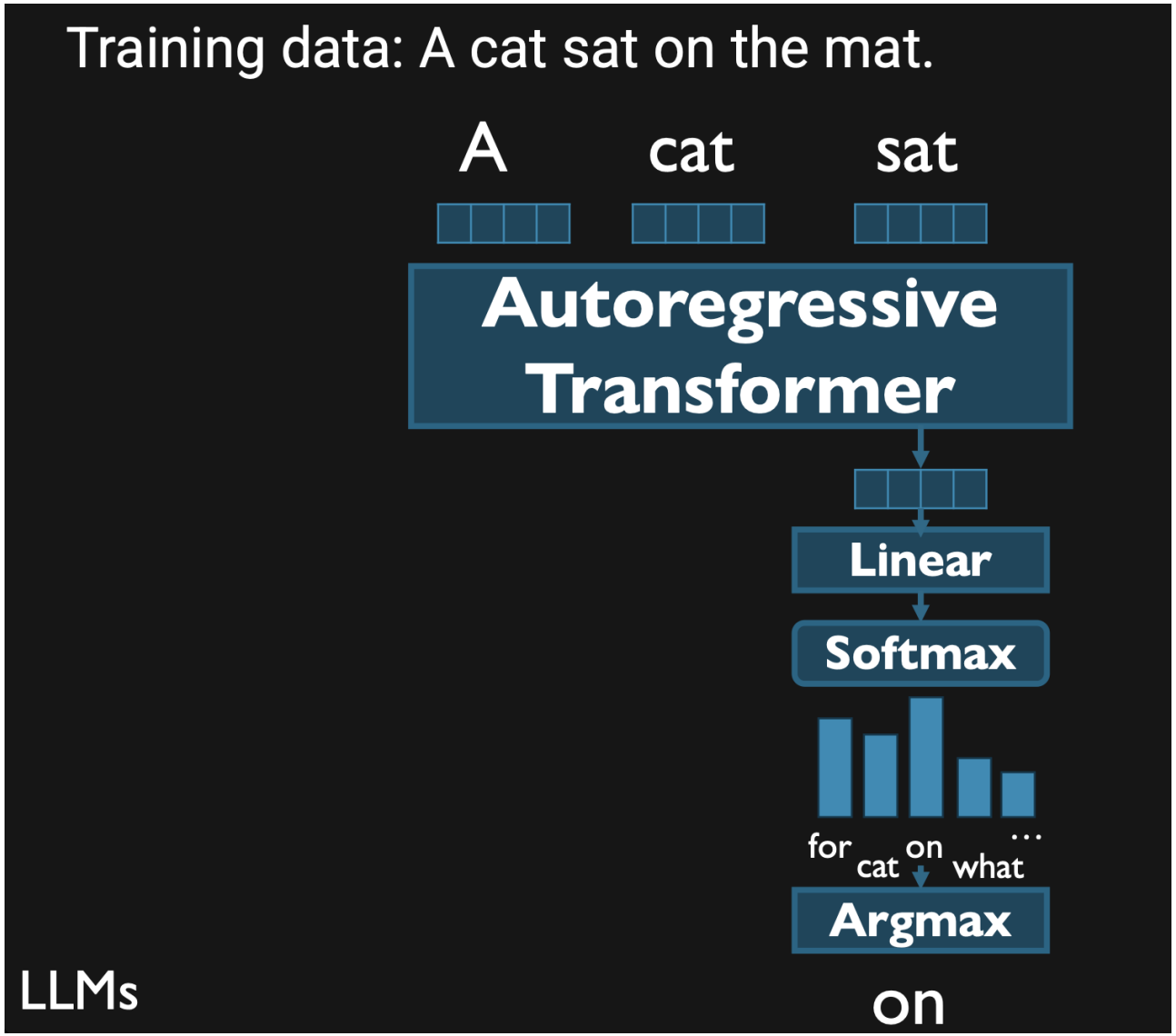
for example,

context

$p(\cdot \mid \text{you, only, live})$

predict next word

Foundational Models



Foundational Models

Large Language Model

$$p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_1, y_3)\ldots = \prod_{i=1}^{n} p(y_i \mid y_1, \ldots, y_{i-1})$$

Pre-Training

Large Unlabeled Dataset

Foundational Models

Large Language Model

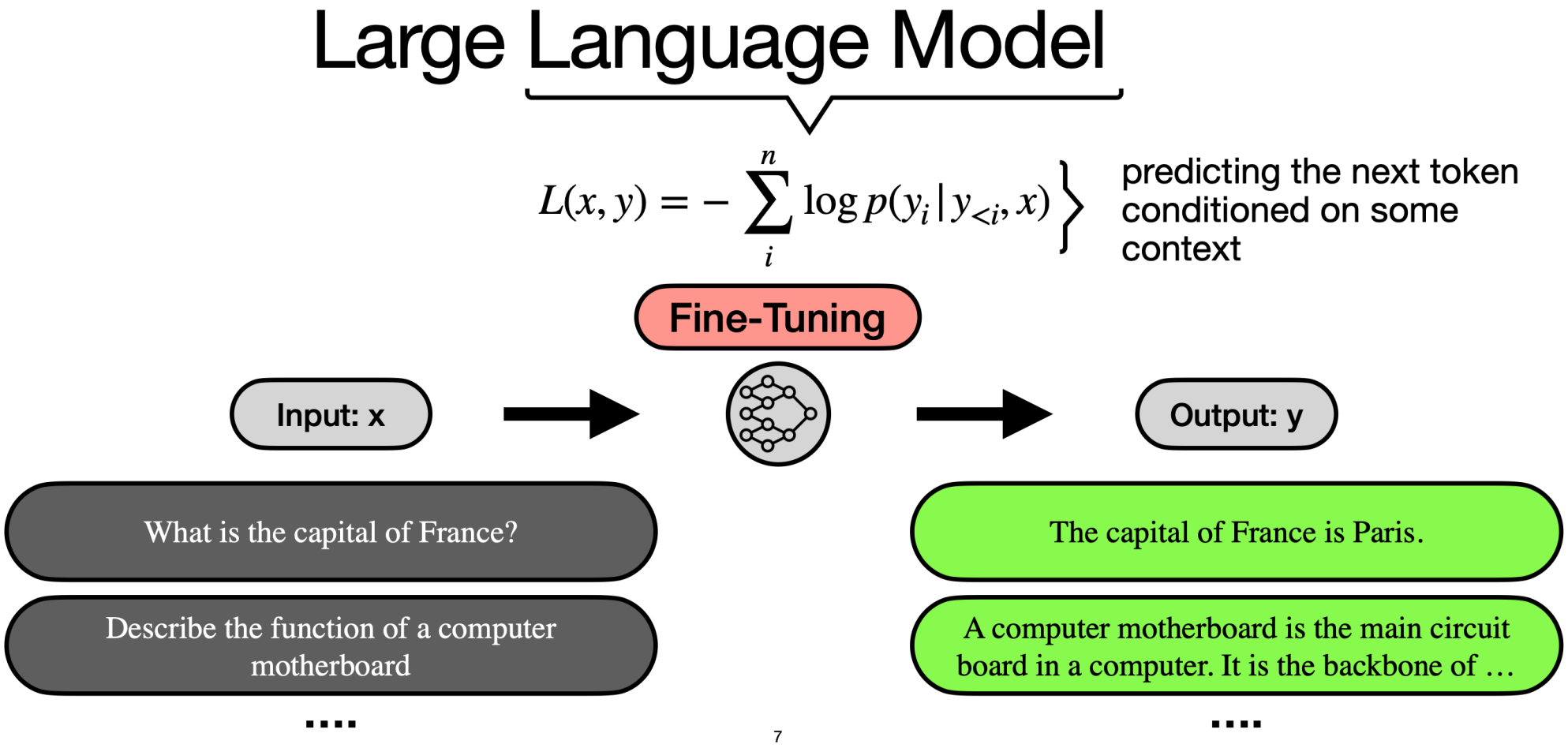
$$L(y) = -\sum_{i}^{n} \log p(y_i \mid y_{<i})$$

next word prediction

Pre-Training

Large Unlabeled Dataset

Foundational Models



Kianté Brantley: https://xkianteb.github.io

Foundational Models

Prompt: This is what I can build from a few sticks to defend against a bear:

GPT-2 (2019)

1. I can build a bear with a stick.
2. I can build a bear with a stick.
3. I can build a bear with a stick.

ChatGPT 4o (2024)

Spiked Barrier Trap

Build the Frame:
Take four sturdy sticks of roughly the same length. Arrange them into a square or rectangular base on the ground. Bind the corners together securely using vines or flexible branches to create a stable frame.

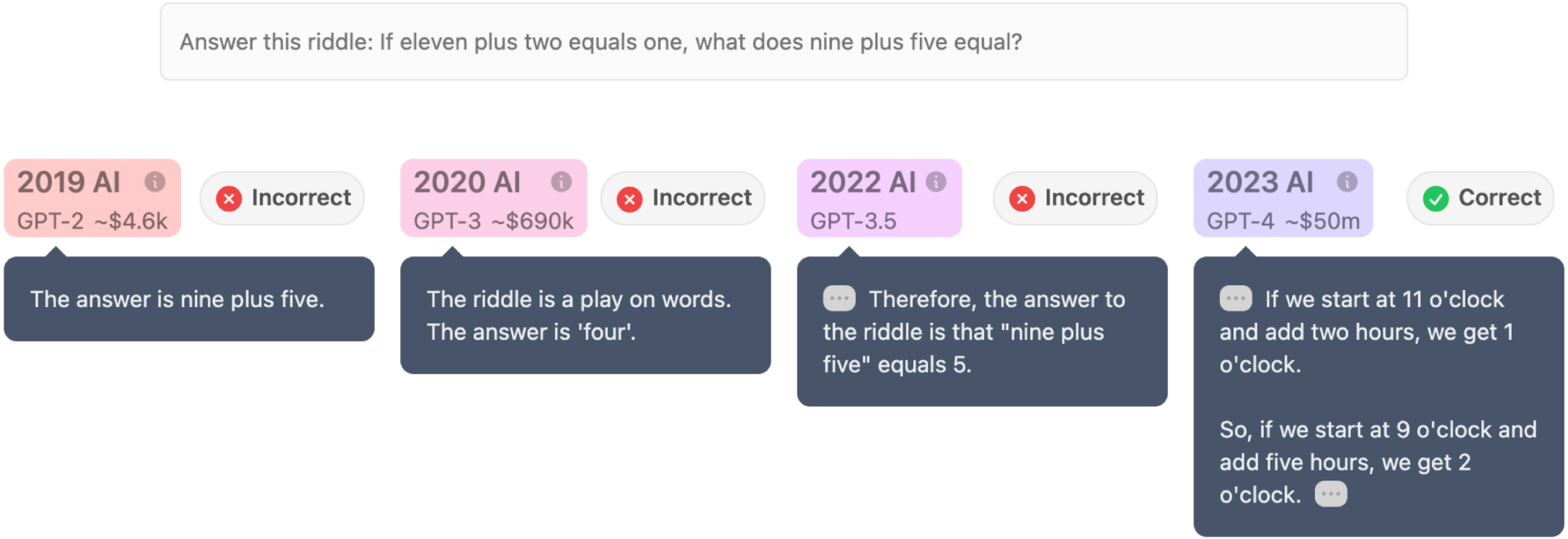
Add Upright Supports:
Attach two more sticks vertically at opposite corners of the base, forming an upright frame. Bind them tightly to the base so they don't wobble.

Create the Spikes:
…

Reinforce the Structure:
…

Camouflage the Trap (Optional):
…

Foundational Models



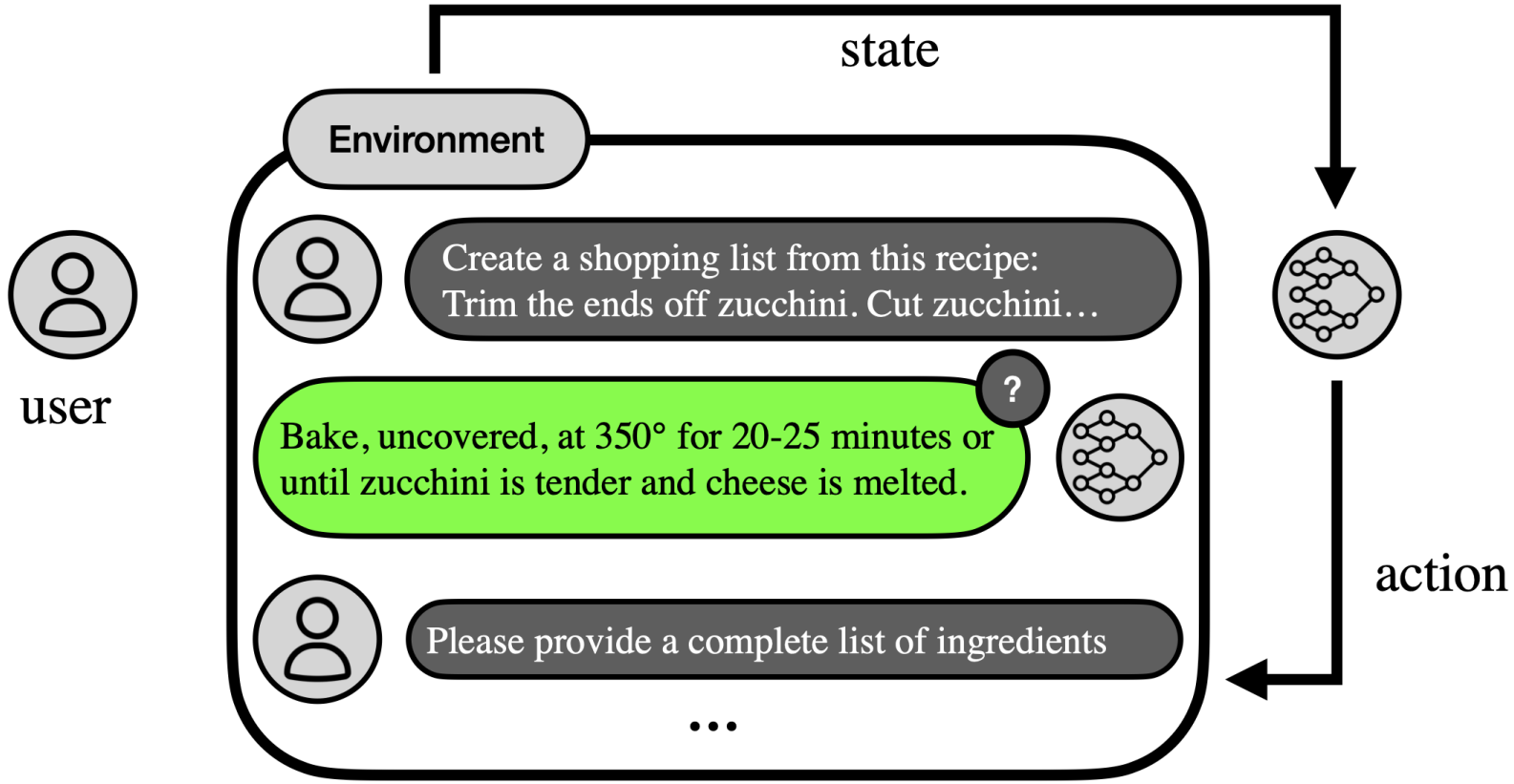
https://www.reddit.com/r/ChatGPT/comments/17h0hyr/i_made_a_site_where_you_can_ask_the_same_question/

Foundational Models: Problems

"Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users."

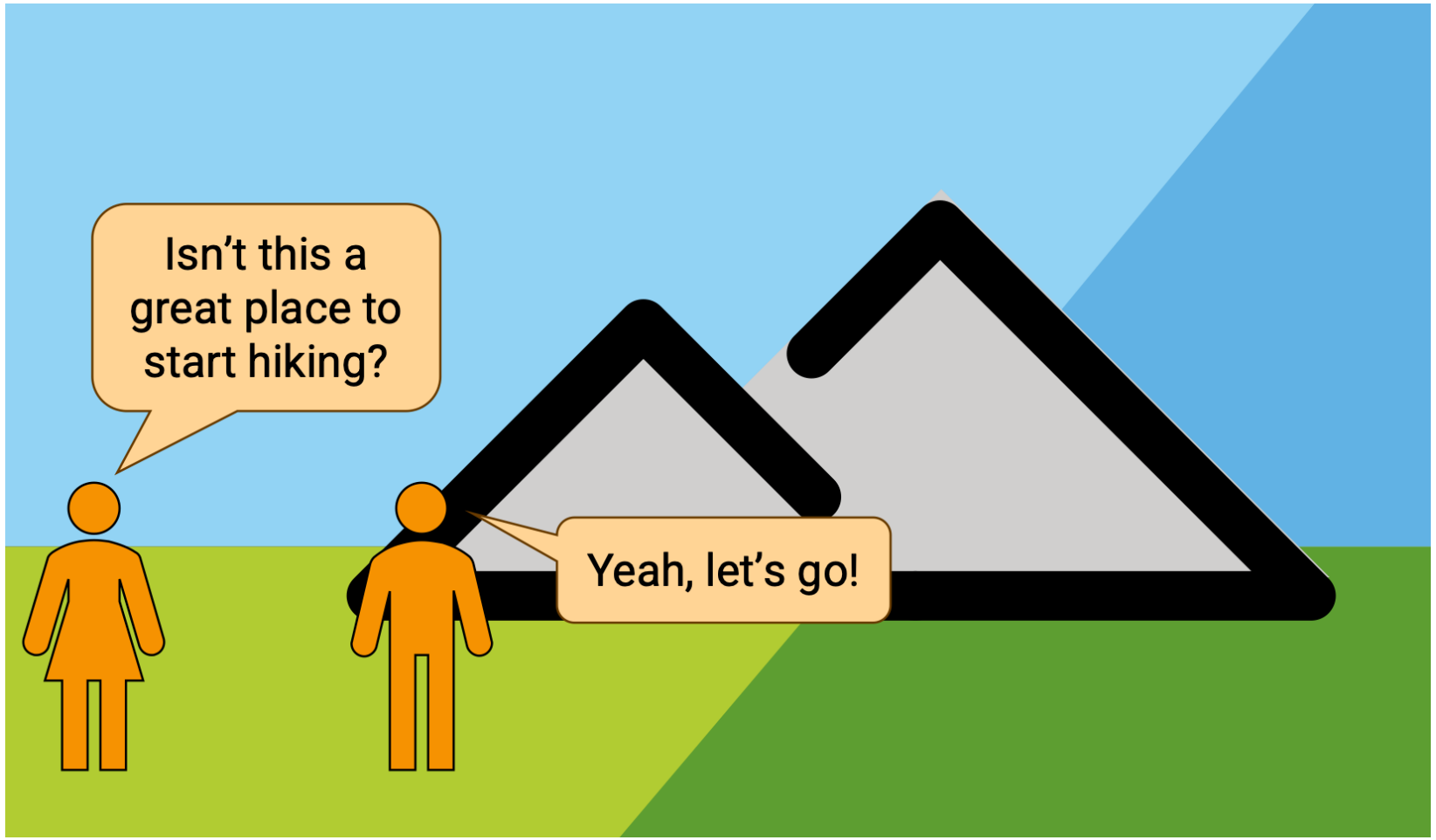
Long Ouyang et al.
Training language models to follow instructions with human feedback
OpenAI 2022

Foundational Models: improvements

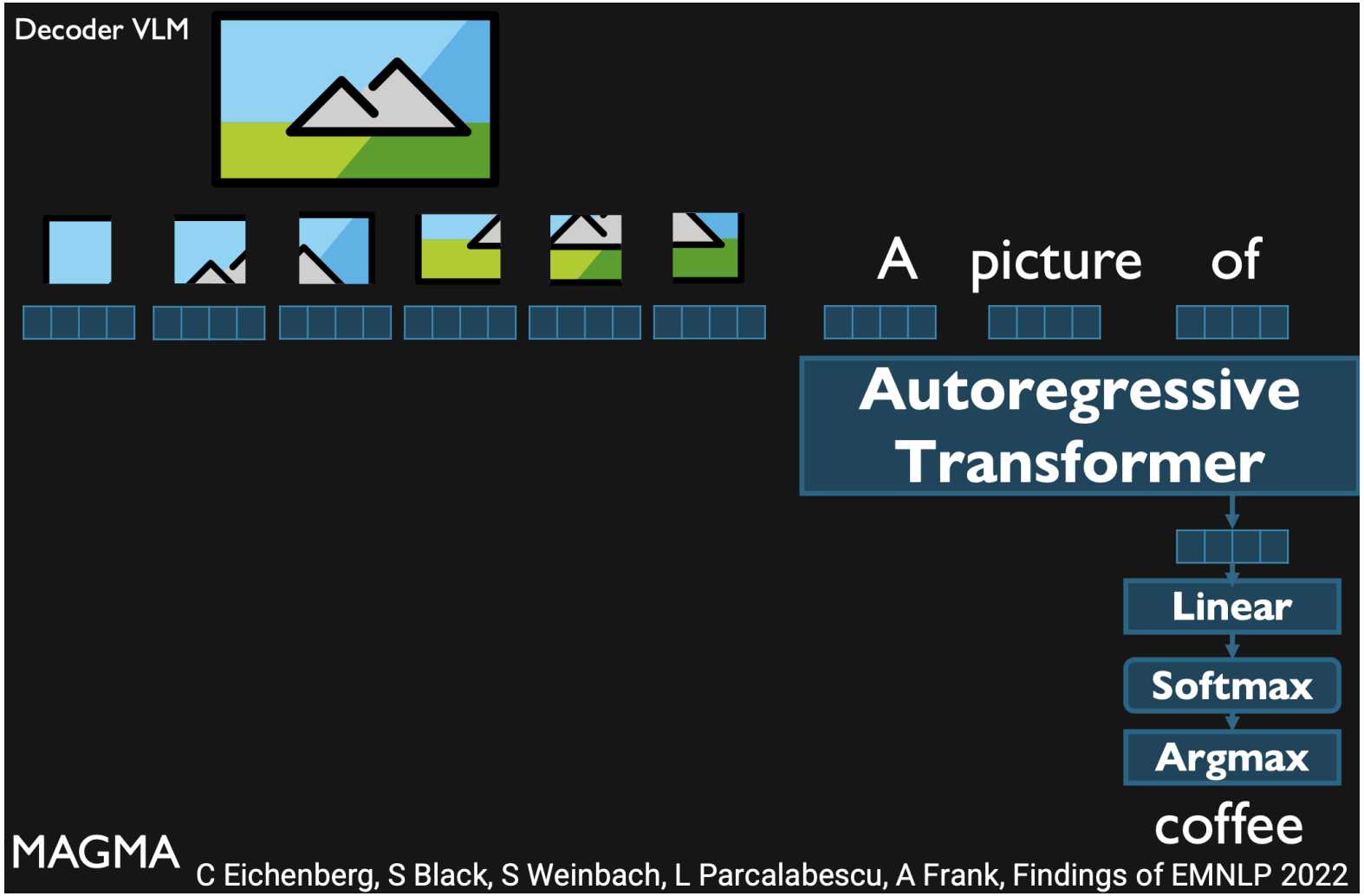


Training language models to follow instructions with human feedback Ouyang et al. 2022

Foundational Models: improvements



Foundational Models: improvements



Dutch language models

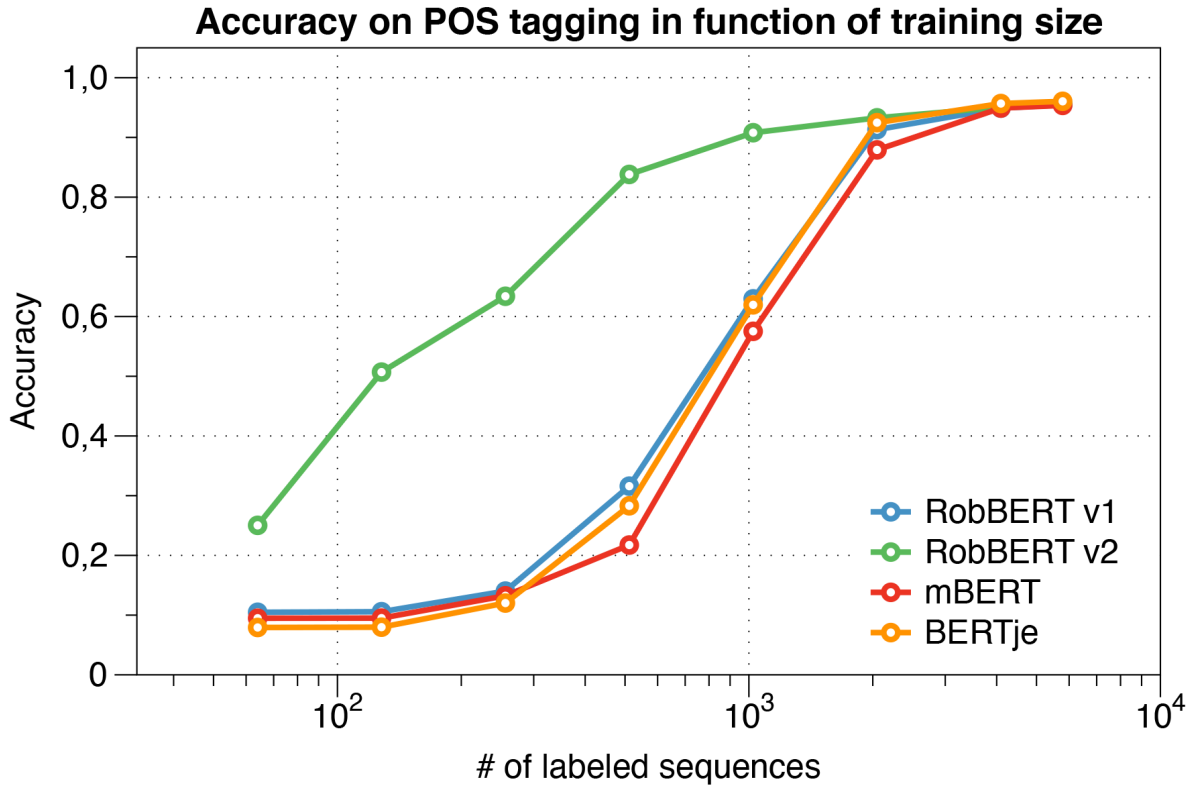
RobBERT

A Dutch RoBERTa-based Language Model

RobBERT: Dutch RoBERTa-based Language Model.

RobBERT is the state-of-the-art Dutch BERT model. It is a large pre-trained general Dutch language model that can be fine-tuned on a given dataset to perform any text classification, regression or token-tagging task. As such, it has been successfully used by many researchers and practitioners for achieving state-of-the-art performance for a wide range of Dutch natural language processing tasks, including:

- Emotion detection
- Sentiment analysis (book reviews, news articles*)
- Coreference resolution
- Named entity recognition (CoNLL, job titles*, SoNaR)
- Part-of-speech tagging (Small UD Lassy, CGN)
- Zero-shot word prediction
- Humor detection
- Cyberbulling detection



Conclusion

- Neural networks are popular methods especially for text mining
- Feed-forward & RNN & CNN
- RNN works better for text data
- Large Language Models such as GPT are based on RNN and attention deep learning layer.

Practical 7

Questions?