Spoken language understanding: from RNNs to Transformers

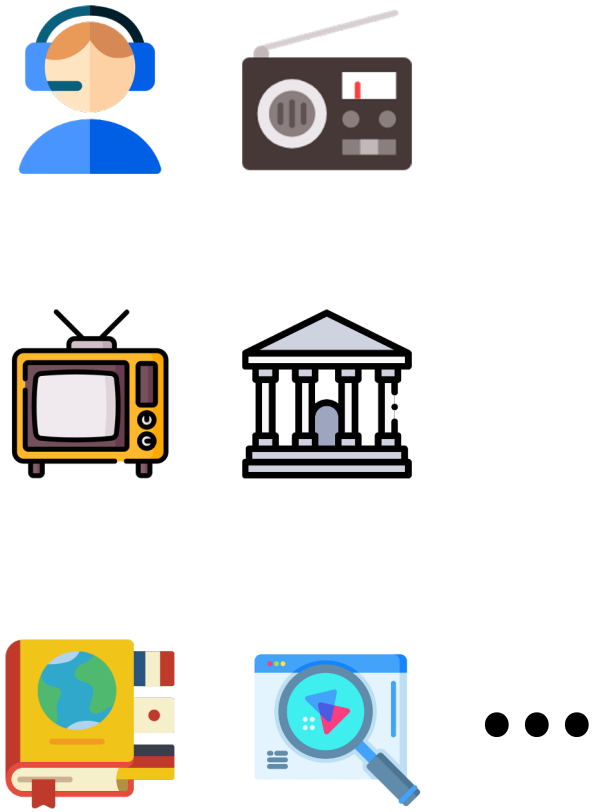
Pablo Ortiz, PhD

Senior Research Scientist, AI group, Telenor Research

pablo.ortiz@telenor.com

NTNU, 24.11.2021

ASR in “real life”



Interruptions
False starts
Noises, laughing, etc
Ungrammatical sentences
Mumbling
Lack of context
Slang and foreign words
Dialects
Several written standards
…

Agenda

1. Introduction and motivation
2. RNNs for sequence-to-sequence problems
    1. (Neural) Machine Translation (NMT)
    2. Automatic Speech Recognition (ASR)
3. Attention mechanisms

PAUSE

1. The Transformer
    1. Attention is all you need
    2. BERT
    3. GPT
    4. Applications: ASR, computer vision

2. RNNs for seq2seq problems: some references

- Sequence transduction with recurrent neural networks (1211.3711).
- Connectionist temporal classification: labelling unsegmented sequence data with RNNs (2006).
- Generating sequences with recurrent neural networks (1308.0850).
- Towards end-to-end speech recognition with recurrent neural networks (2014).
- **Learning phrase representations using RNN encoder-decoder for statistical machine translation (1406.1078).**
- Sequence to sequence learning with neural networks (1409.3215).
- On the properties of neural machine translation: encoder-decoder approaches (1409.1259).
- Gradient flow in recurrent nets: the difficulty of learning long-term dependencies (2001).
- An empirical study of smoothing techniques for language modelling (1999).
- Unsupervised feature learning for audio classification using convolutional deep belief networks (2009).
- First-pass large vocabulary continuous speech recognition using bidirectional recurrent DNNs (1408.2873).
- **Deep Speech (1412.5567 and 1512.02595).**

2.1 Intro: (n-gram) language models (LMs)

$$p(s) = p(w_1)p(w_2|w_1)p(w_3|w_1\,w_2)\cdots p(w_l|w_1\cdots w_{l-1}) = \prod_{i=1}^{l} p(w_i|w_1\cdots w_{i-1})$$

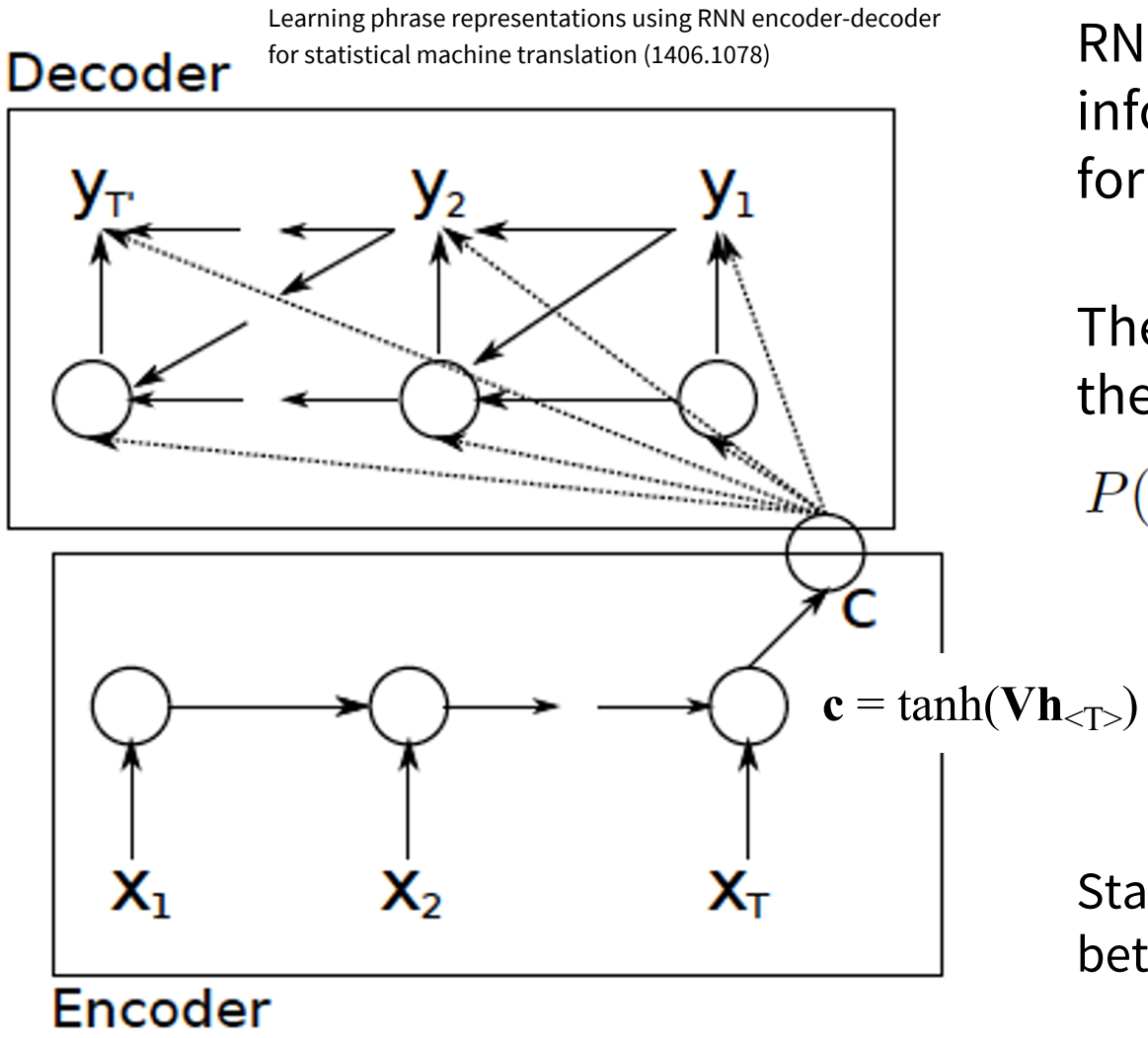
Jeg koser meg skikkelig på italiensk restaurant

3-gram approximation:

P(Jeg koser meg skikkelig på italiensk restaurant) =
= P(Jeg) P(koser|Jeg) P(meg|Jeg koser) P(skikkelig|koser meg) … P(restaurant|på italiensk)

What if P(skikkelig|koser meg) = 0 ??? **Smoothing** An empirical study of smoothing techniques for language modelling (1999).

2.1 NMT with RNN Encoder-Decoder



RNN units learn to pass along (or drop) information that is important (or irrelevant) for the future

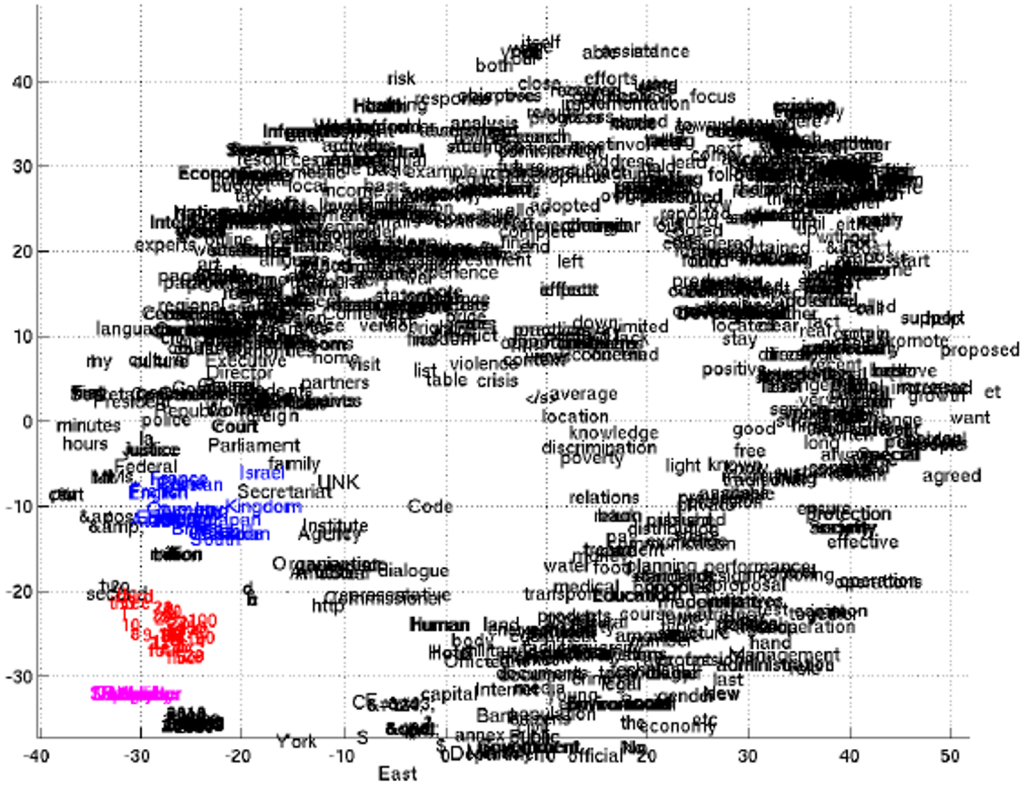
The decoder is like an LM conditioned on the input:

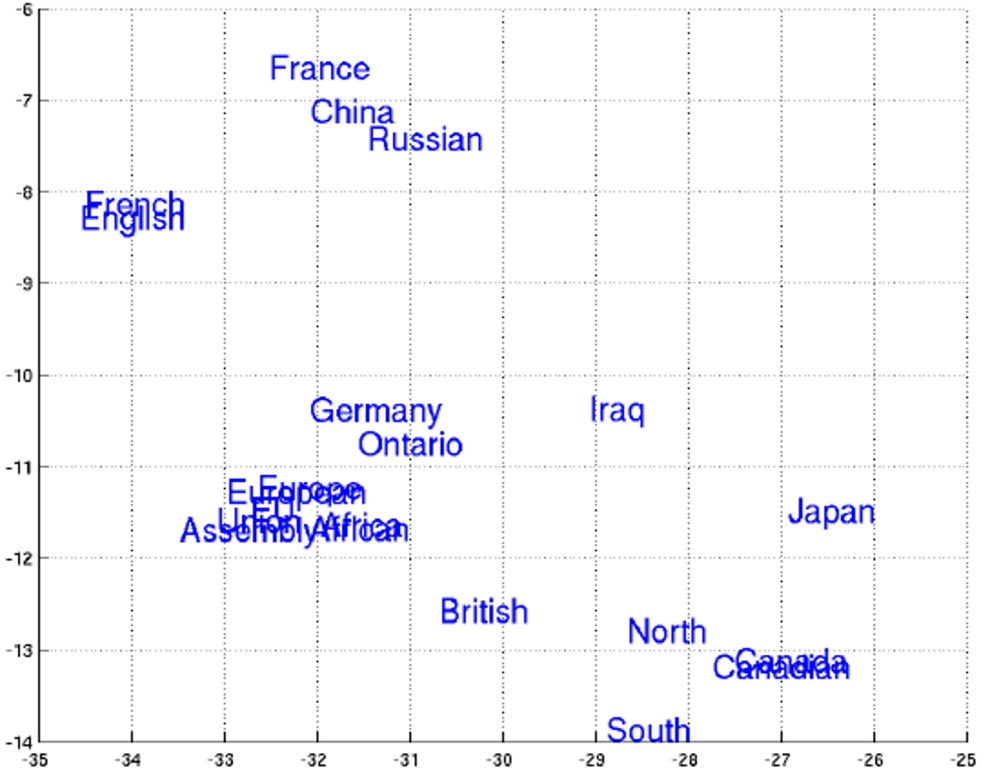
$$P(y_t|y_{t-1}, y_{t-2}, \ldots, y_1, \mathbf{c}) = g\left(\mathbf{h}_{\langle t \rangle}, y_{t-1}, \mathbf{c}\right)$$

$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}, \mathbf{c}\right)$$

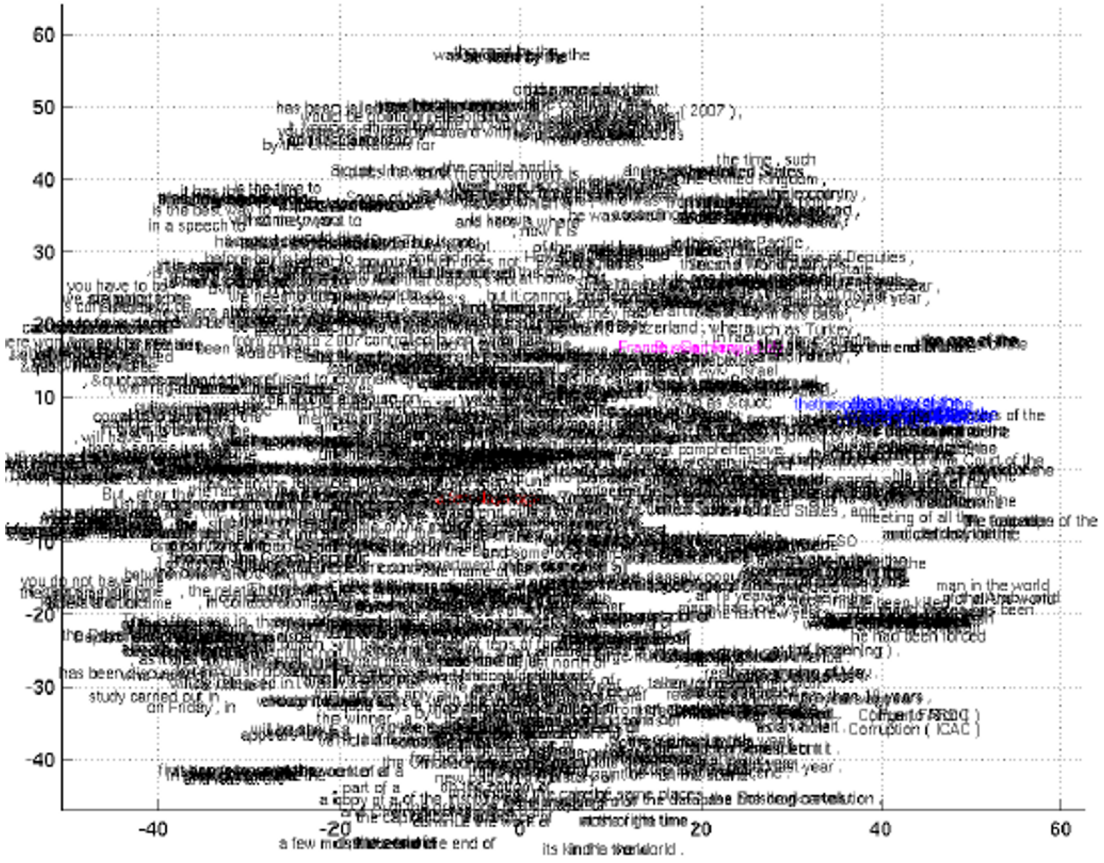
Stacks of LSTMs with inverted source work better (1409.3215)

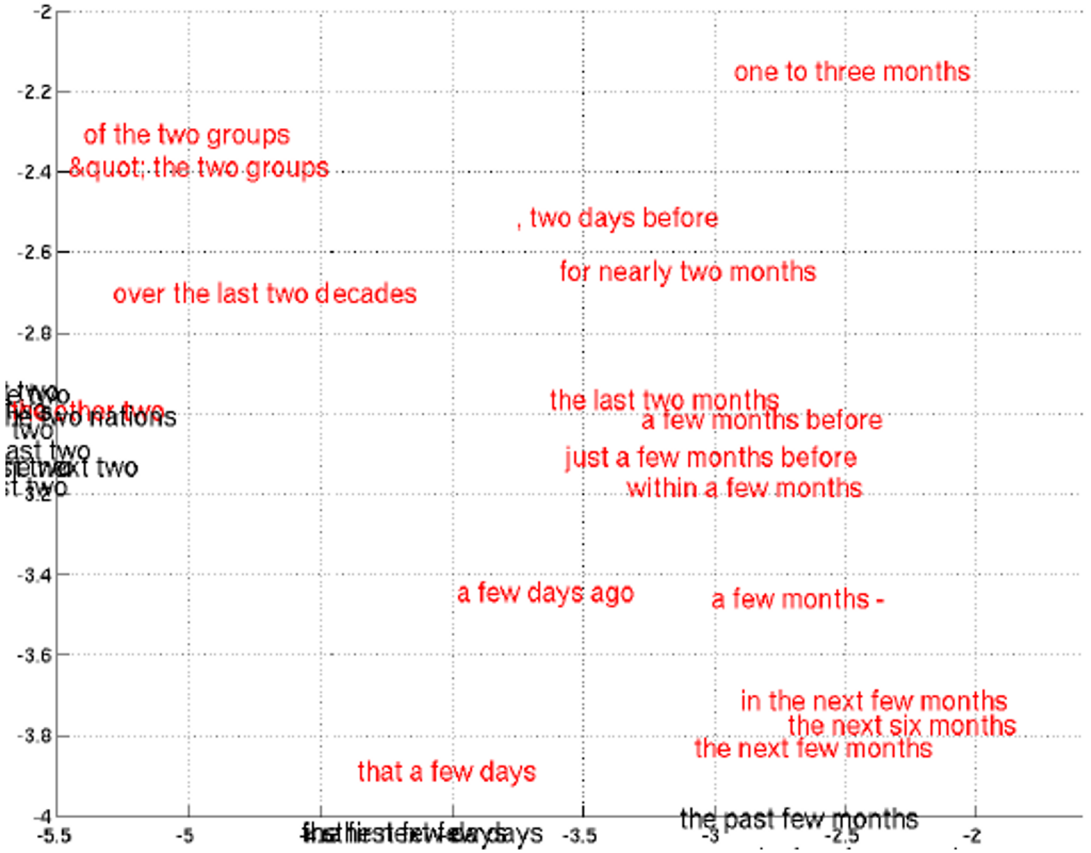
2.1 NMT: learned word representations





2.1 NMT: learned phrase representations





For reflection: syntactics, semantics and context.

ASR

2.2 ASR: aligning input and output

CTC

[Connectionist temporal classification: labelling unsegmented sequence data with RNNs (2006)]

$$p(Y \mid X) = \sum_{A \in \mathcal{A}_{X,Y}} \prod_{t=1}^{T} p_t(a_t \mid X)$$

The CTC conditional **probability** | **marginalizes** over the set of valid alignments | computing the **probability** for a single alignment step-by-step.

https://distill.pub/2017/ctc/

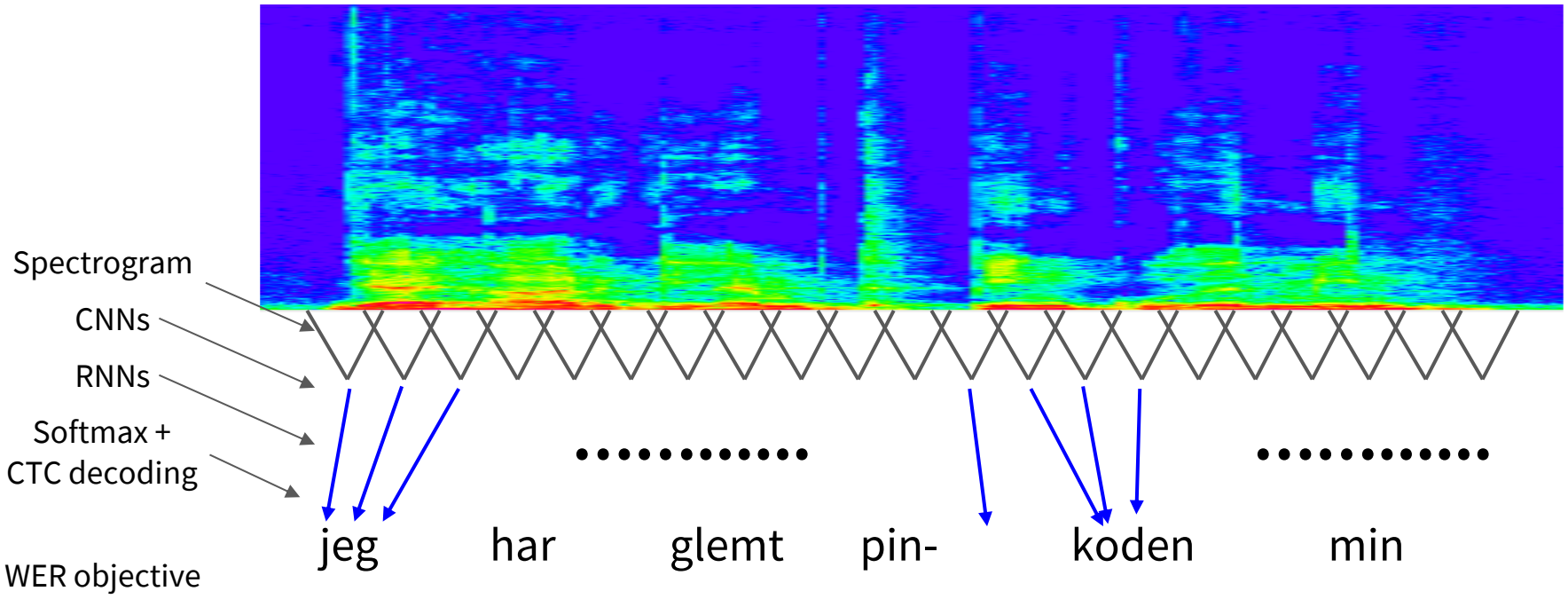
- L(output) < L(input)
- Assumes independence between outputs (doesn't learn an LM - good for transfer learning)
- Monotonic alignments (OK for ASR, not for MT)

RNN Transducer

Sequence transduction with recurrent neural networks (1211.3711)

- Output sequences of all lengths
- Models dependencies between outputs

2.2 ASR: full DS2 model

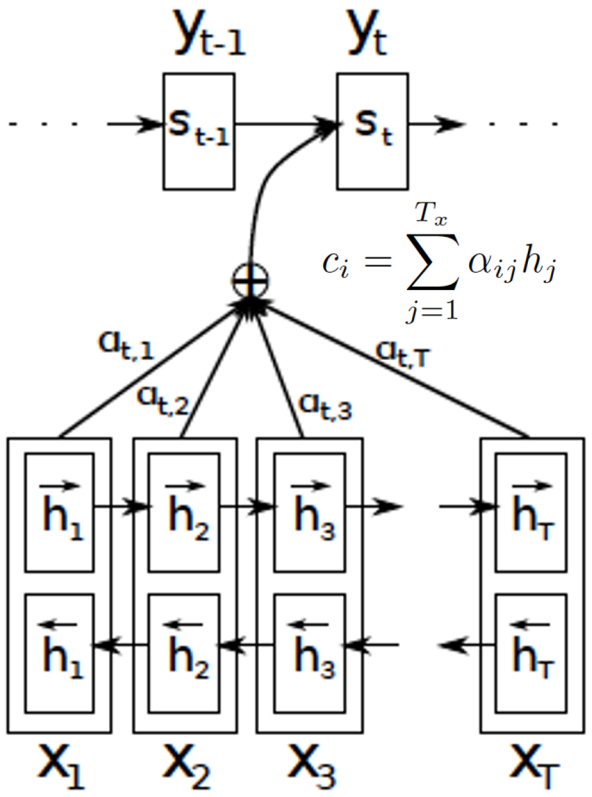


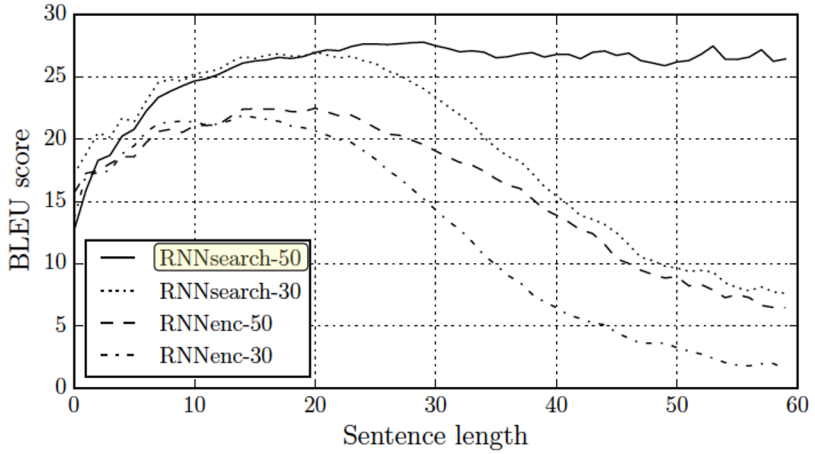
$$Q(c) = \log(\mathbb{P}(c|x)) + \alpha \log(\mathbb{P}_{\text{lm}}(c)) + \beta \text{ word\_count}(c)$$

Attention

3. Attention mechanisms

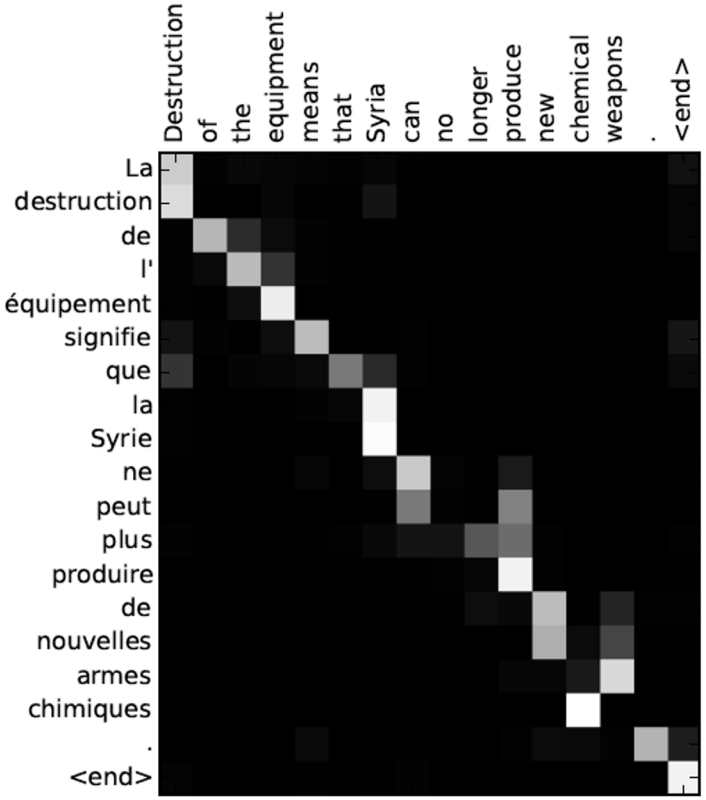
- **Neural machine translation by jointly learning to align and translate (1409.0473)**
- End-to-end continuous speech recognition using attention-based RNN: first results (1412.1602)
- End-to-end attention-based large vocabulary speech recognition (1508.0439)
- Attention-based models for speech recognition (1506.07503)
- Listen, Attend and Spell (1508.01211)





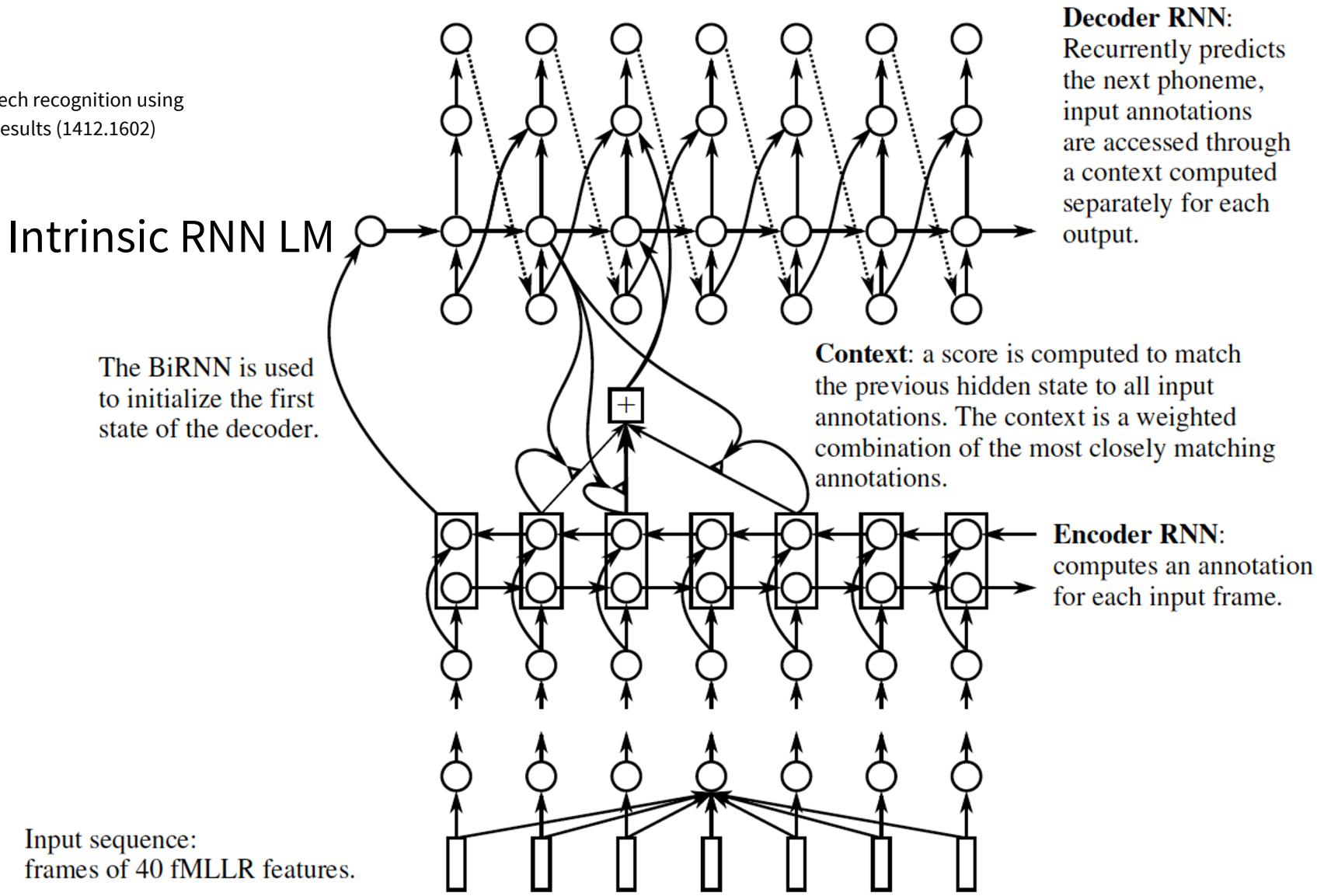
$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$



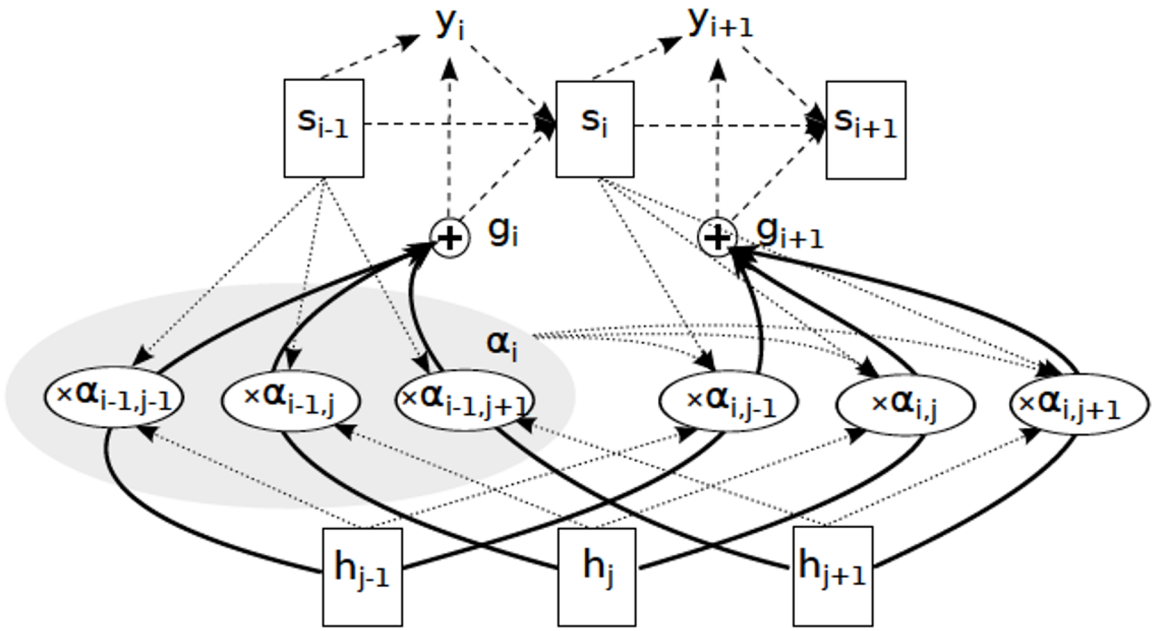
3. Attention mechanisms in ASR

End-to-end continuous speech recognition using attention-based RNN: first results (1412.1602)



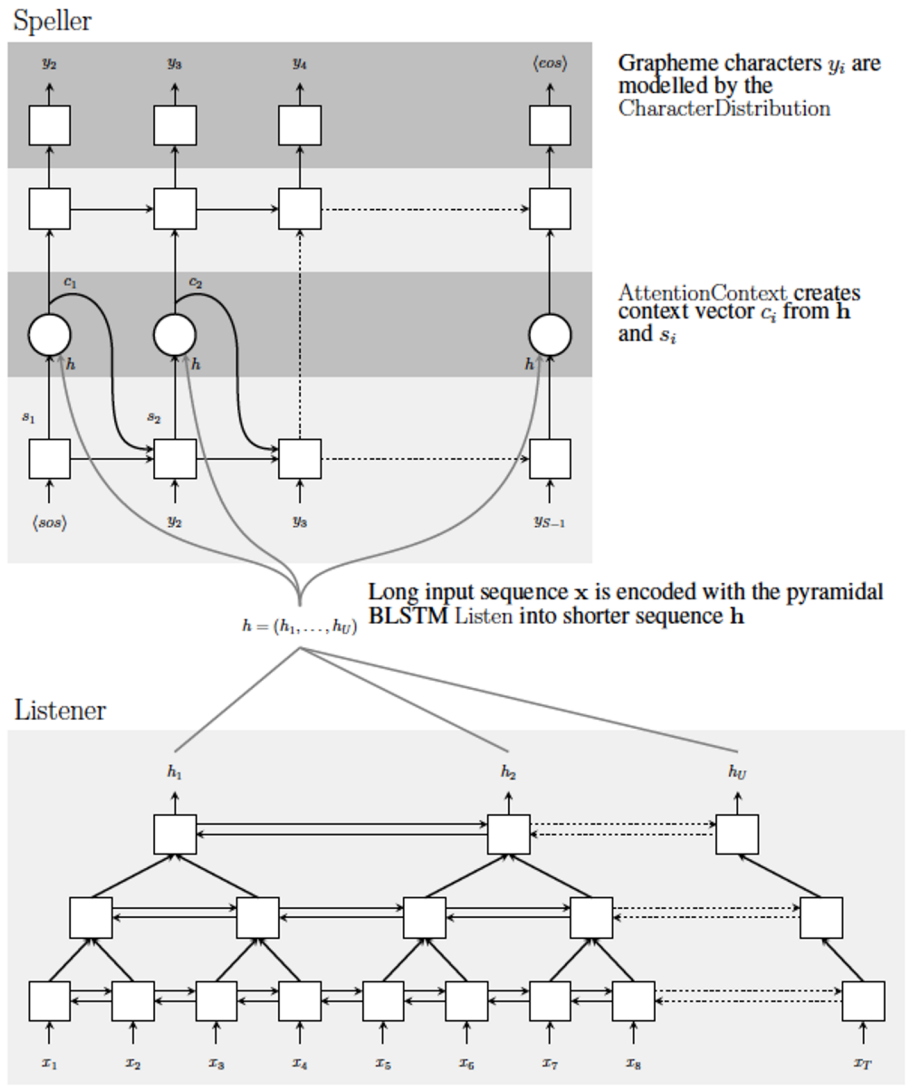
3. Attention in ASR 2.0

Attention-based models for speech recognition (1506.07503)



The encoder has limits capturing contextual information. To avoid the issue of similar speech fragments we use location-based attention to select a short list of frames and use content-based attention on them.

Listen, Attend and Spell (1508.01211)



Noisy predictions as training material for regularization during inference. Output characters for the first time in this type of models.

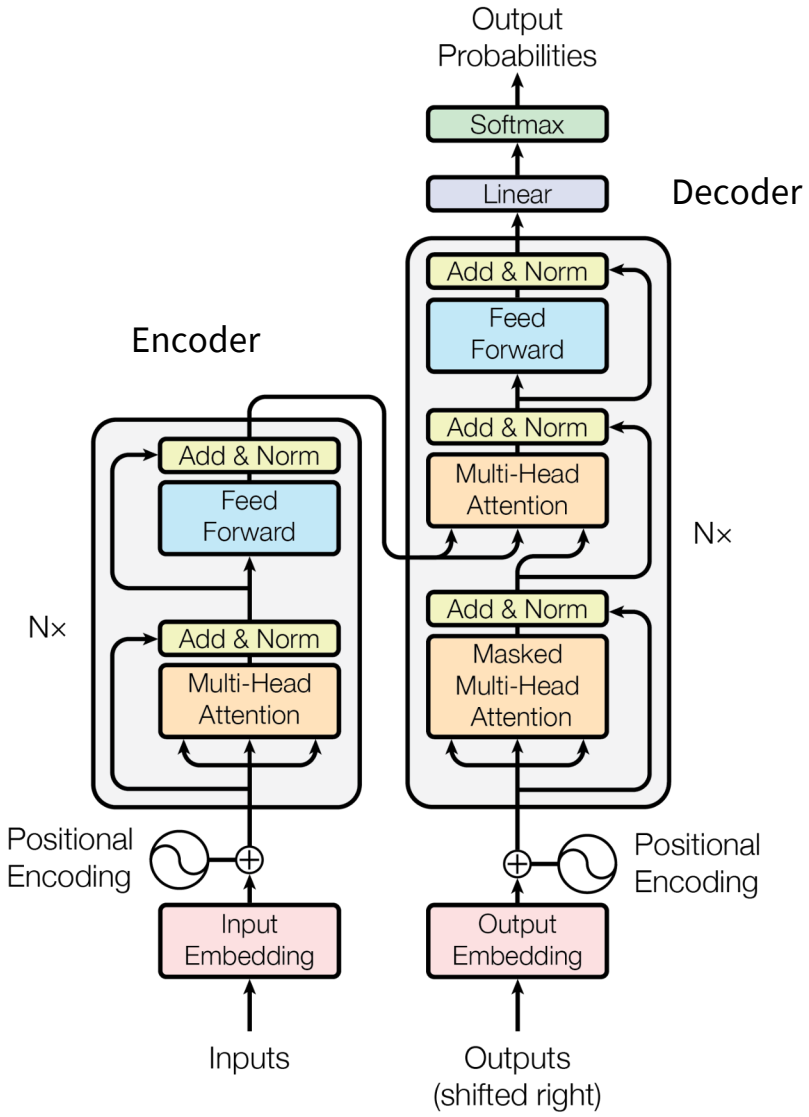
The Transformer

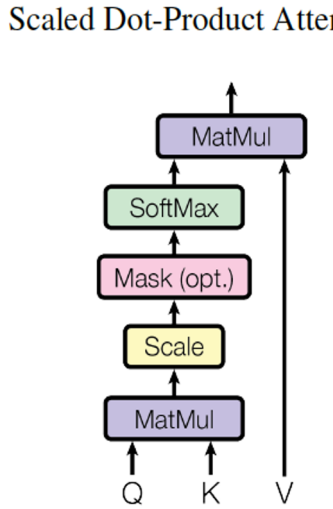
4. Transformer-based models

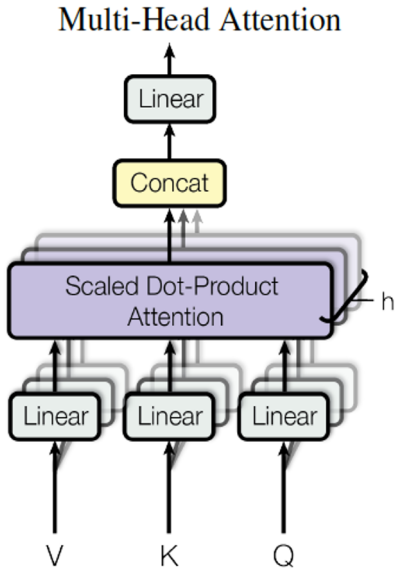
- **Attention is all you need (1706.03762)**
- Improving language understanding by generative pre-training (GPT) (2018)
- **BERT: pre-training of deep bidirectional transformers for language understanding (1810.04805)**
- Speech-transformer: a no-recurrence seq2seq model for speech recognition (2018)
- Transformer-XL: attentive language models beyond a fixed-length context (1901.02860)
- Analyzing the structure of attention in a transformer language model (1906.04284)
- Convolutional self-attention networks (2019)
- Language models are unsupervised multitask learners (GPT-2) (2019)
- **Generative pre-training from pixels (2020)** Take a look at Image GPT
- **Language models are few-shot learners (GPT-3) (2005.14165)**
- Conformer: convolution-augmented transformer for speech recognition (2005.08100)
- Convolutions and self-attention: re-interpreting relative positions in pre-trained language models (2106.05505)
- Bayesian attention belief networks (2106.05251)
- A comparative study on neural architectures and training methods for Japanese speech recognition (2106.05111)
- And hundreds keep on coming! Let's take a look at https://github.com/huggingface/transformers

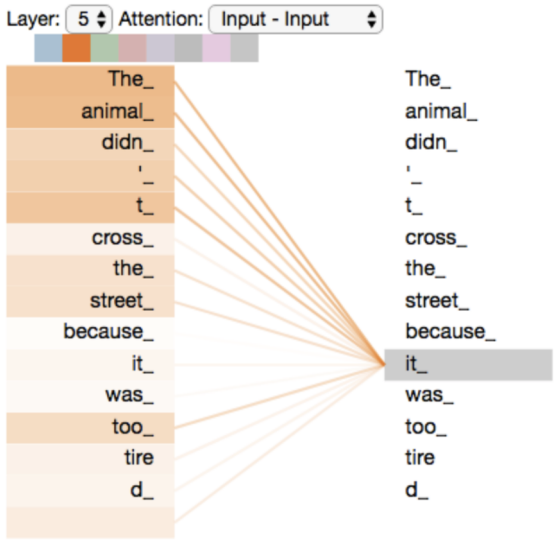
4.1. Attention is all you need

[The illustrated transformer]
[The annotated transformer]



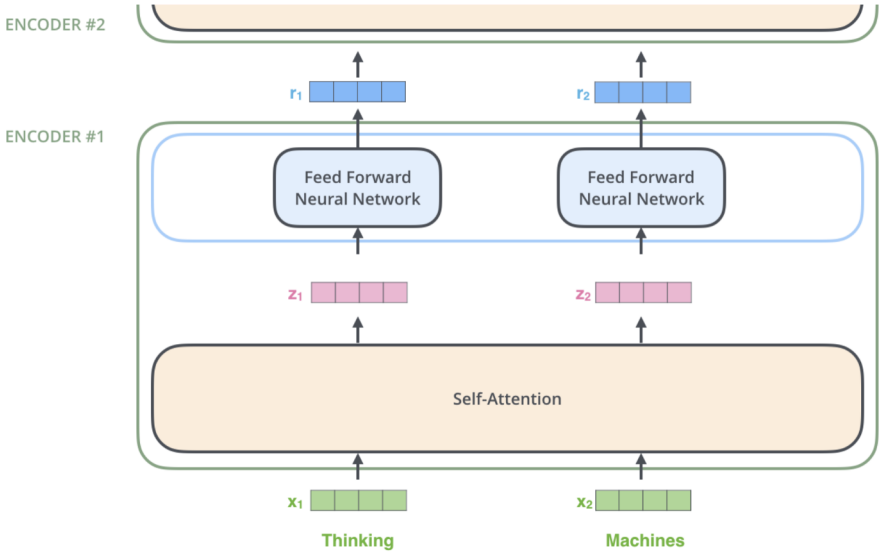




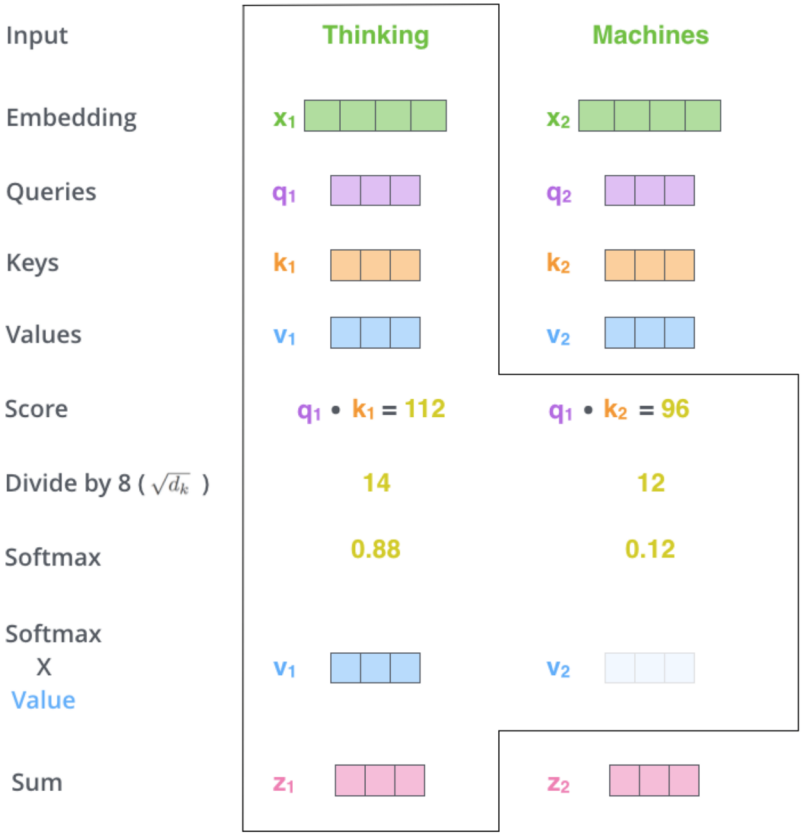


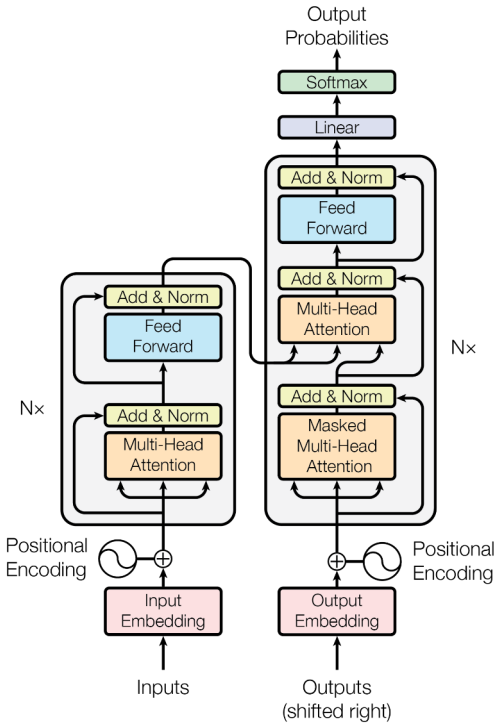
Generates the embedding dynamically for each context, this is how self-attention during encoding replaces RNNs

4.1. Attention is all you need: encoding

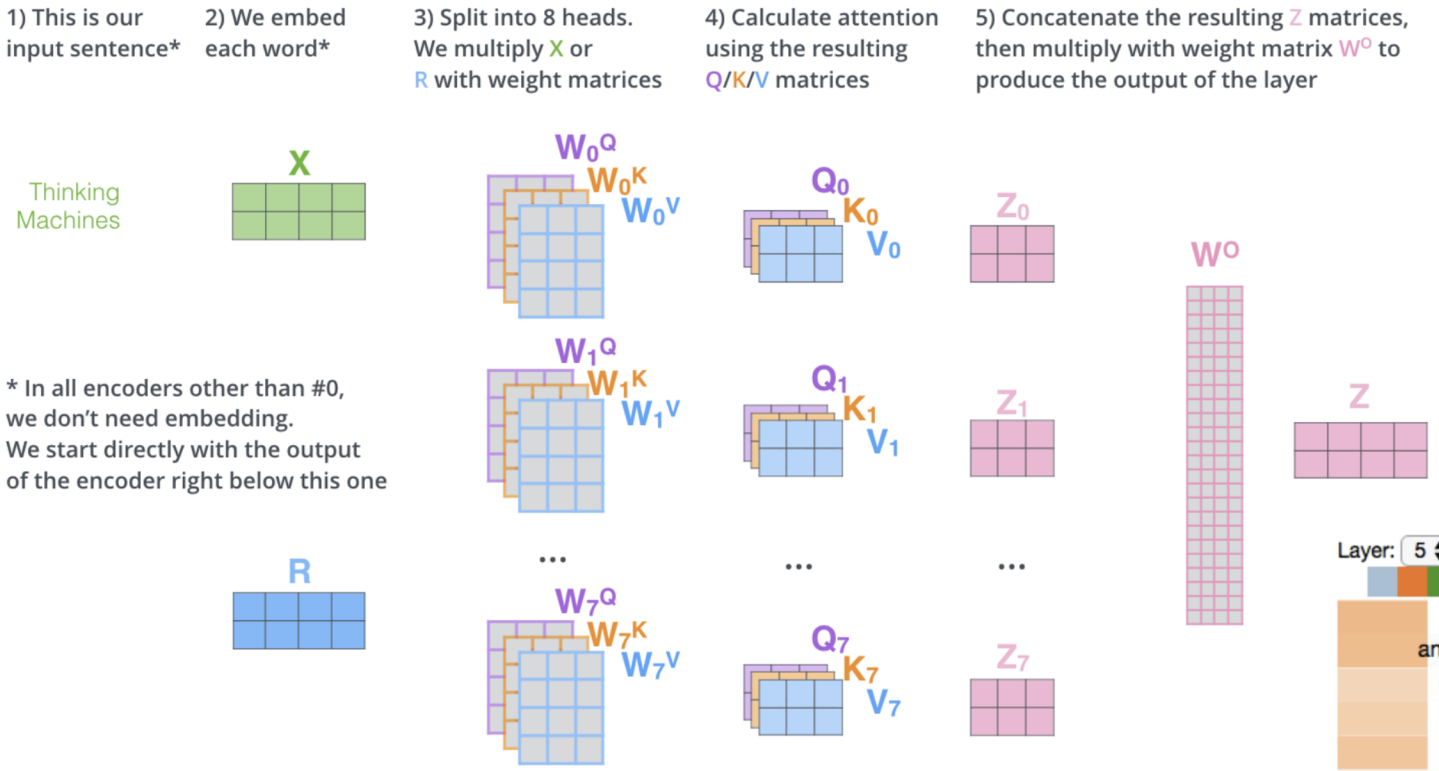


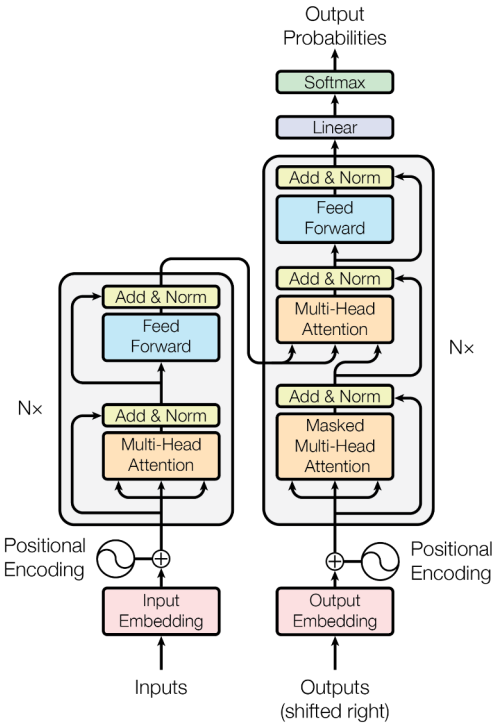
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

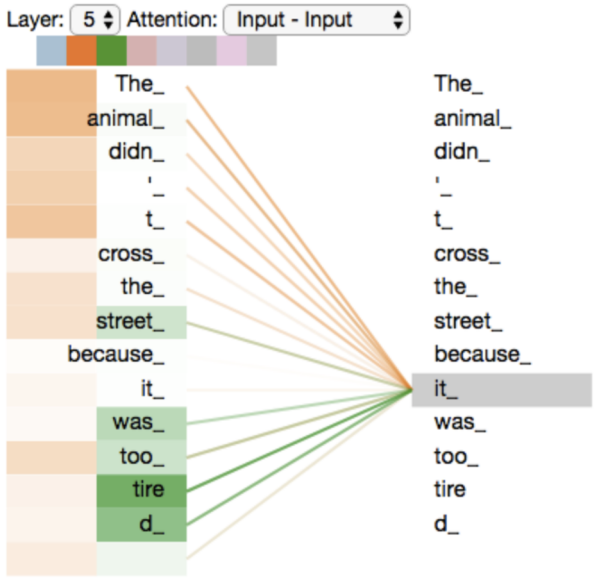




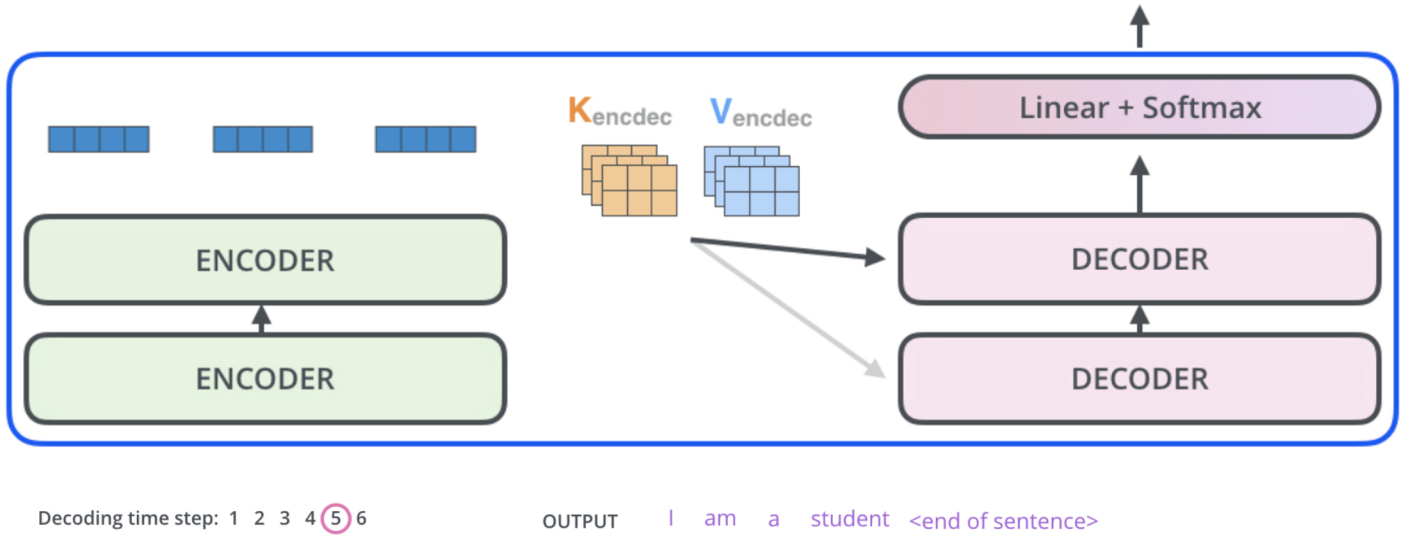
4.1. Attention is all you need: multi-head

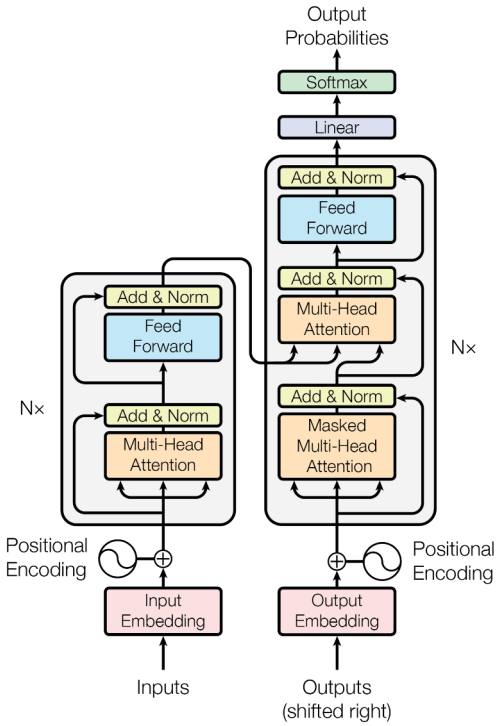


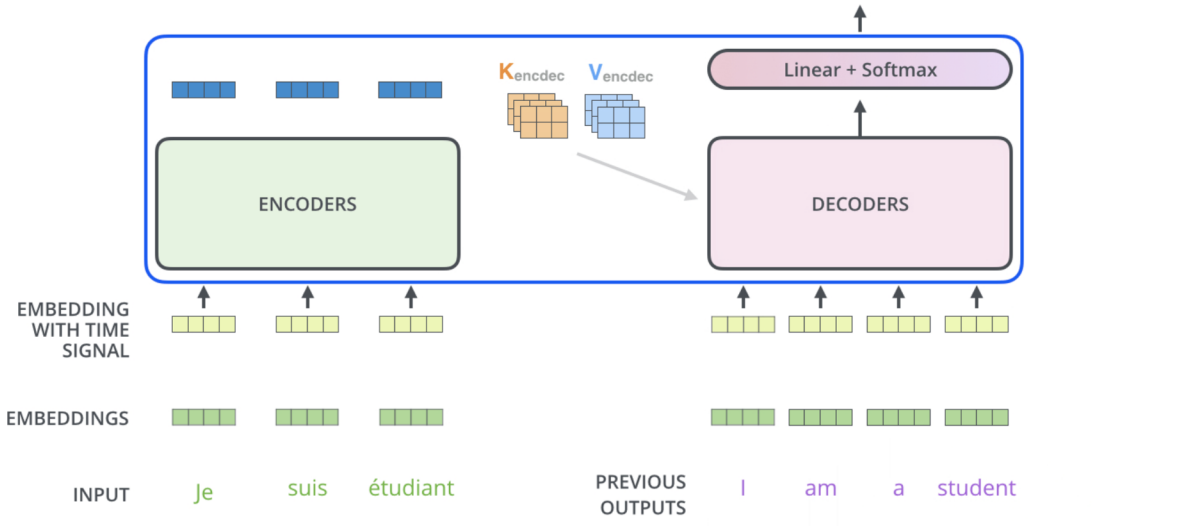


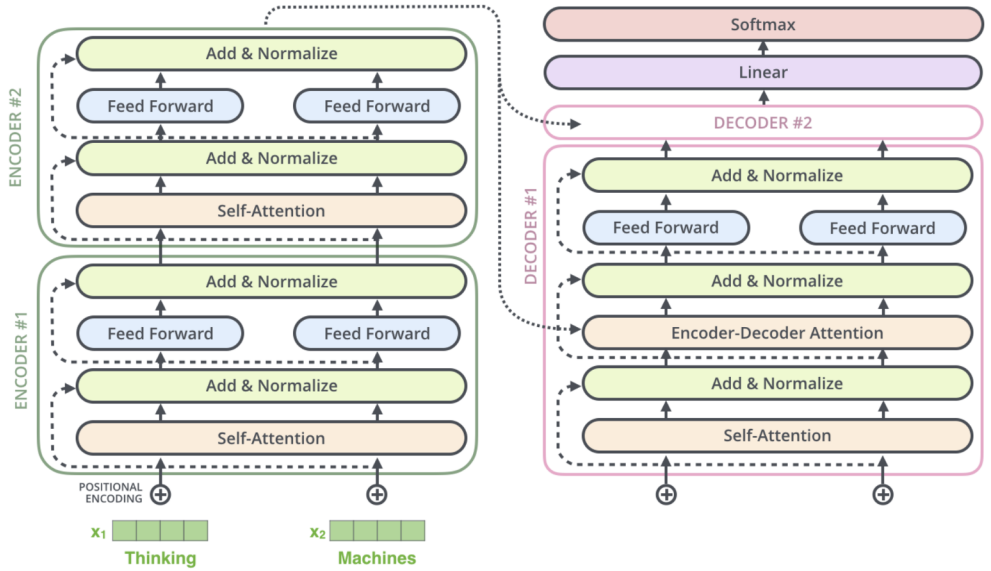


4.1. Attention is all you need: the decoder

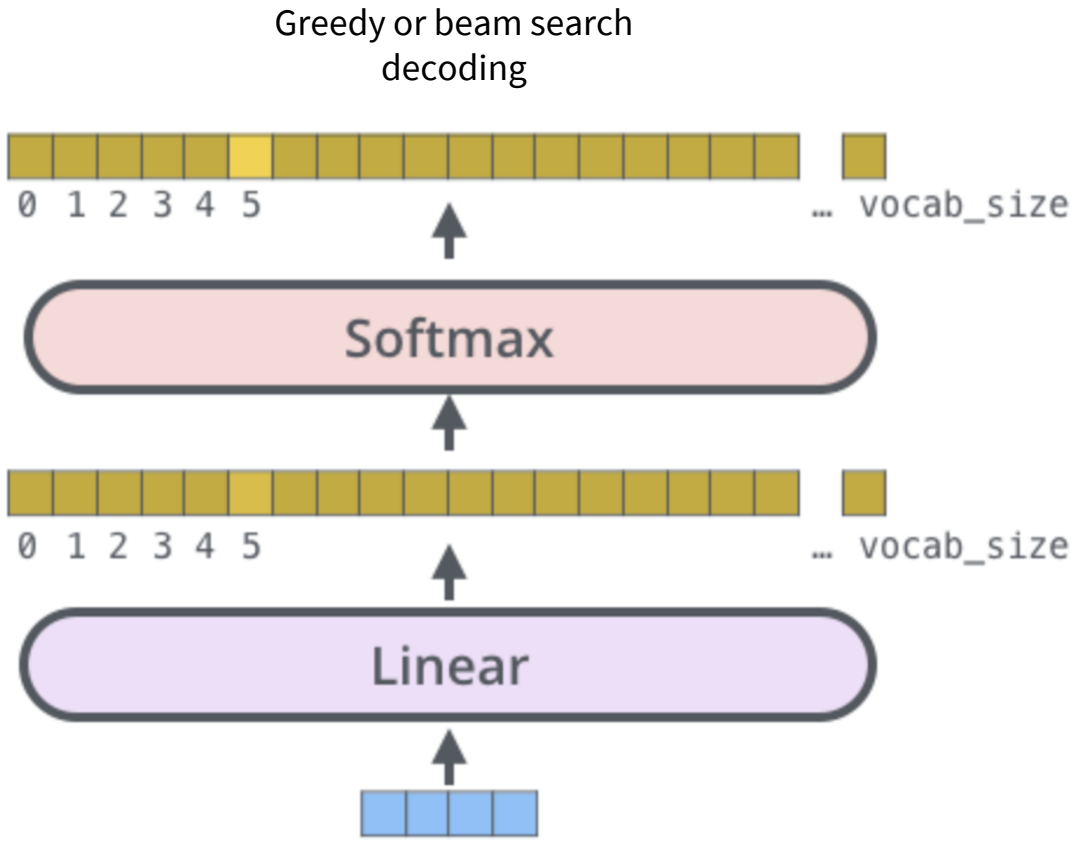


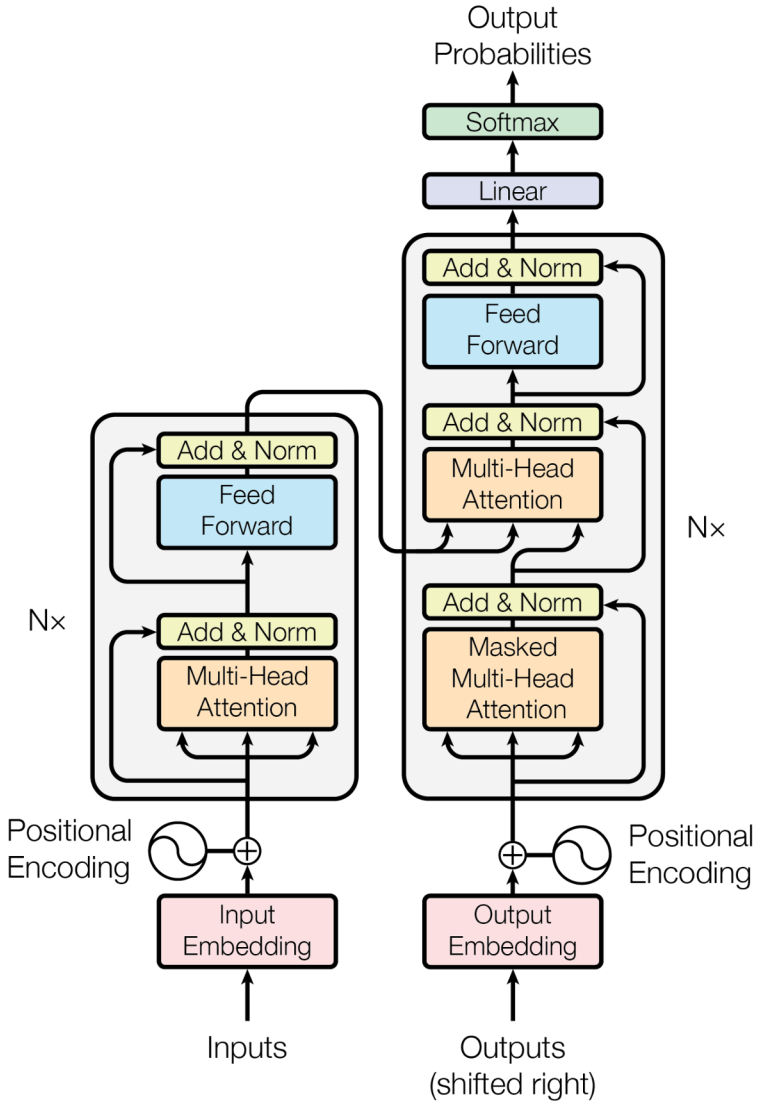




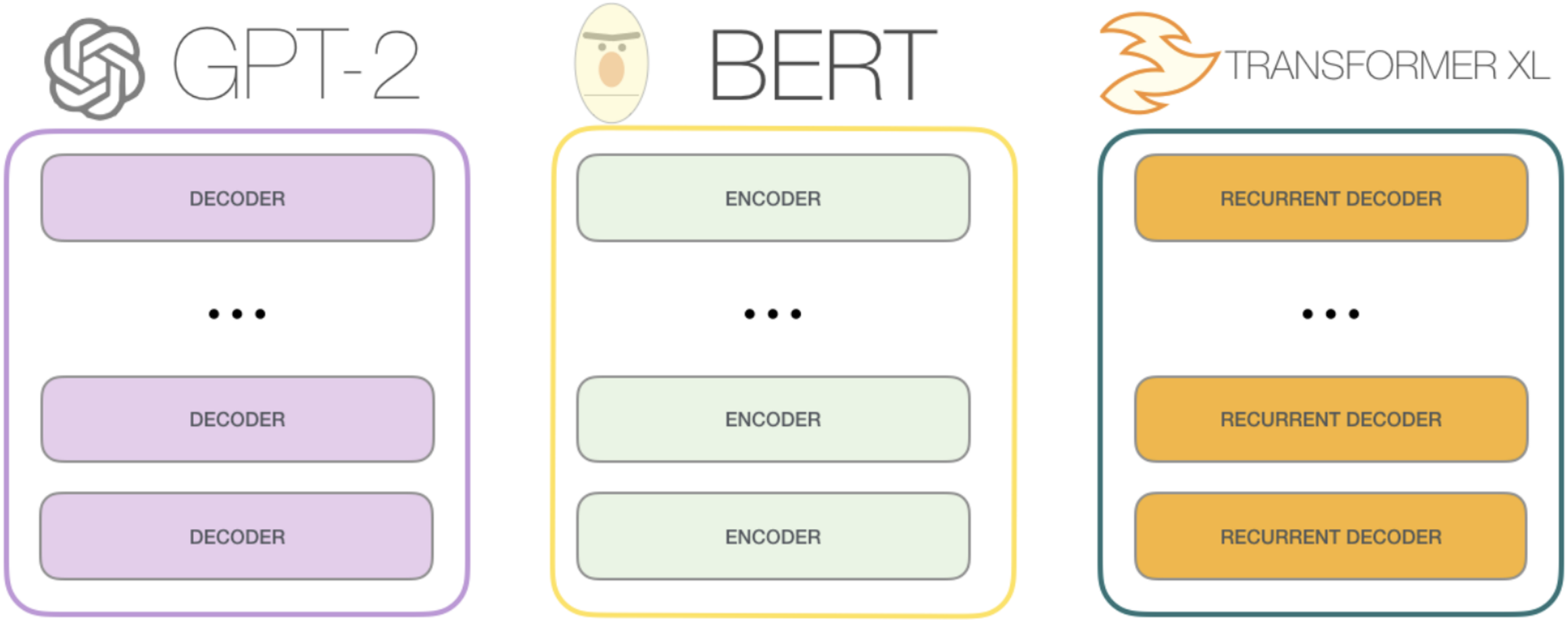


4.1. Attention is all you need: the decoder

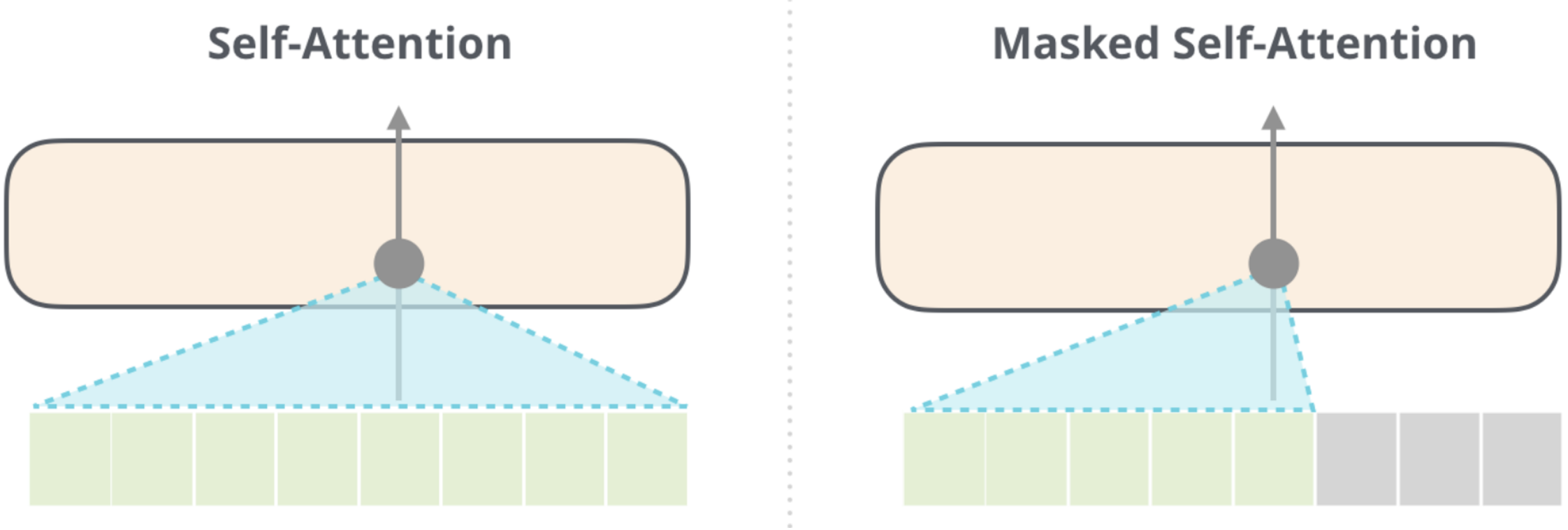




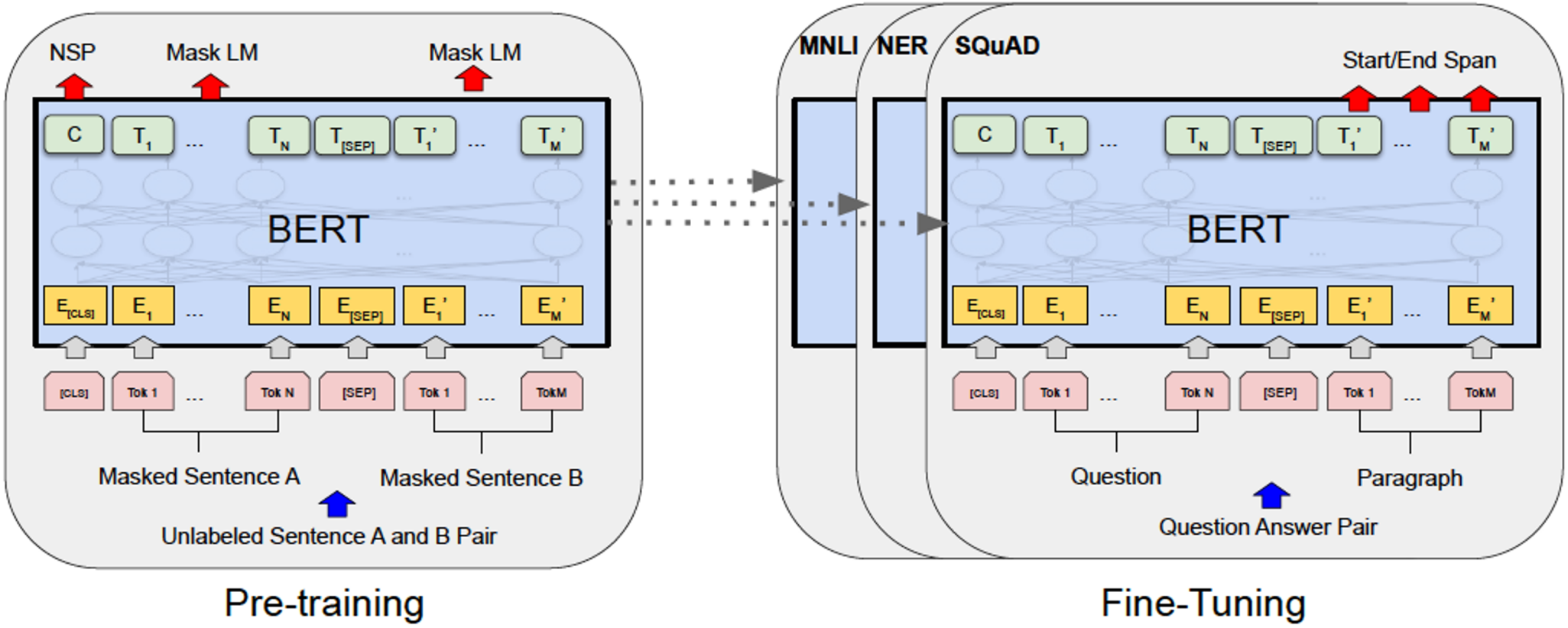
4.2. After the transformer...



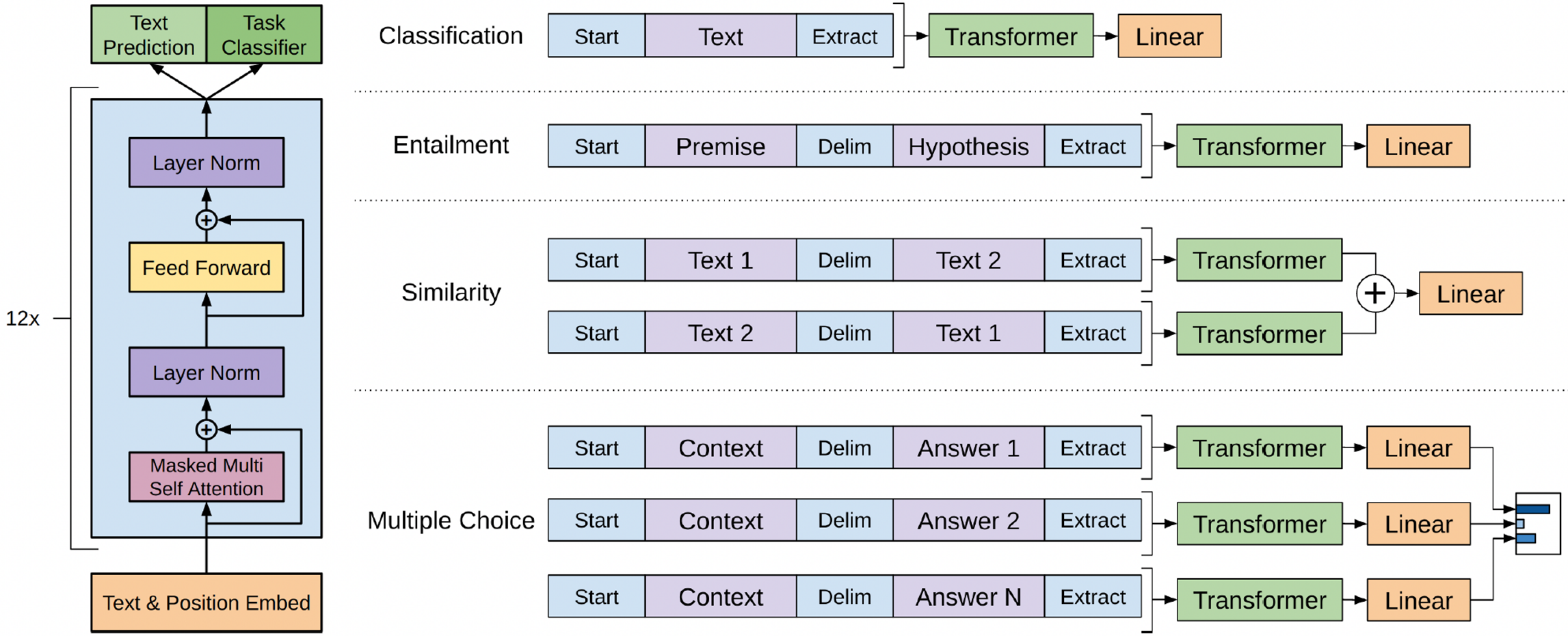
4.2. Encoder or decoder?



4.2. BERT: bidirectional encoder representations from transformers



4.3. GPT

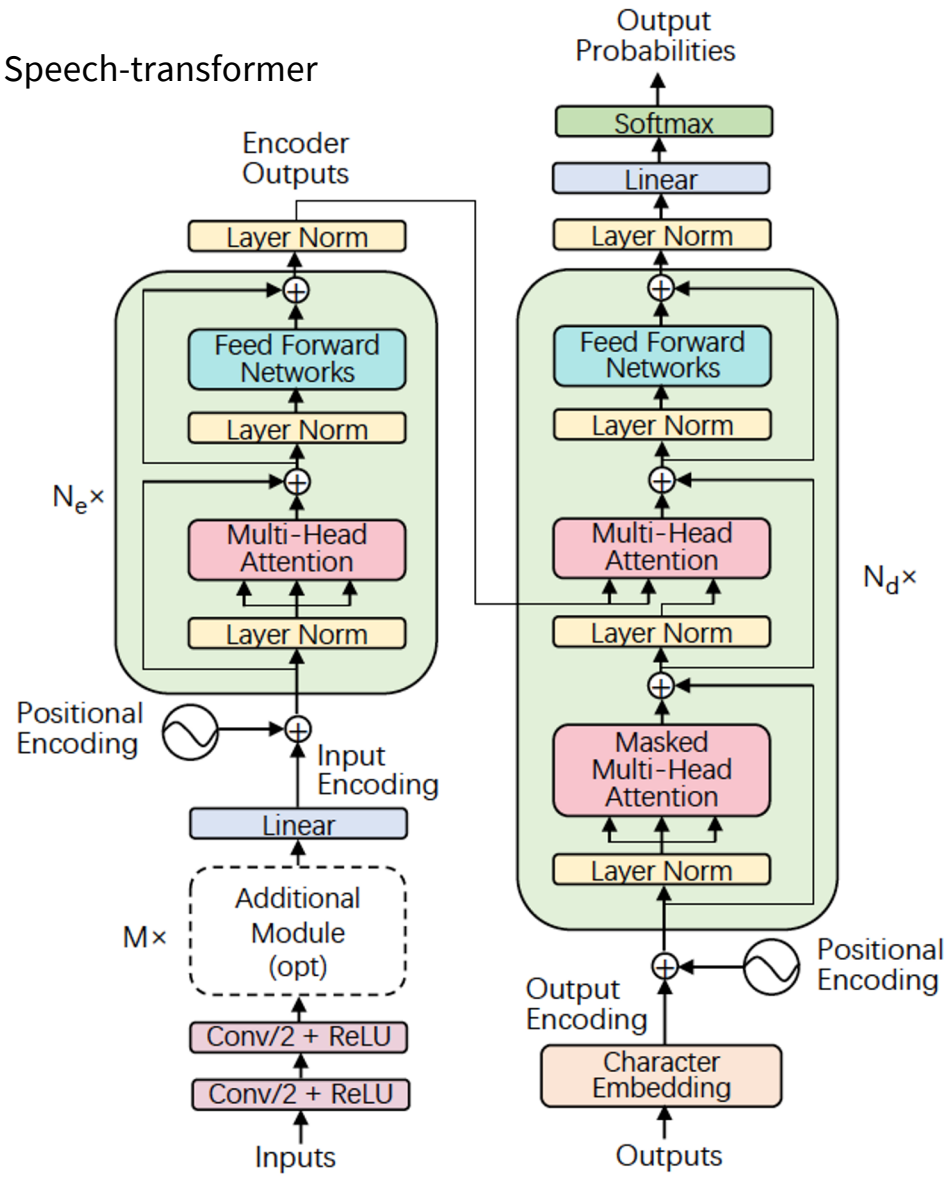

Text Prediction
Task Classifier
Layer Norm
Feed Forward
Layer Norm
Masked Multi Self Attention
12x
Text & Position Embed
Classification
Start
Text
Extract
Transformer
Linear
Entailment
Start
Premise
Delim
Hypothesis
Extract
Transformer
Linear
Similarity
Start
Text 1
Delim
Text 2
Extract
Start
Text 2
Delim
Text 1
Extract
Transformer
Transformer
Linear
Multiple Choice
Start
Context
Delim
Answer 1
Extract
Start
Context
Delim
Answer 2
Extract
Start
Context
Delim
Answer N
Extract
Transformer
Transformer
Transformer
Linear
Linear
Linear

4.3. GPT-2: 1.5B parameters (GPT-3 has 175B)

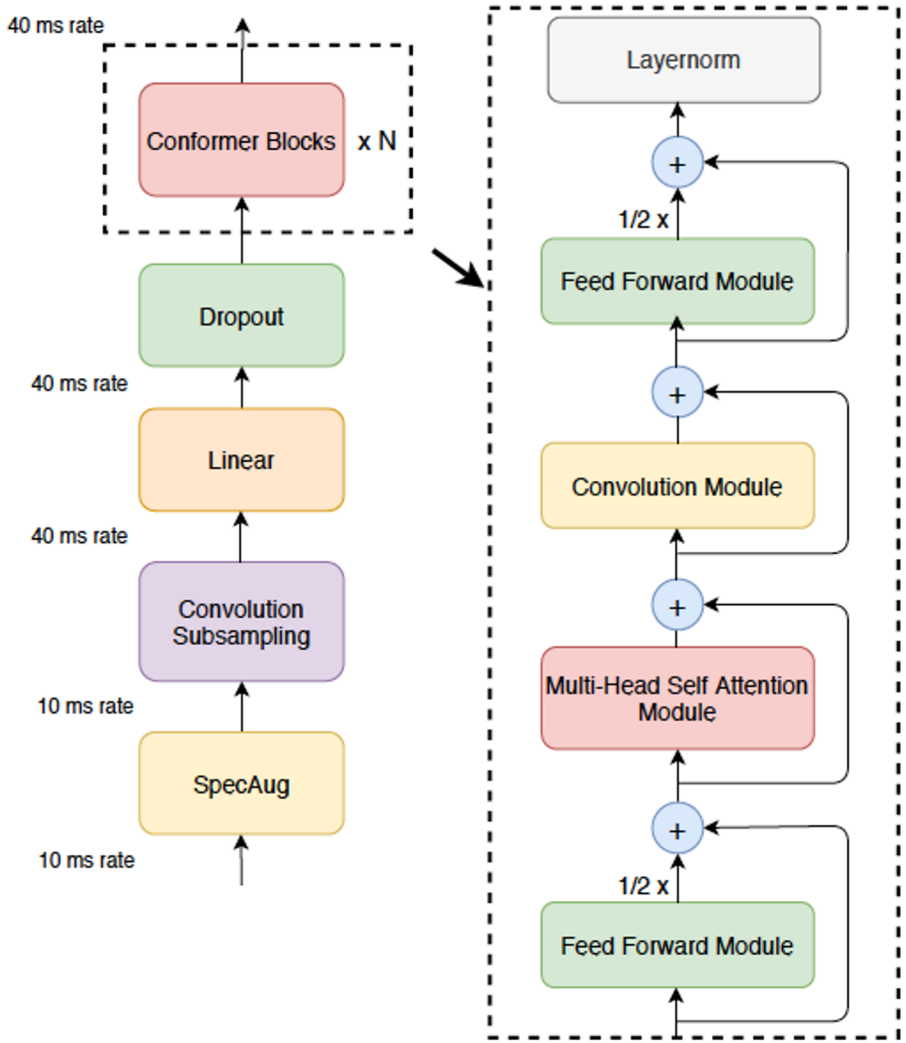
SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	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.

4.4. Applications: ASR

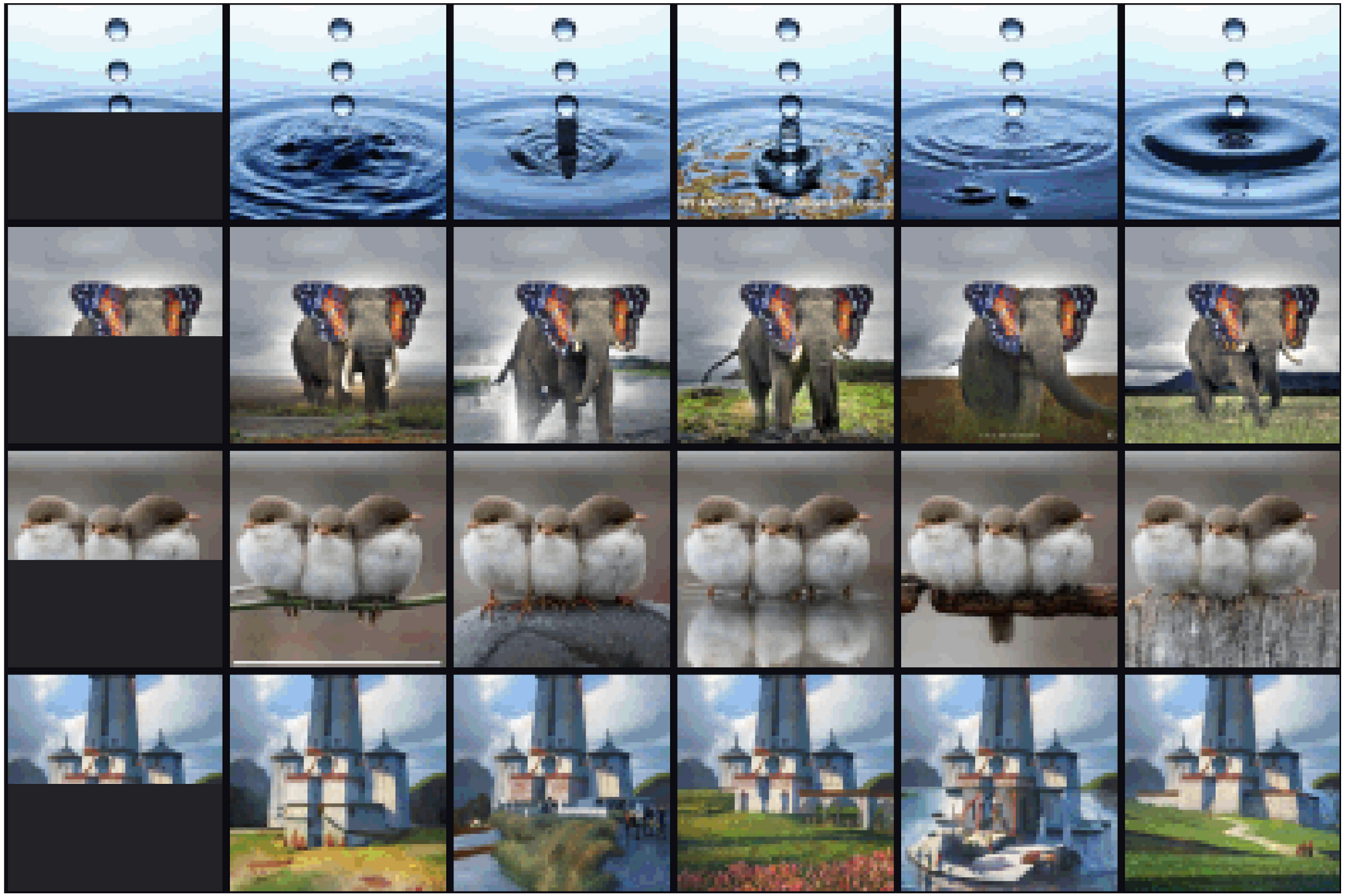
Speech-transformer



Conformer encoder (+LSTM decoder)



4.4. Applications: Image generation - Image GPT



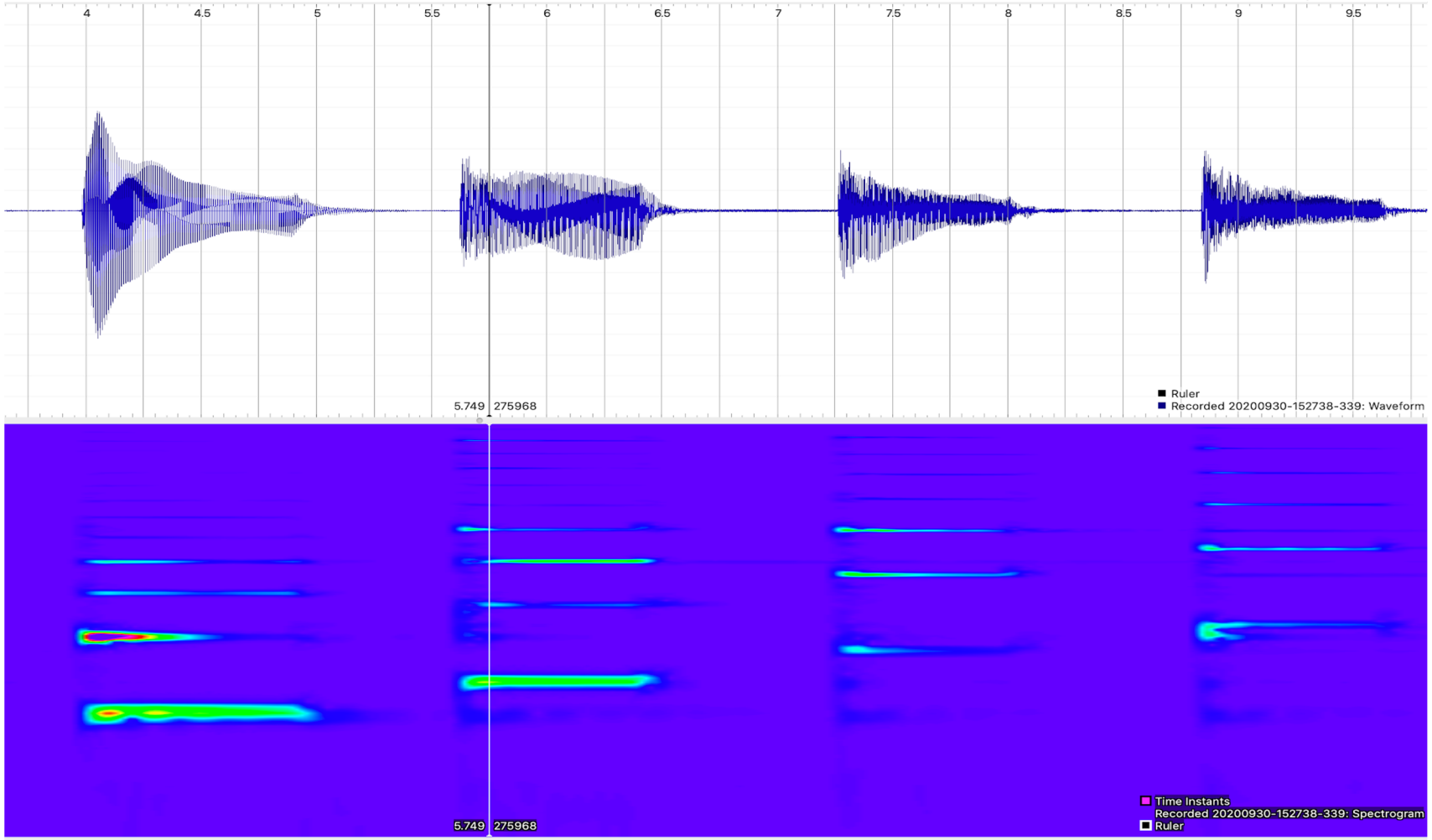
Recap

- Sequence to sequence problems can be approached with RNNs.
- Attention improves modelling of long-range dependencies and parallelization. End-to-end networks with implicit LM.
- Pre-training transformer models on massive corpora allows for “easy” fine-tuning with little data.
- Transformers can be used in images, music, speech, text...

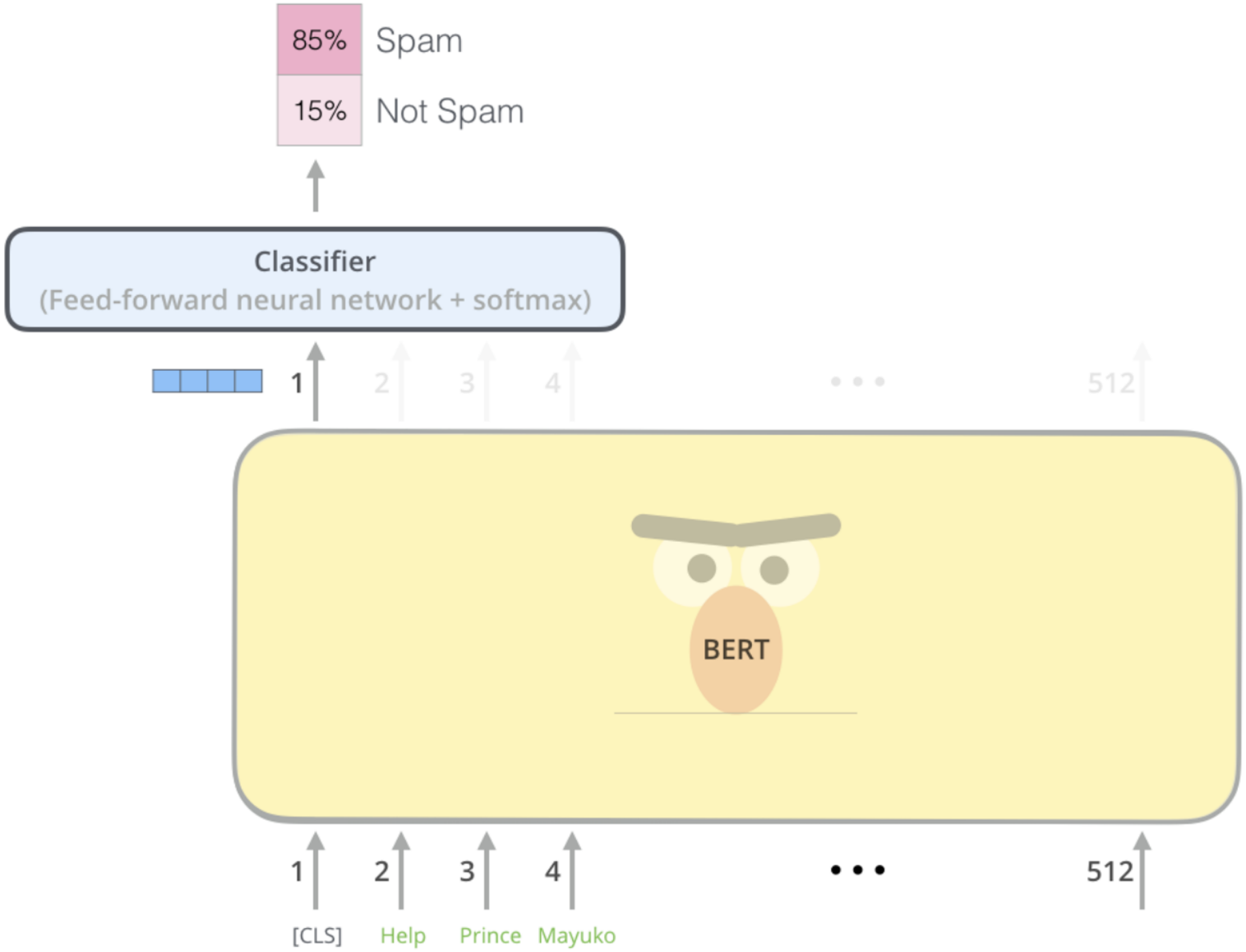
Extra

2.2 ASR: from HMMs to end-to-end with DNNs

Acoustic and language modelling $p(s) = p(w_1)p(w_2|w_1)p(w_3|w_1\,w_2)\cdots p(w_l|w_1\cdots w_{l-1}) = \prod_{i=1}^{l} p(w_i|w_1\cdots w_{i-1})$



4.2. BERT


85%
Spam
15%
Not Spam
Classifier
(Feed-forward neural network + softmax)
1
2
3
4
512
BERT
1
2
3
4
512
[CLS]
Help
Prince
Mayuko