ELEC-C5220
Lecture 8:
Automatic Speech Recognition

Machine learning in information technology

Aalto University
School of Electrical
Engineering

Lauri Juvela

7.3.2024

Lecture 8 content

- Automatic speech recognition (ASR)
- The sequence alignment problem
- Attention mechanism in neural networks
- Transformer neural net
- ASR with Transformers

Transformer neural network

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Aalto University
School of Electrical
Engineering

Whisper ASR

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] **Jong Wook Kim** [*1] **Tao Xu** [1] **Greg Brockman** [1] **Christine McLeavey** [1] **Ilya Sutskever** [1]

Abstract

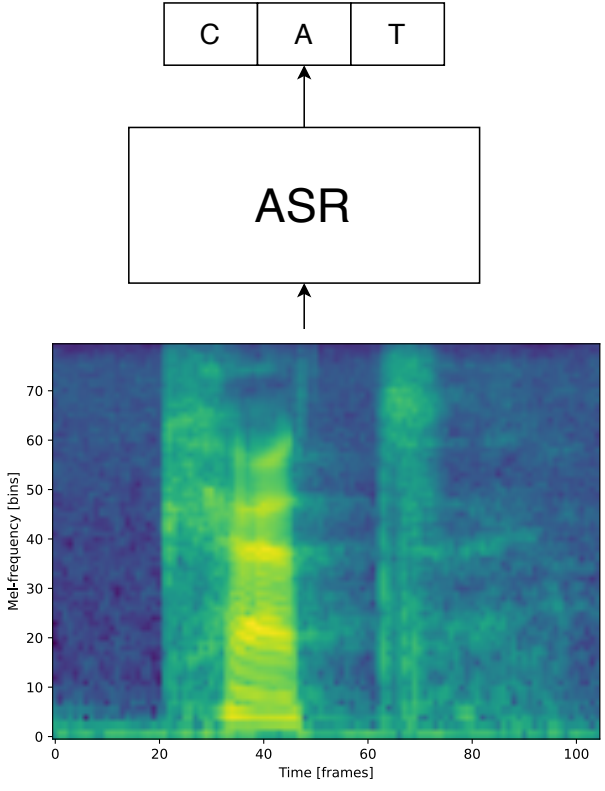
We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy when classifying the same objects on seven other natural image datasets. A model that achieves "superhuman" performance when trained on a dataset can still make many basic errors when evaluated on another, possibly precisely because it is exploiting those dataset-specific quirks that humans are oblivious to (Geirhos et al., 2020).

Aalto University
School of Electrical
Engineering

Automatic Speech Recognition (ASR)

- Task: transcribe text from speech
- Requires knowledge about speech signal – Acoustic Model
- Requires knowledge about language – Language Model
- How to combine the two?



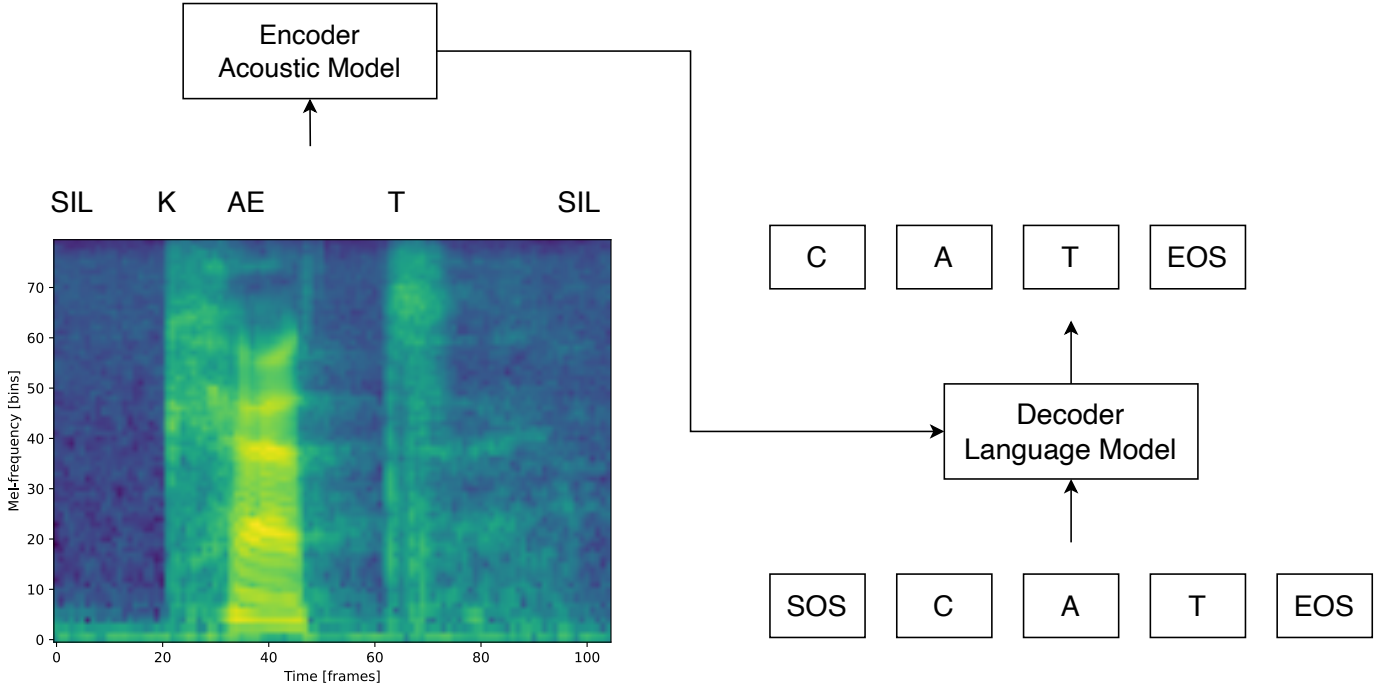
ASR applications

- Captioning and transcription
- Speech interfaces
- Voice chatbots for customer service
- Conversational AI
- …

ASR related fields

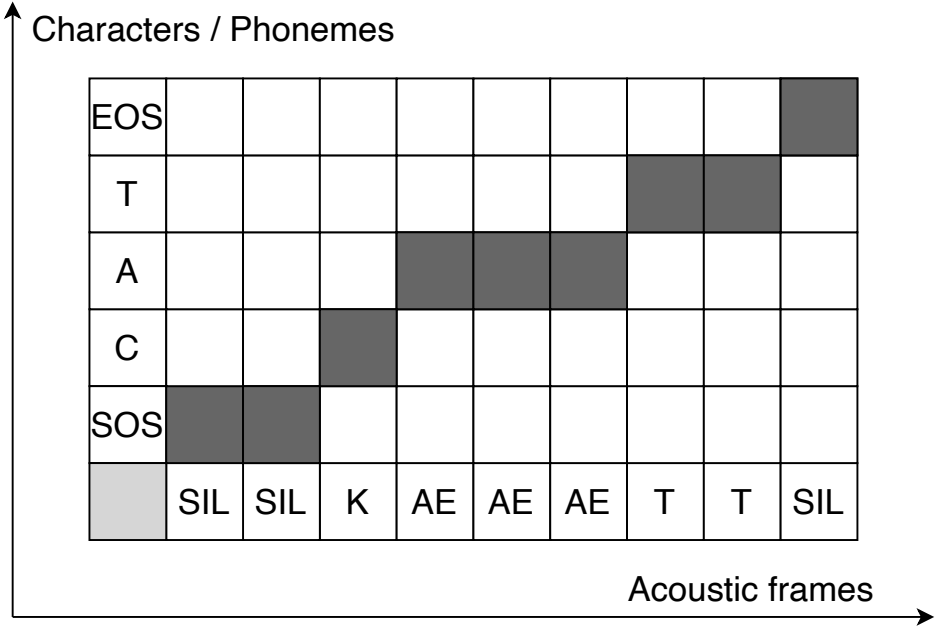
- Language modeling
- Machine translation
- Speech synthesis
- Speaker recognition
- All these use neural networks for solving sequence-to-sequence tasks!

Acoustic Model Encoder and Language Model Decoder in ASR



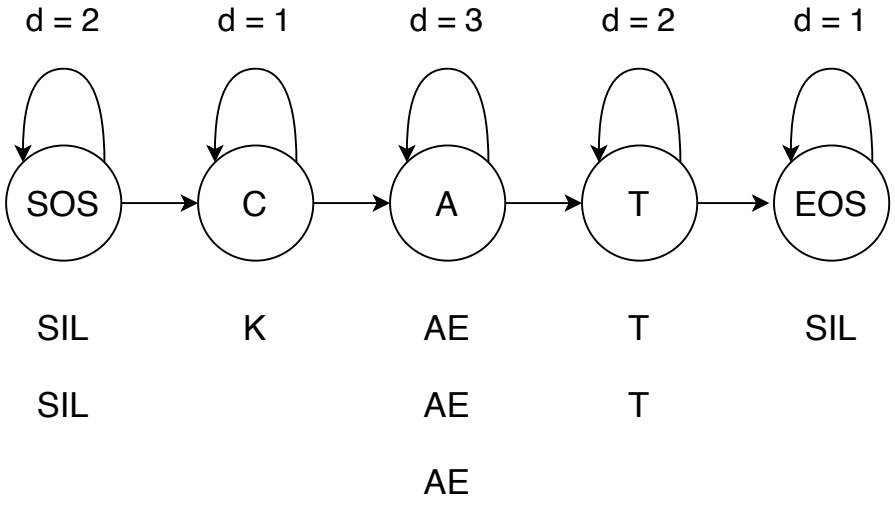
The alignment problem

- **Phoneme duration is usually multiple time-frames**
- **Alignment is monotonic: both speech and text move forward in time**
- **Durations are variable**
- **How to align the two sequences?**



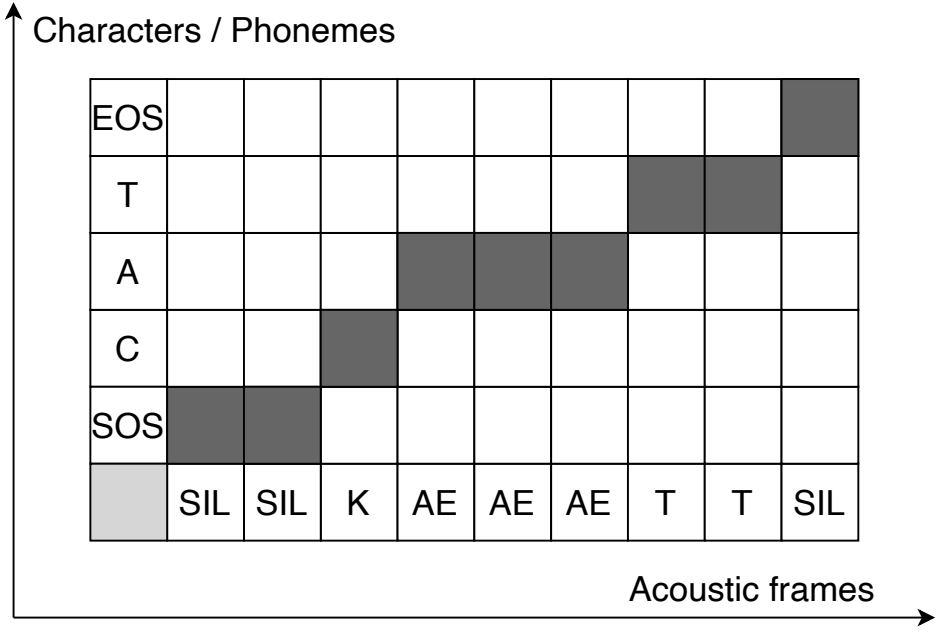
Hidden Markov Models (HMMs)

- **Classic model in ASR**
- **Phonemes (~letters if Finnish) emit acoustic frames**
- **Phoneme emits a frame on every time-step for it's duration, then state moves to next phoneme**
- **Still useful in alignment**



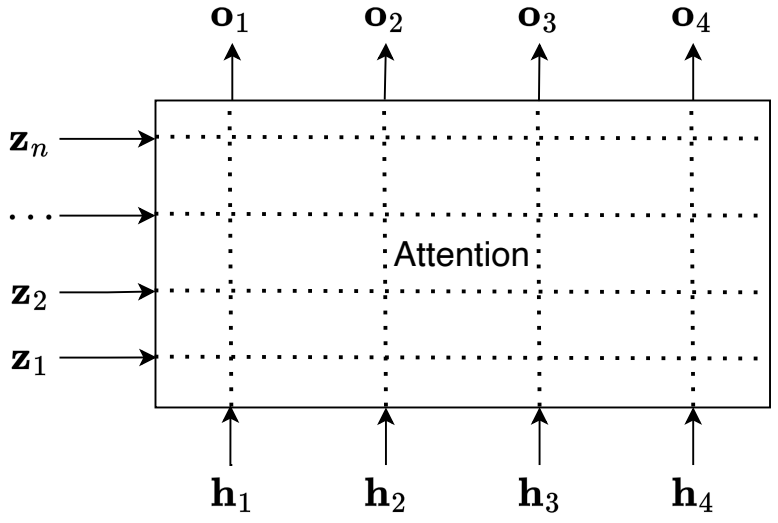
Predicting alignment with neural networks

- **How to generate this kind of aligment matrix in general?**
- **Requires some kind of distance metric**
- **Attention mechanism does exactly this!**



Attention

- Embed each sequence in d-dimensional space
- Compute attention weights as dot products between each time step on both sequences
- Normalise with softmax
- Apply attention weights to map between the sequences

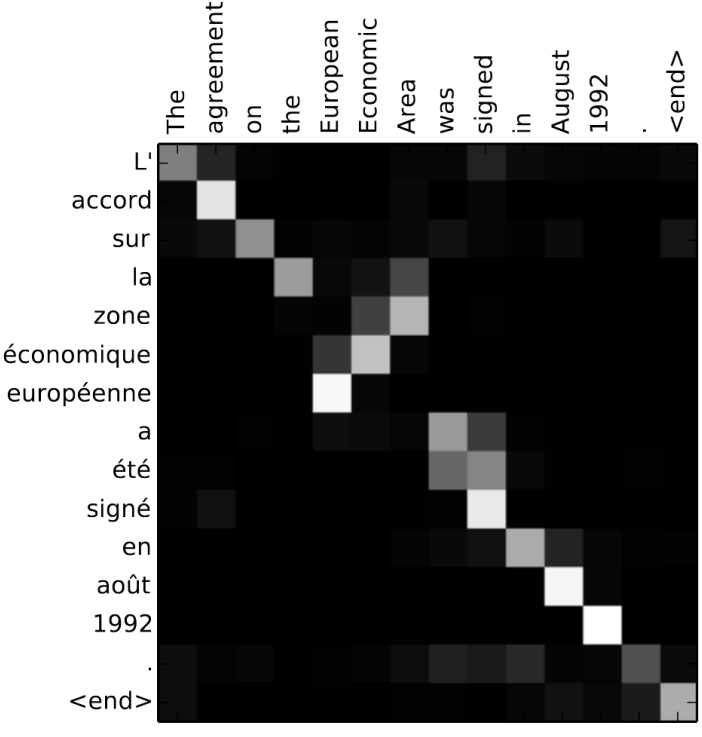


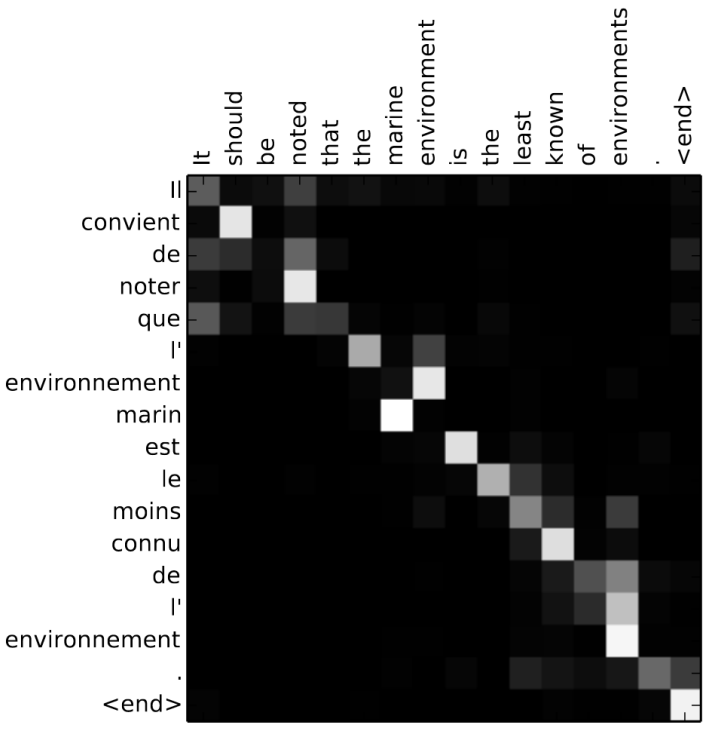
$$\mathbf{z}_j, \mathbf{h}_i \in \mathbb{R}^d$$

$$\alpha_{i,j} = \frac{\exp(\mathbf{z}_j^T \mathbf{h}_i / \sqrt{d})}{\sum_{j'=1}^{n} \exp(\mathbf{z}_{j'}^T \mathbf{h}_i / \sqrt{d})}$$

$$\mathbf{o}_i = \sum_{j=1}^{n} \alpha_{i,j} \mathbf{z_j}$$

Attention weights visualised (machine translation example)



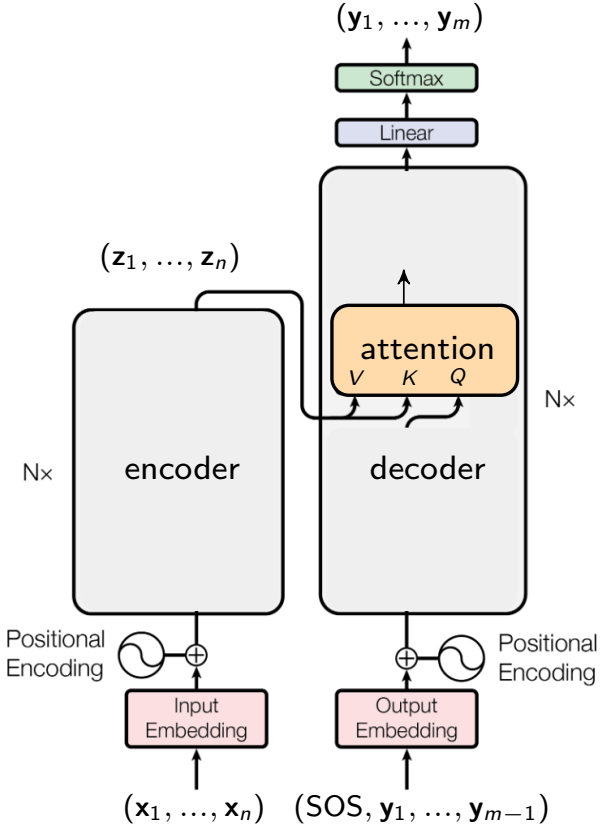


Attention – Query, Key, Value

- **Generalises self and cross attention**
- **Dot product measures similarity between key and query**
- **In cross attention, key and value are the same vector**

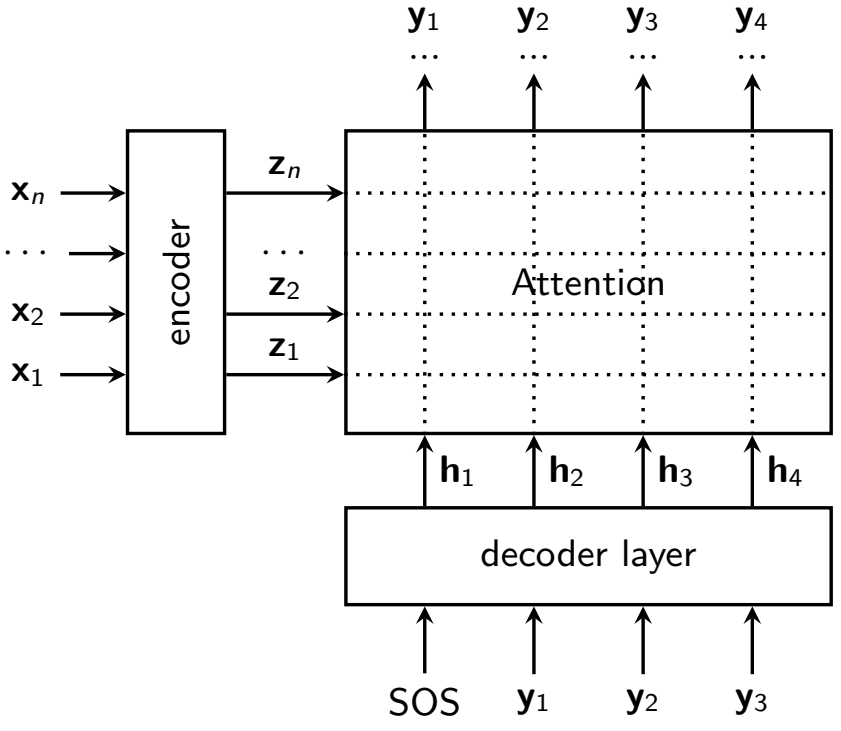
$$\mathbf{o}_i = \sum_{j=1}^{n} \alpha_{ij} \mathbf{z}_j \qquad \mathbf{o}_i = \sum_{j=1}^{n} \alpha_{ij} \mathbf{v}_j$$

$$\alpha_{ij} = \frac{\exp(\mathbf{z}_j^\top \mathbf{h}_i / \sqrt{d_k})}{\sum_{j'=1}^{n} \exp(\mathbf{z}_{j'}^\top \mathbf{h}_i / \sqrt{d_k})} \qquad \alpha_{ij} = \frac{\exp(\mathbf{k}_j^\top \mathbf{q}_i / \sqrt{d_k})}{\sum_{j'=1}^{n} \exp(\mathbf{k}_{j'}^\top \mathbf{q}_i / \sqrt{d_k})}$$



ASR with attention

- **Decoder Language Model predicts next token given previous tokens using masked Self-Attention**
- **Decoder receives information about the speech signal from the Encoder Acoustic Model using Cross-Attention**



Scaled dot-product attention

- **Scalar form**

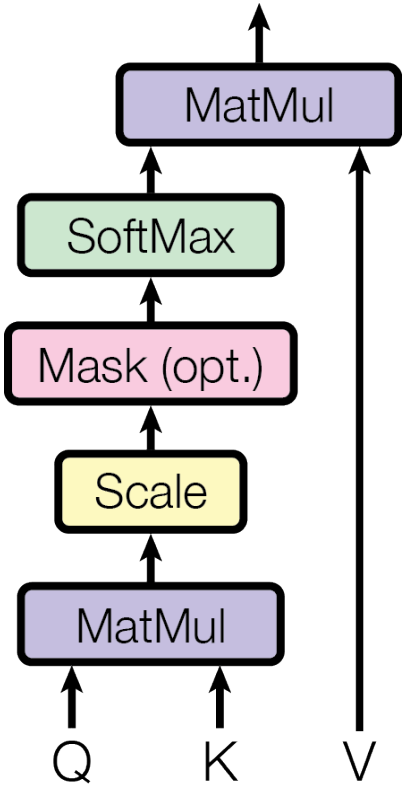
$$\mathbf{o}_i = \sum_{j=1}^{n} \alpha_{ij} \mathbf{v}_j$$

$$\alpha_{ij} = \frac{\exp(\mathbf{k}_j^\top \mathbf{q}_i / \sqrt{d_k})}{\sum_{j'=1}^{n} \exp(\mathbf{k}_{j'}^\top \mathbf{q}_i / \sqrt{d_k})}$$

- **Matrix form**

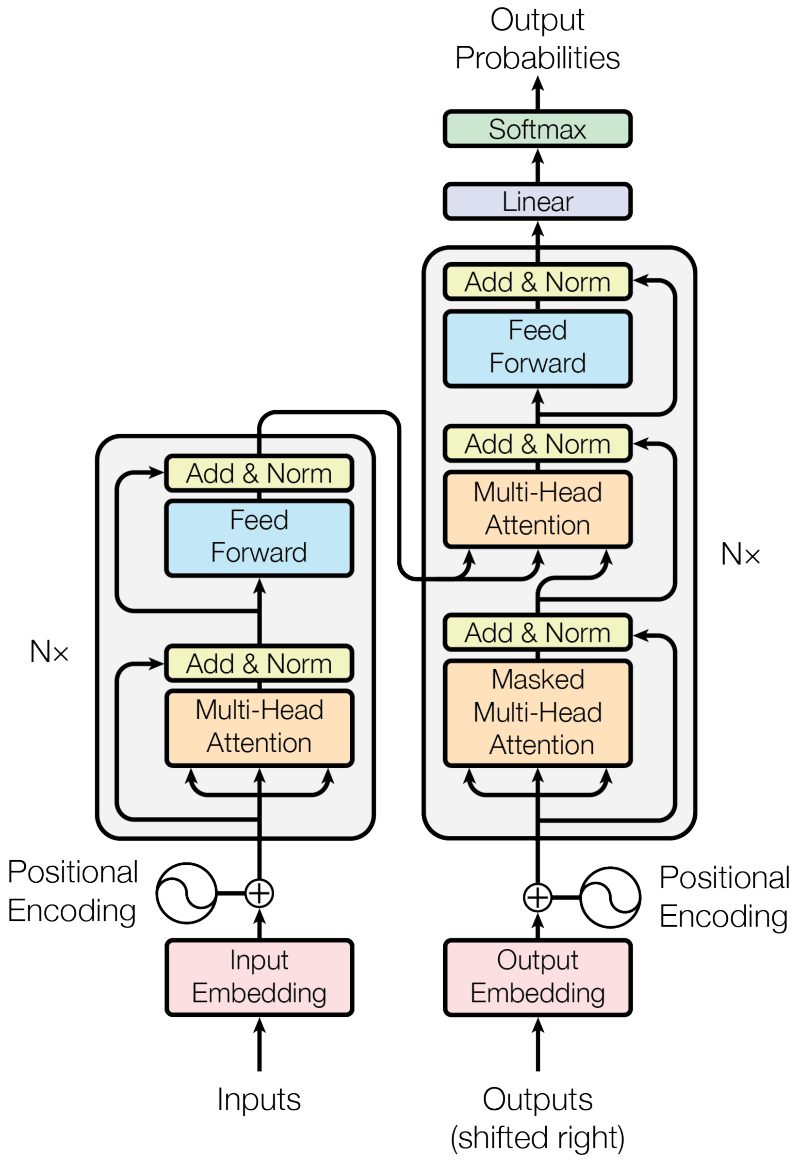
$$\text{attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

$$\mathbf{V} \in \mathbb{R}^{m \times d_v},\ \mathbf{Q} \in \mathbb{R}^{n \times d_k},\ \mathbf{K} \in \mathbb{R}^{m \times d_k}$$



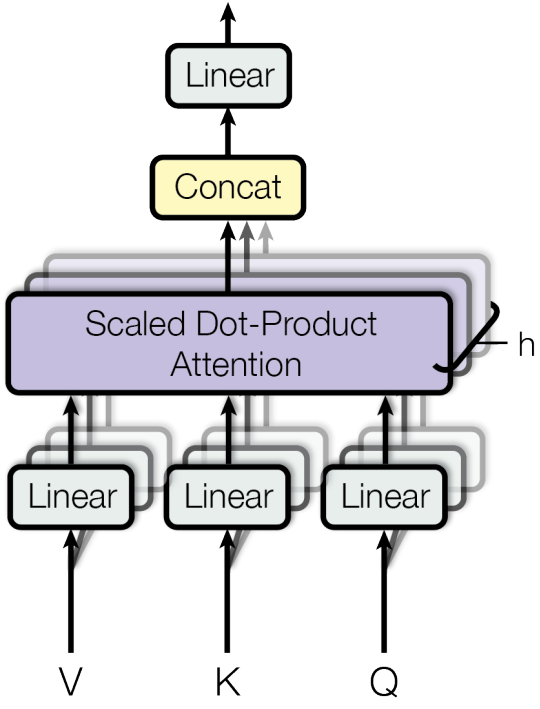
Transformer

- **Good general tool for sequence-to-sequence problems**
- **Backbone of LLMs, including the Generative Pretrained Transformer (GPT) family of models**



Multi-Head Attention

- **Transformers use multiple attention layers in parallel**

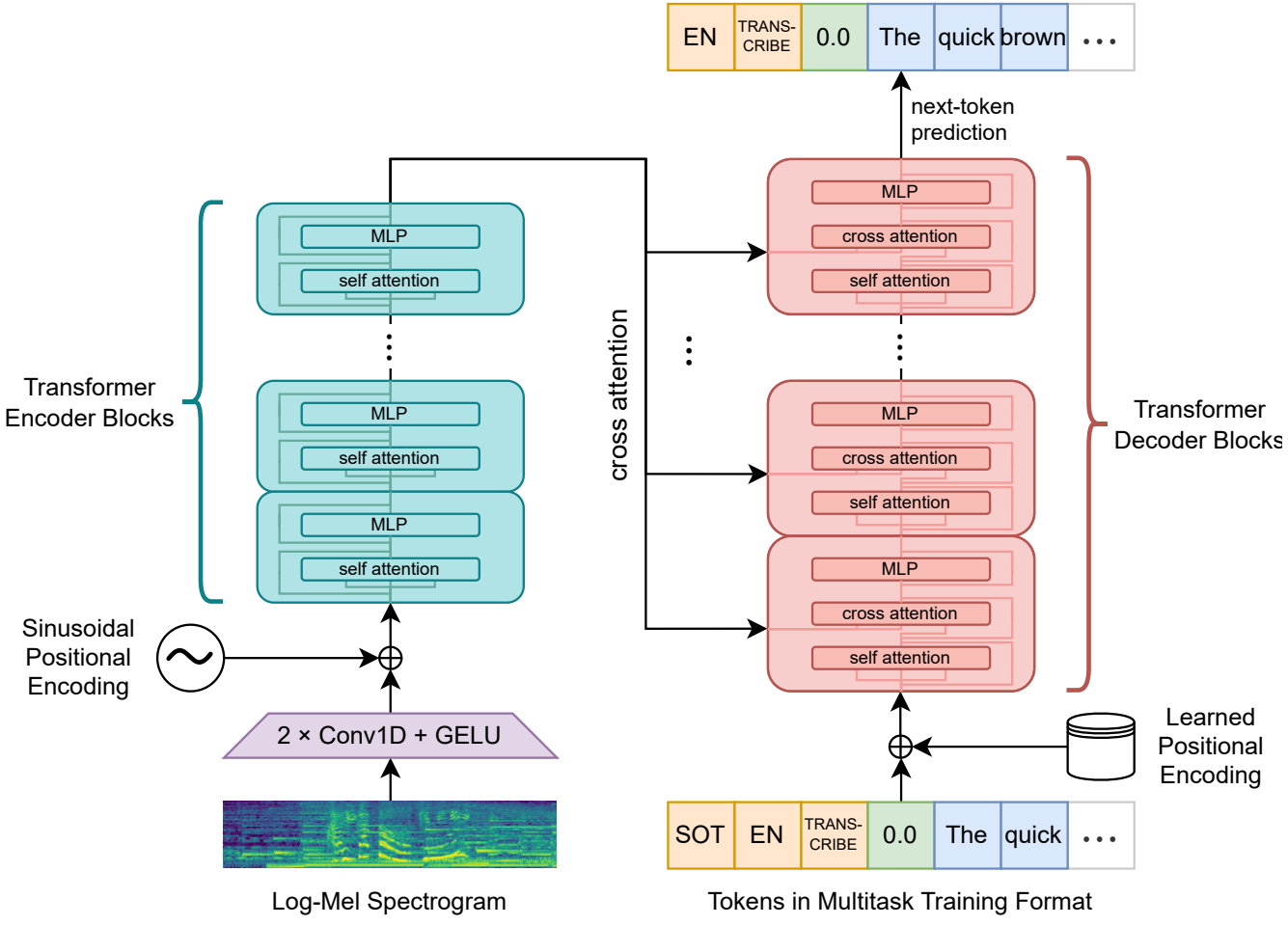


$$\text{head}_i = \text{attention}(QW_i^Q, KW_i^K, VW_i^V)$$

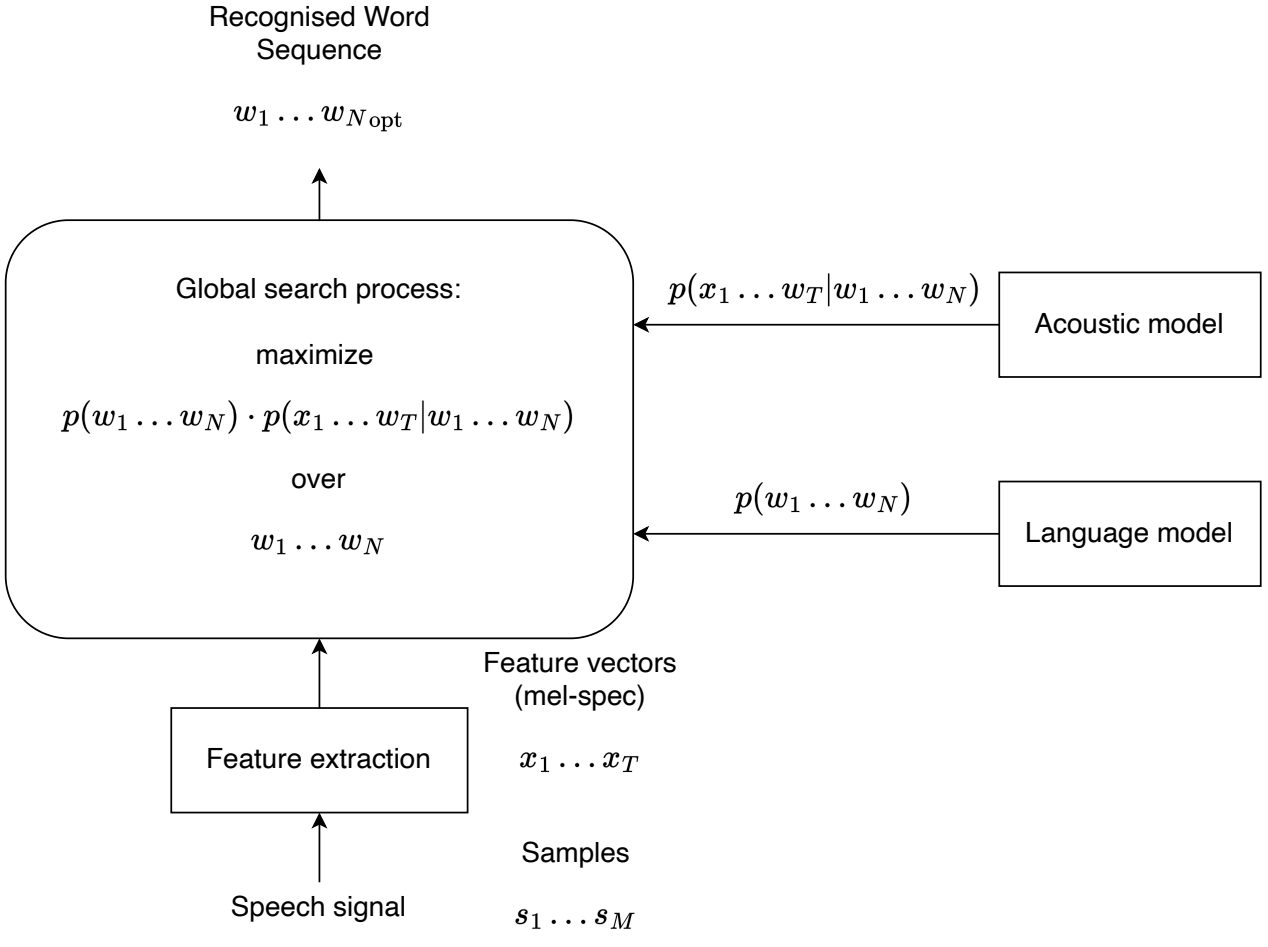
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$

$$\mathbf{V} \in \mathbb{R}^{m \times d_v},\ \mathbf{Q} \in \mathbb{R}^{n \times d_k},\ \mathbf{K} \in \mathbb{R}^{m \times d_k},$$
$$\text{head}_i \in \mathbb{R}^{n \times d_i},\ \text{output} \in \mathbb{R}^{n \times d_k}.$$

Whisper ASR architecture



ASR is a search problem



Weekly exercise

- Dissect a pre-trained Whisper ASR model
- Let’s try to visualise how cross attention between speech and text works
- Re-implement the language model decoder – Similar to last week’s autoregressive text generation!

Reading list: Transformer

Attention is All you Need

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

https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Reading list: Whisper ASR

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever

https://arxiv.org/abs/2212.04356

Aalto University
School of Electrical
Engineering

Reading list: Attention – explained well with historical context

Aalto Deep Learning Course (CS-E4890), Lecture 6: Attention-based models

Alexander Ilin

https://mycourses.aalto.fi/mod/resource/view.php?id=1013764

Aalto University
School of Electrical
Engineering

Lecture 8 summary

- Automatic speech recognition (ASR)
- The sequence alignment problem
- Attention mechanism in neural networks
- Transformer neural net
- ASR with Transformers