# Sequence Learning

## Pre-training and Important transformer architectures

**Sebastian Bayerl, Korbinian Riedhammer**

TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

# Todays episode features:

- Pretraining/Fine-tuning / Semi-Supervise learning refresher
- BERT
- Vision transformer
- wav2vec 2.0
- data2vec

# Idea

## Pretraining

- Multiple problems share common features / properties
- Transfer knowledge from one domain to other domain
- Transfer knowledge from one modality to another modality/domain

# Terminology

- Pre-training
  - Supervised
  - Semi-Supervised
    - Self supervised
    - Self training
- Fine-tuning
- Transfer learning
- Representation learning

# Terminology

## Supervised and Semi-Supervised Pretraining

- Supervised:
  - all training data has labels
- Semi supervised
  - a small part of the data has labels
  - knowledge about the class distribution, use of clustering / bayes-rule to assesing labels

# Self training

I like cats

Help me, I don’t know stuff

I hate dogs

**Train supervised model with all data you have!**

Supervised Model

# Self training

Supervised model

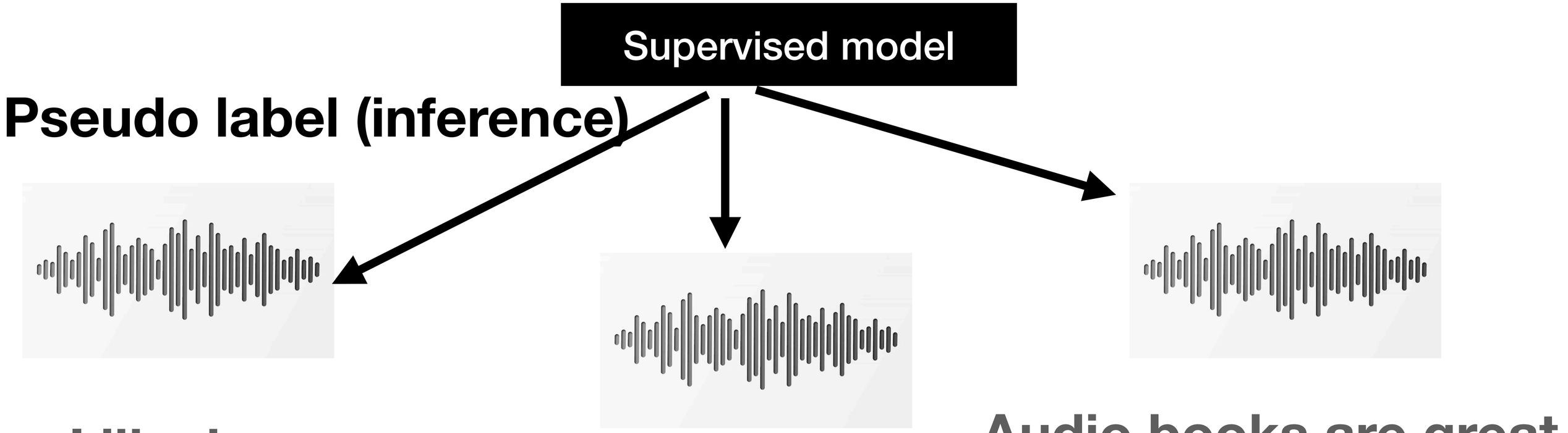
**Get quality unlabeled audio data**

I like burgers

Audio books are great

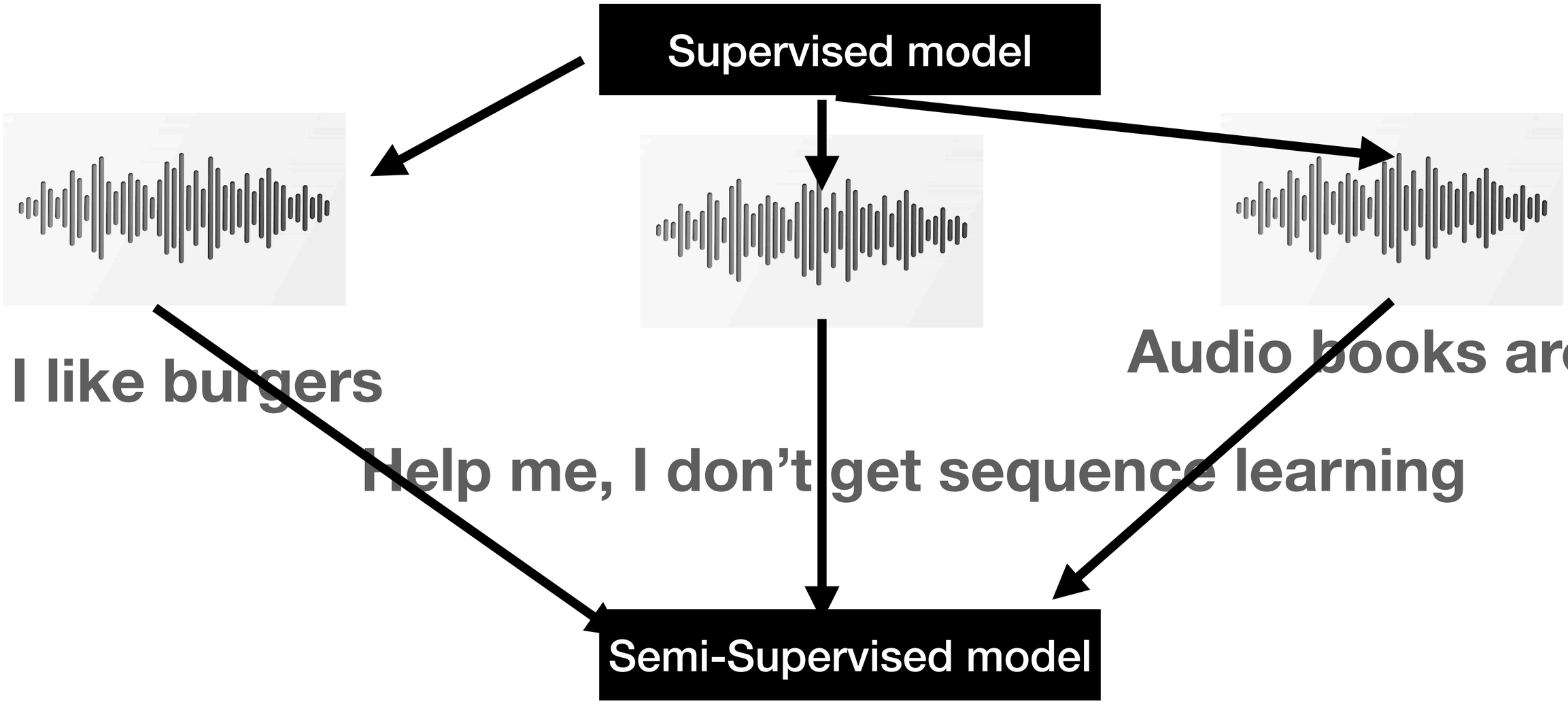
Help me, I don’t get sequence learning

# Self training



# Self training

## Get quality unlabeled audio data

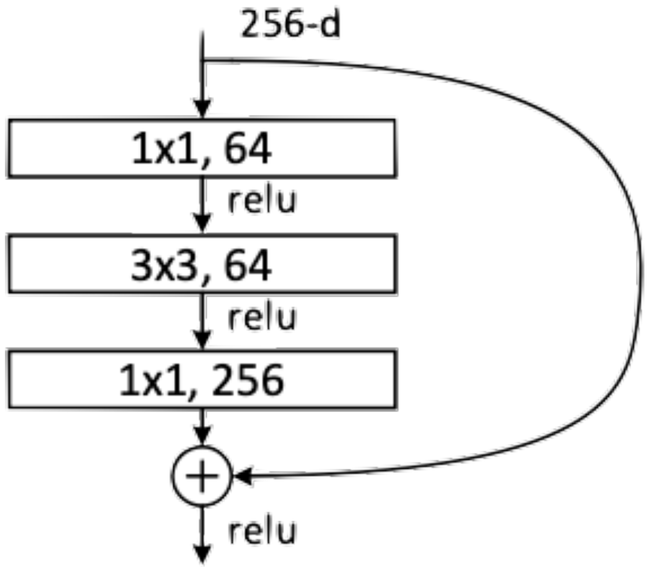


# Pretraining

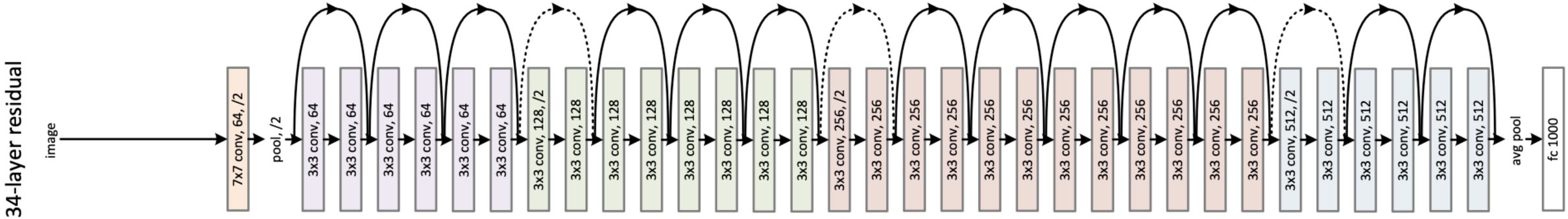
**Sebastian Bayerl, Korbinian Riedhammer**

TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

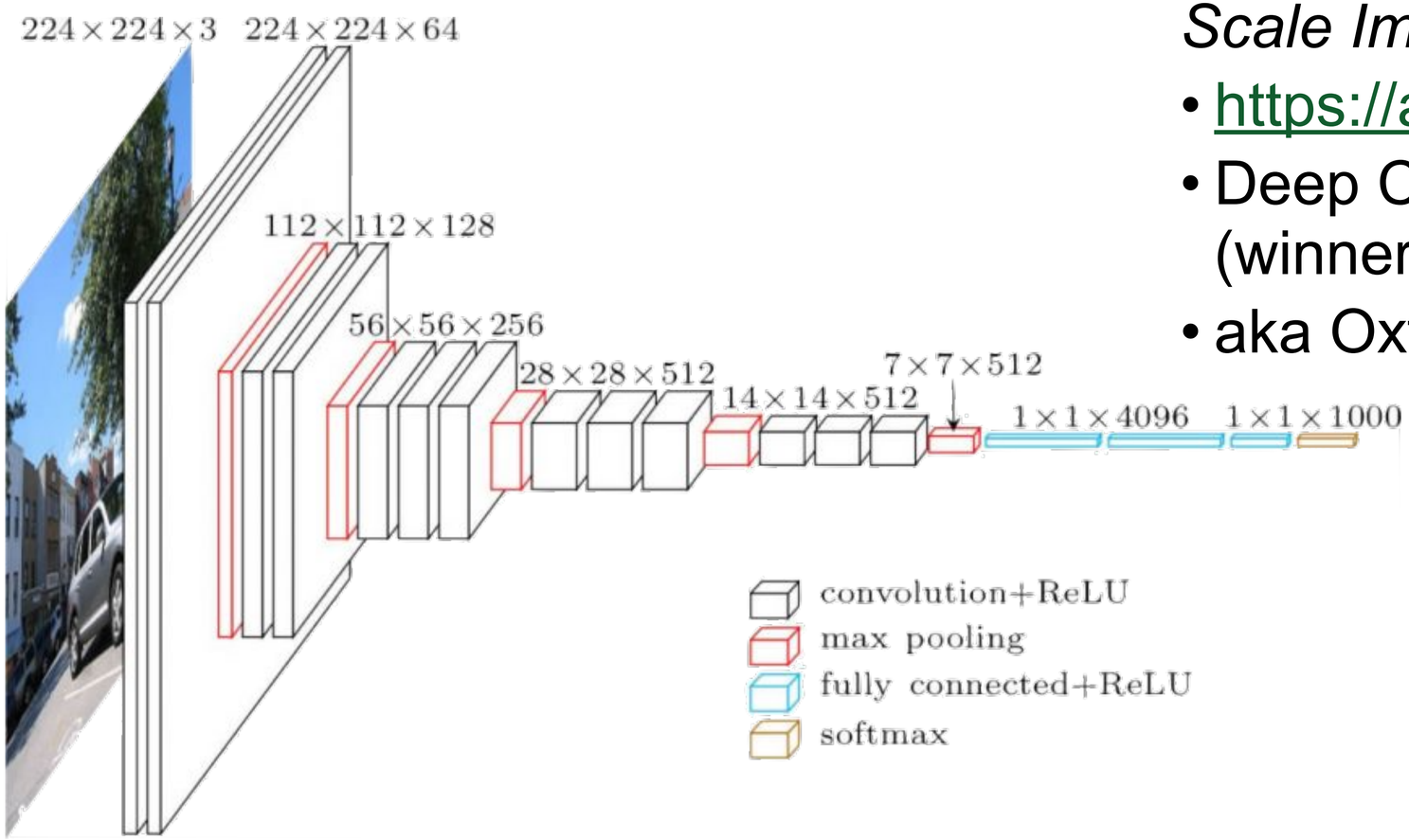
# Image processing




- He et al., 2015: *"Deep Residual Learning for Image Recognition"*
  - https://arxiv.org/pdf/1512.03385.pdf
  - Deep ConvNet with residual connections
  - aka "ResNet" (resnet50)



# Image processing




- Simonyan and Zisserman, 2015: *"Very Deep Convolutional Networks for Large-Scale Image Recognition"*
  - https://arxiv.org/pdf/1409.1556.pdf
  - Deep ConvNet trained on ImageNet (winner 2014!)
  - aka OxfordNet, VGG16

https://neurohive.io/en/popular-networks/vgg16/

# Audio/Speech Processing

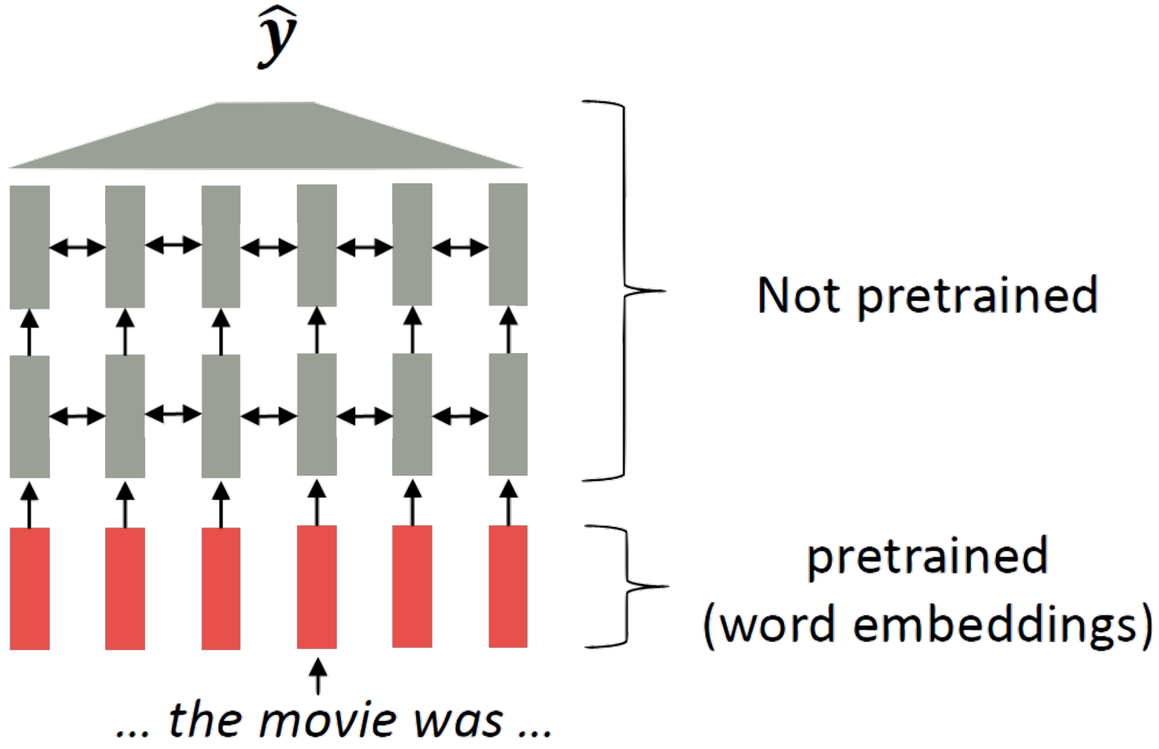
## Representation Learning

- Snyder et al., 2018: “X-Vectors: Robist DNN Embeddings for Speaker Recognition”.
  - https://www.danielpovey.com/files/2018_icassp_xvectors.pdf
  - Use TDNN features and global pooling
  - Contrastive Loss
  - Classification of Age/Sex, medical conditions, etc.

# NLP

- Up to 2017: start with pre-trained word embeddings (no context!, but somewhat semantic relationships, e.g., word2vec)
- Learn the context in LSTM/ RNN etc.

random initialization



# Transformers for Text

**Sebastian Bayerl, Korbinian Riedhammer**

TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

# BERT

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

**Semi-supervised Learning Step**

**Model:**


**Dataset:**

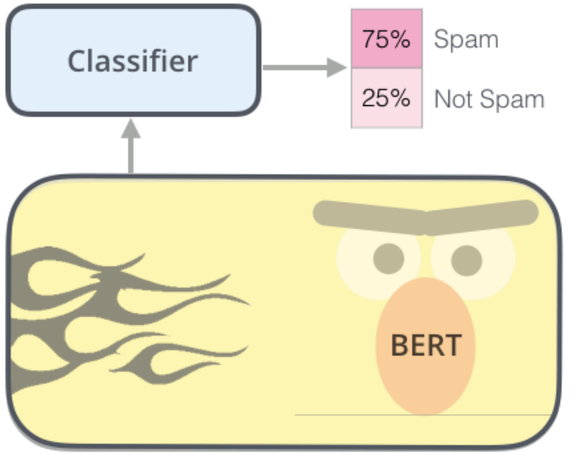

**Objective:**

Predict the masked word
(langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.

**Supervised Learning Step**

**Model:**
(pre-trained in step #1)

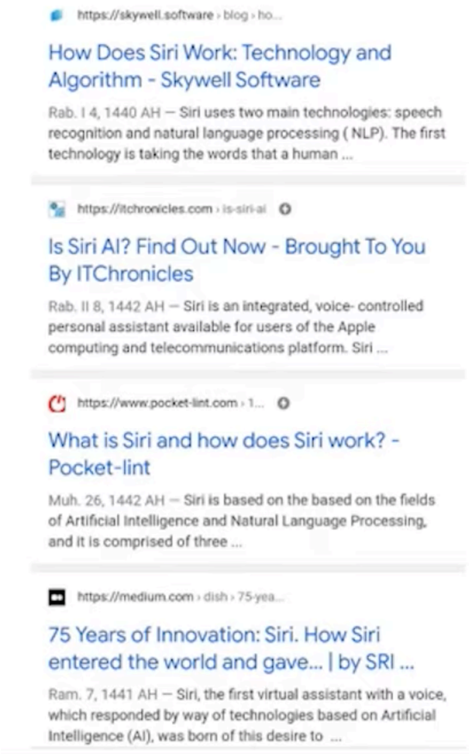



**Dataset:**

| Email message | Class |
|---|---|
| Buy these pills | Spam |
| Win cash prizes | Spam |
| Dear Mr. Atreides, please find attached… | Not Spam |

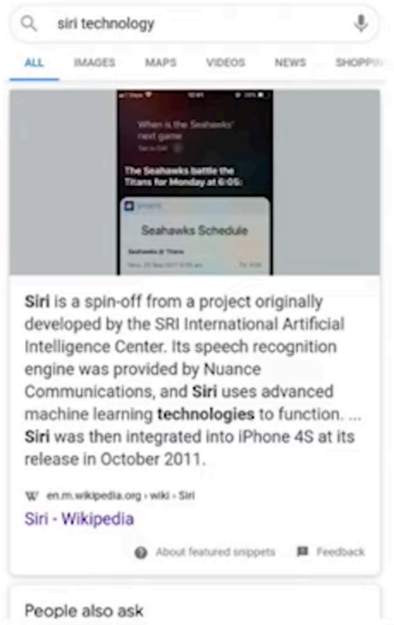
# BERT

1




**Text encoding**
**Similarity retrieval**

2




**Text Summarization**

3


**Question Answering**

# BERT

- Systems and systems like BERT helped to achieve current SOTA of NLU

- 2018/19 Breakthrough word/ language representation through transformers
- Downstream tasks profit from pretraining that already learned the structure of language

- Powers google search -> Notion of order in queries

# BERT

- **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
- Pre-training from unlabeled text
- Conditioning on left and right context of a word
- Resulting in model that can be fine tuned
- Just one additional layer needed
- Multiple NLP tasks:
  - Question answering
  - Sentiment analysis
  - Answer generation
  - Summarization
  - …

# BERT – Contextualized word embeddings

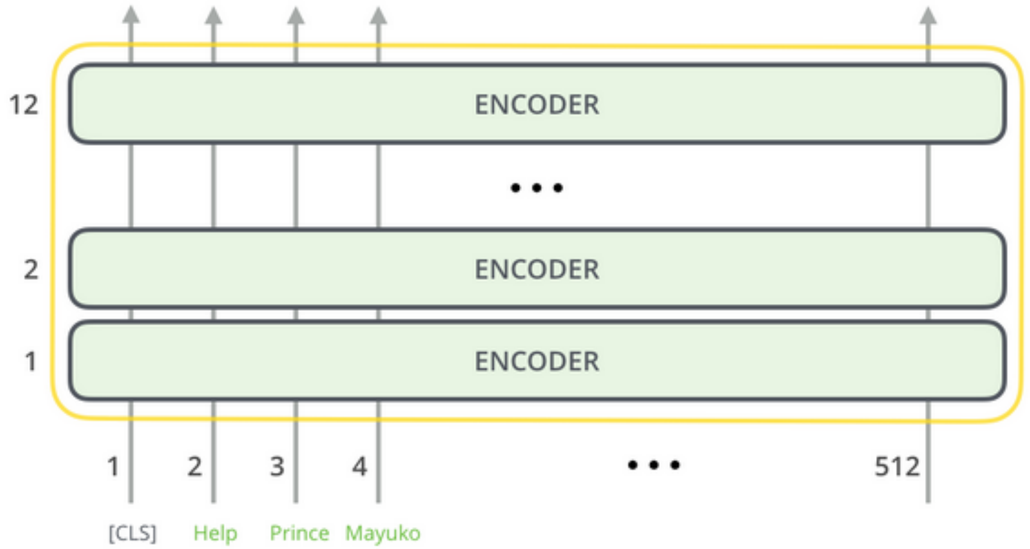
## prerequsites

- ELMo (Embeddings from language models)
- Needs to see the complete sentence
- Trained to predict the next word in a sequence (Language Modelling ;) )
- Forward and Backward language models (LSTM-based)
- -> contextualized Word embedding by grouping together the hidden states

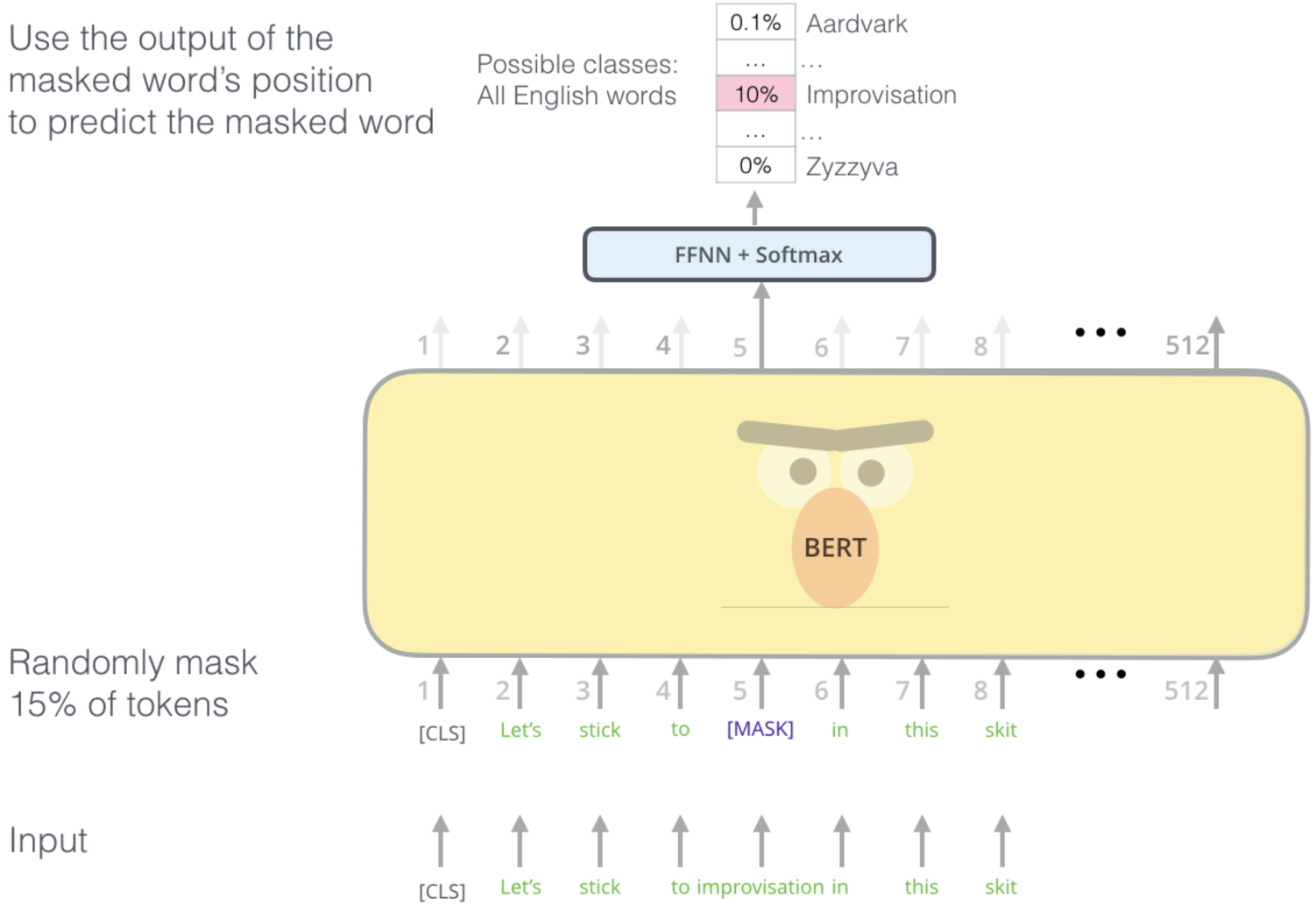
# BERT – Model architecture

- BERT in essence is a trained Transformer Encoder stack
- 12 / 24 encoder layers (BERT Base, BERT Large)

- Inputs start with a class token (task-dependent),
  each layer applies self-attention

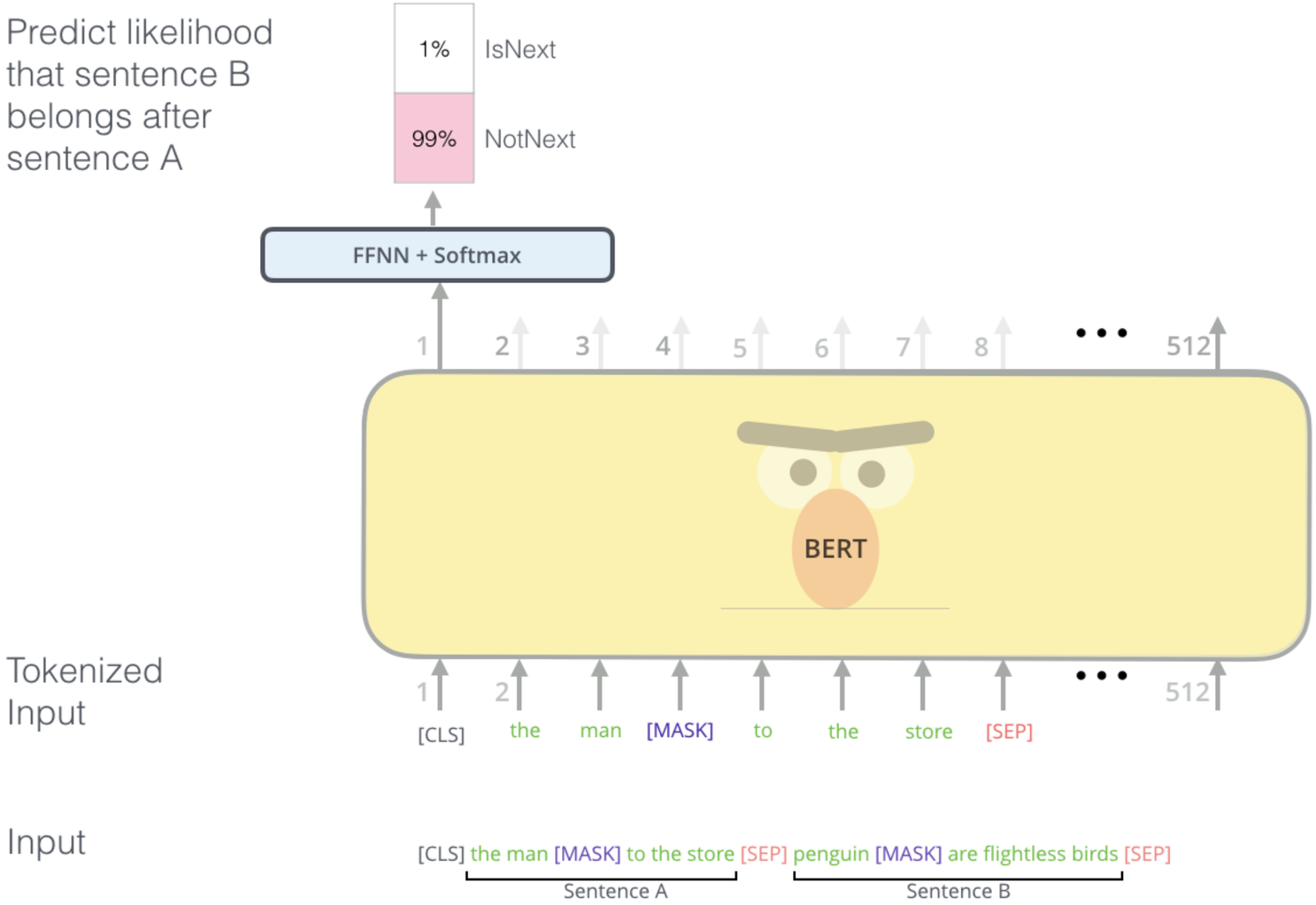


# BERT – Model architecture




https://jalammar.github.io/illustrated-bert/

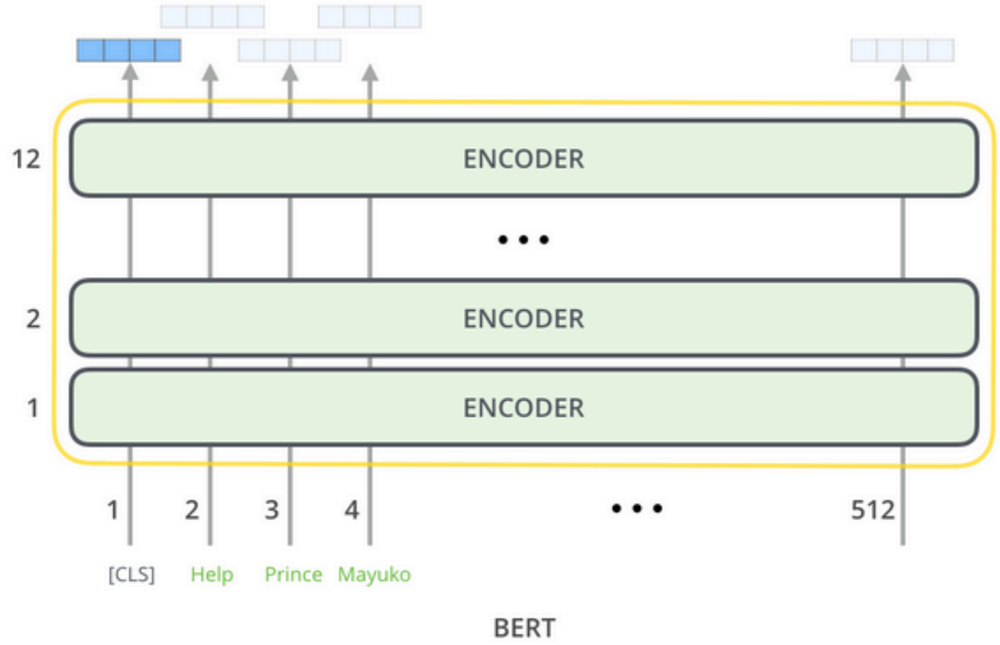
# BERT- two sentence task




https://jalammar.github.io/illustrated-bert/

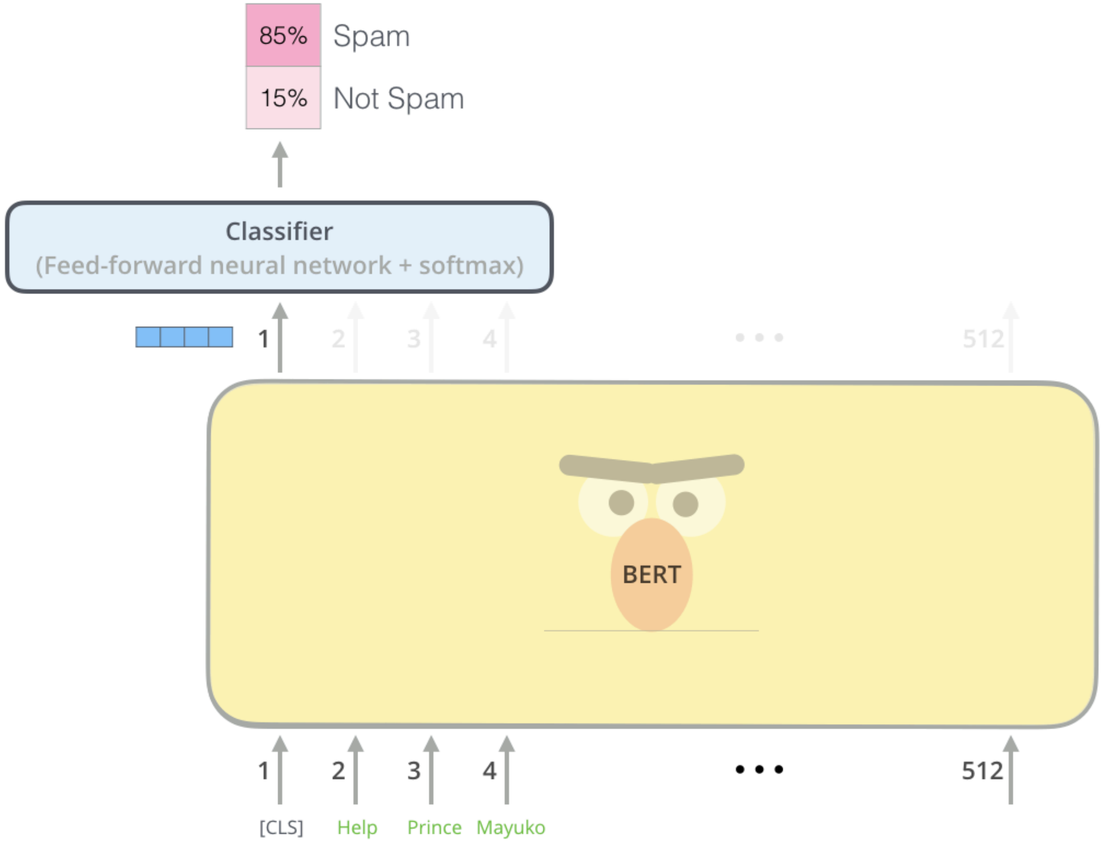
# BERT – Model architecture

- Each position outputs a vector of size 768 (BERT Base)



# BERT – Model architecture

- Vectors can now be used for downstream tasks, e.g. SPAM classification

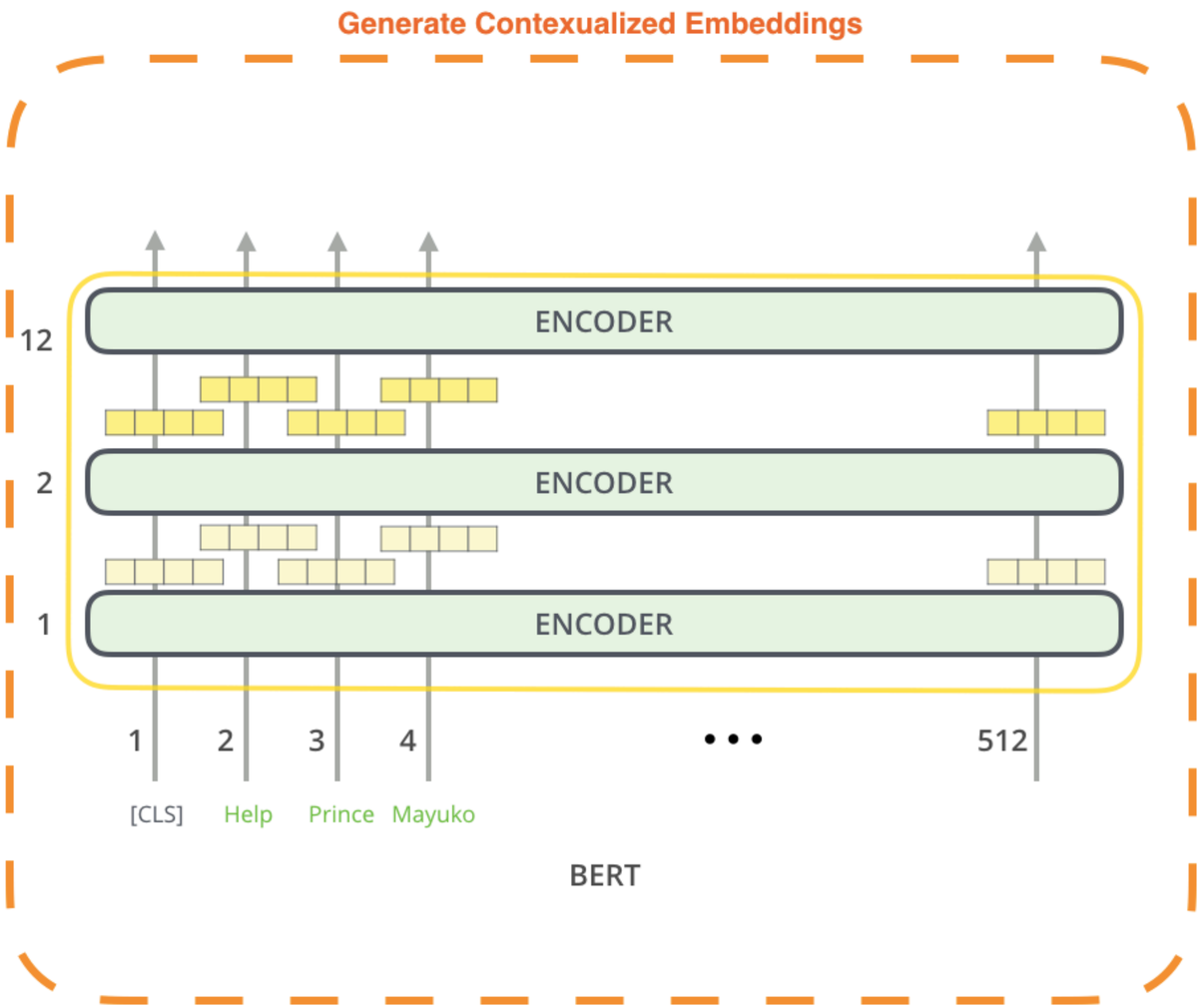



https://jalammar.github.io/illustrated-bert/

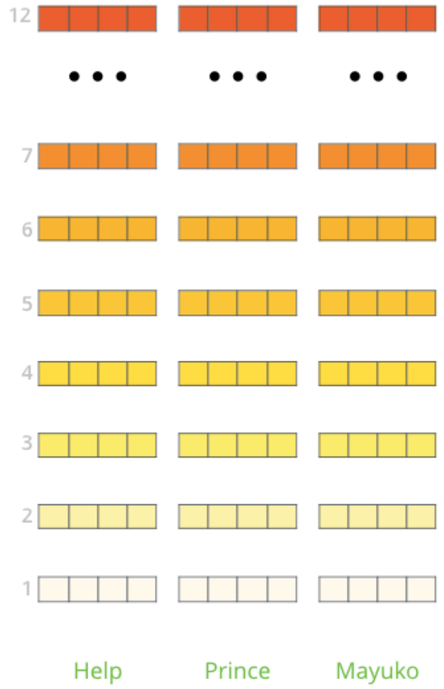
# BERT – Model architecture

- BERT uses contextualized word-embeddings
- Trained with transformer encoders
- Bidirectional would allow each word to indirectly see itself
- -> masked language modelling
- Next Sentence Prediction (NSP)

# BERT- Feature Extrator




The output of each encoder layer along each token's path can be used as a feature representing that token.




**But which one should we use?**

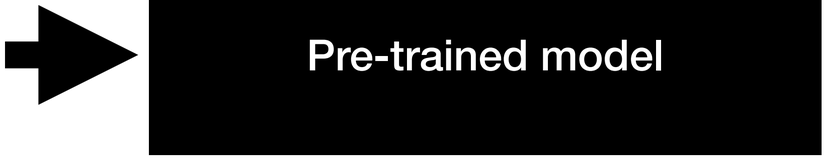
https://jalammar.github.io/illustrated-bert/

# What about audio?

How to learn good representations of audio data?

# What would be a good representation?

## Speech recognition

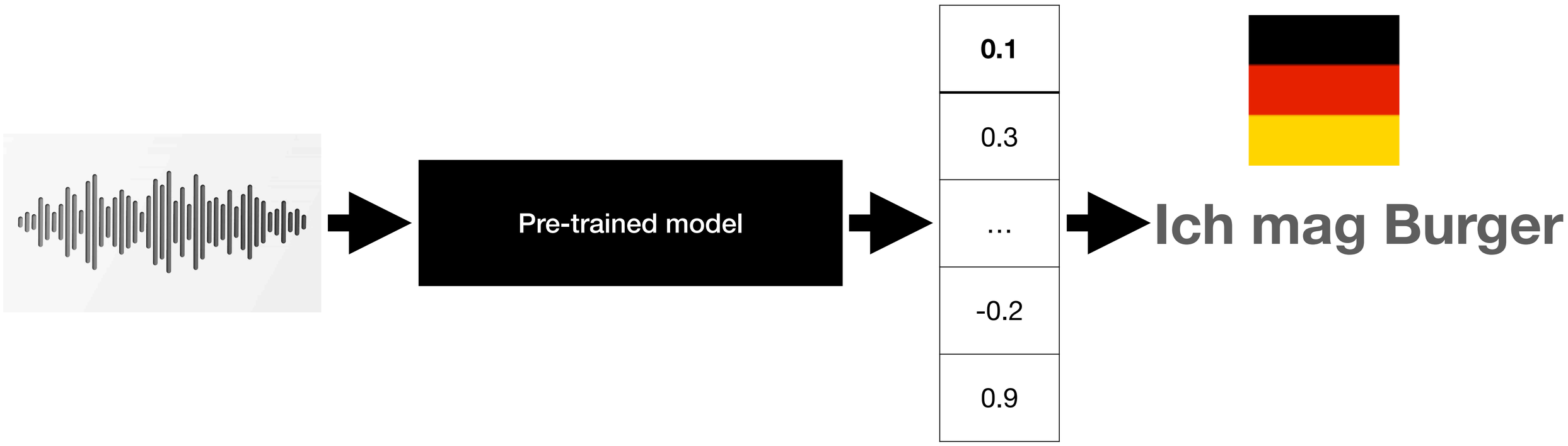



| |
|---|
| **0.1** |
| 0.3 |
| … |
| -0.2 |
| 0.9 |

I like burgers

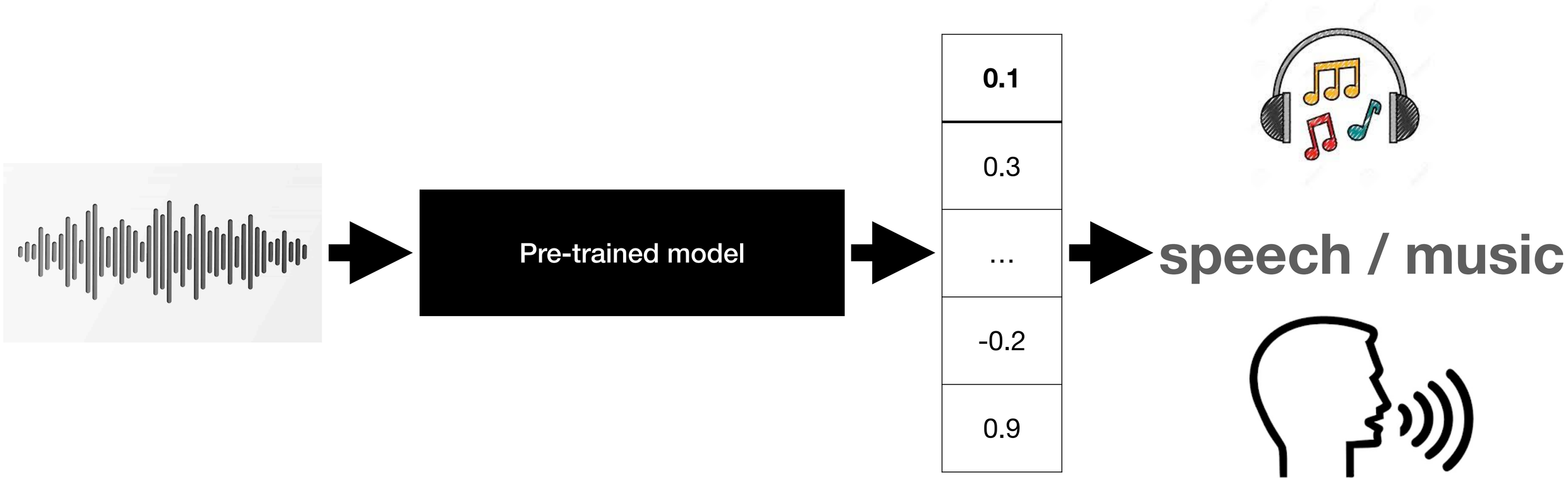
# What would be a good representation?

## Speech translation



# What would be a good representation?

## Audio event detection



# Possible Paradigms

## Unsupervised / Self-supervised Pre-training

- Learn good representations without labels
- NLP: Predict occluded parts of sentences
- Vision/ Image: make representations invariant to augmentations

# Classical ASR

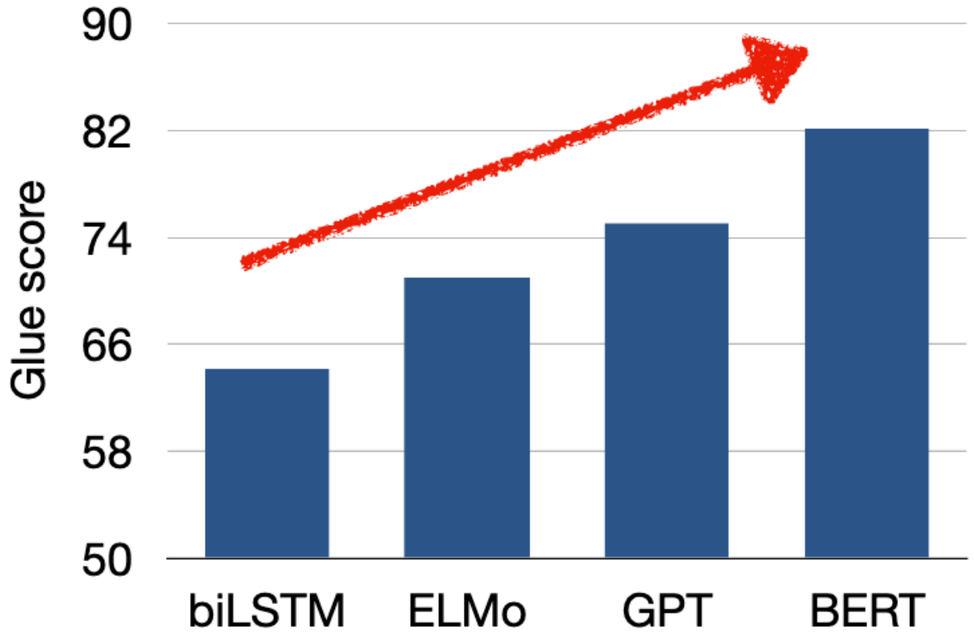
## Training ASR models

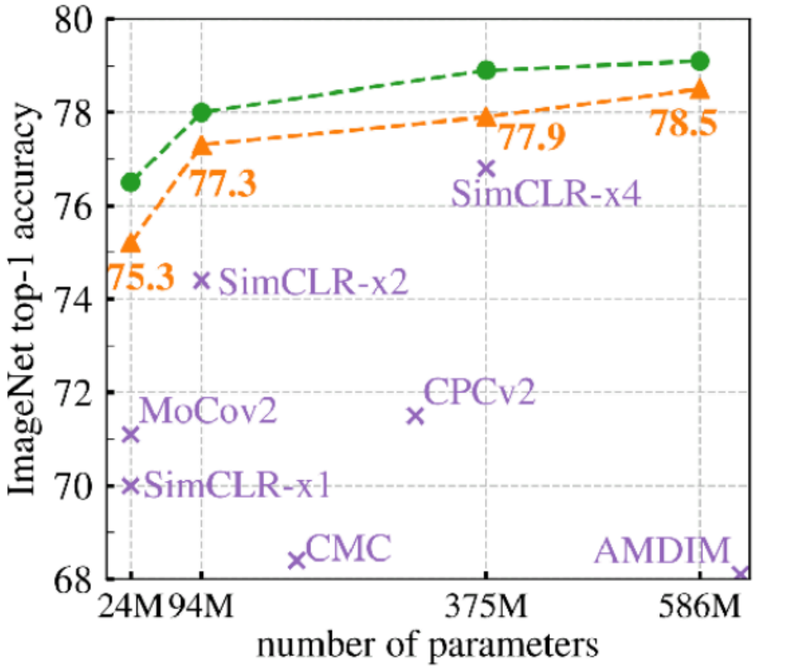
**I like burgers**

- Training on 1000s of hours of data for good systems
- Many languages, dialects, domains etc.

# Pretraining

## In other fields…



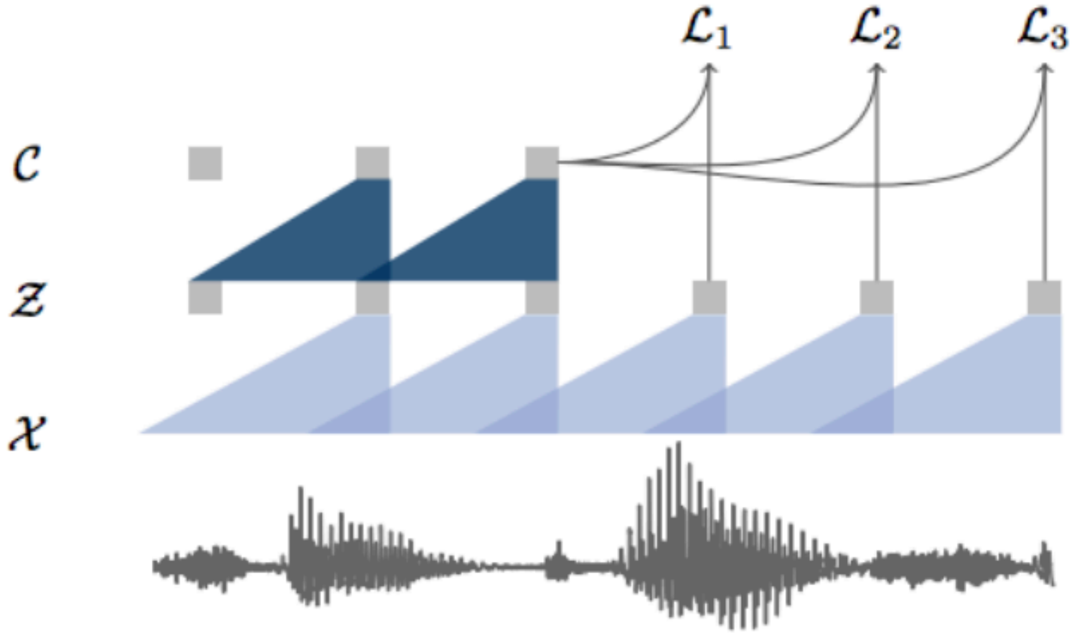
# wav2vec 2.0

Learning good representations of
audio data
from unlabeled audio!

# wav2vec 2.0

## Gradual development wav2vec

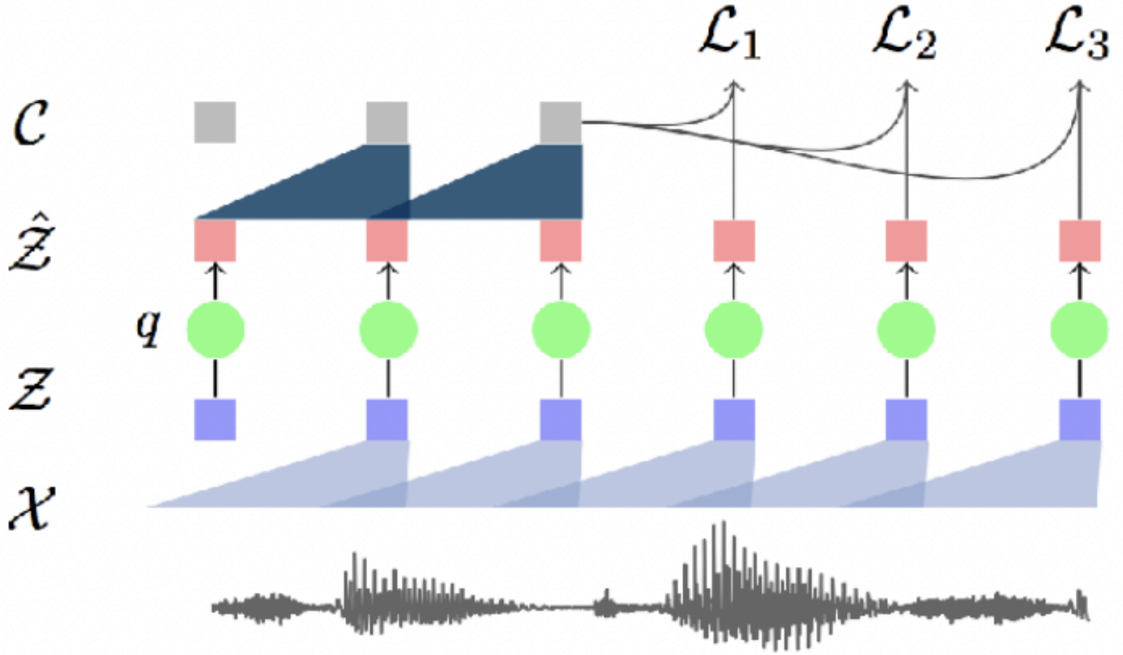
- fully convolutional
- binary cross entropy loss
- representations for ASR tasks



# wav2vec 2.0

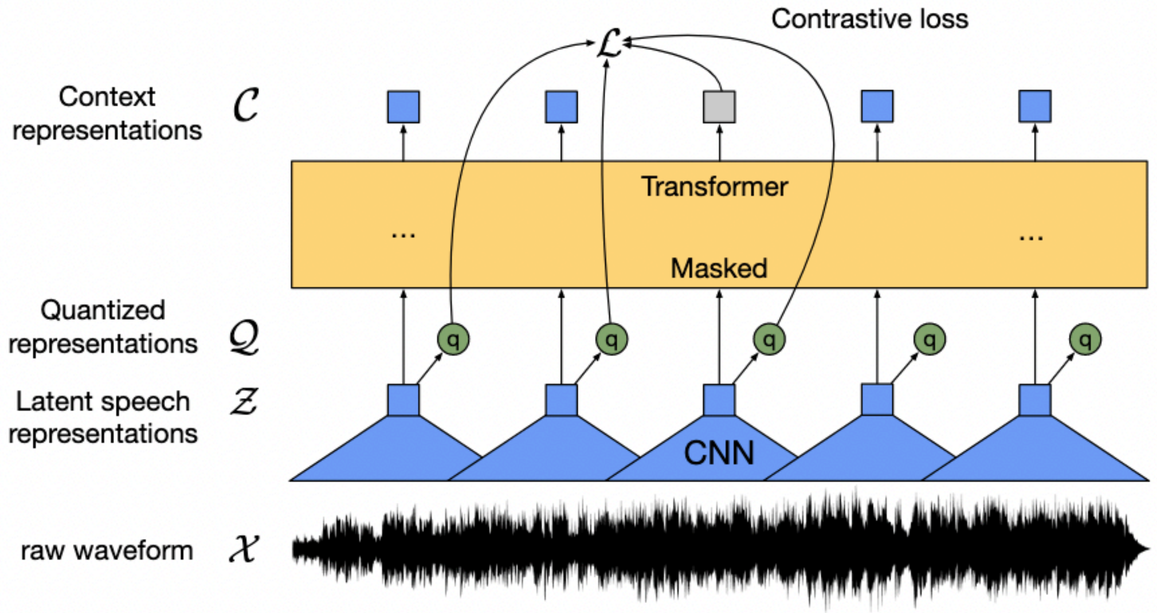
## gradual step: vq-wav2vec

- vector quantize to discover **discrete latent speech representations**
- Learn contextualized representations on top of quantized speech
- Product quantization of discrete units
- Quantization via Gumbel and K-means
- VQ enables use of NLP-style models
- Different to vq-vae: context in latent space prediction vs. data reconstruction



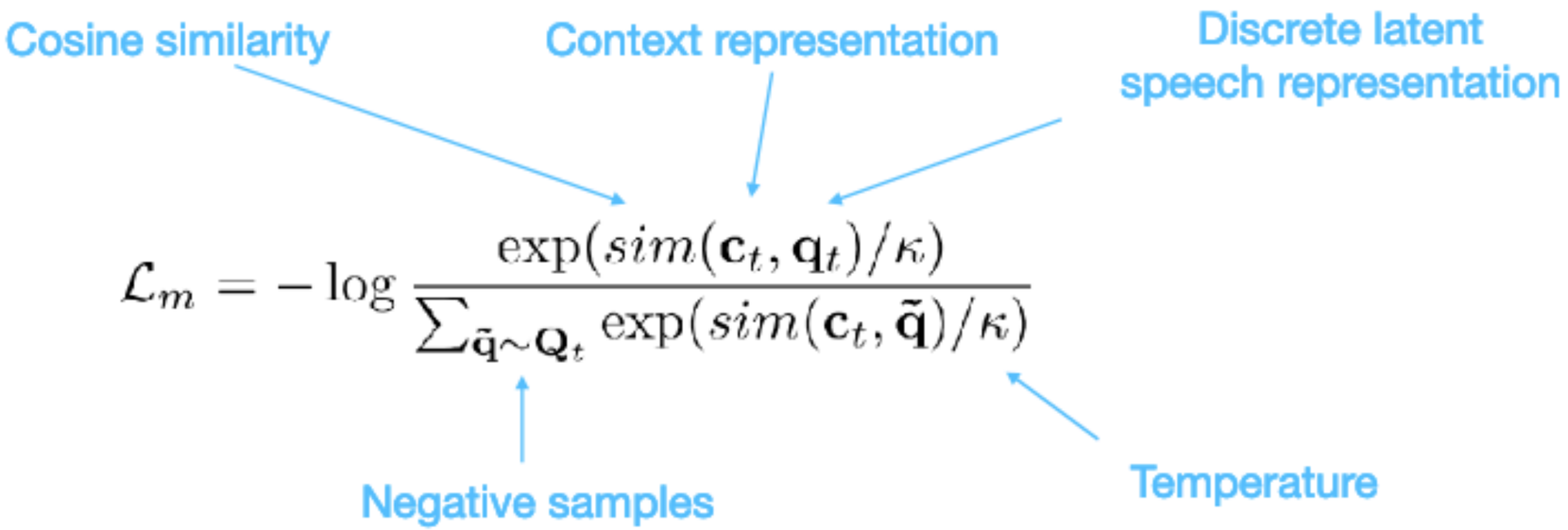
# wav2vec 2.0

- Joint VQ & context **representation learning**
- Bi-directional **contextualized representations**
- **Contrastive task**
- Vector quantized targets
- **Fine-tuned on labeled data**



# wav2vec 2.0

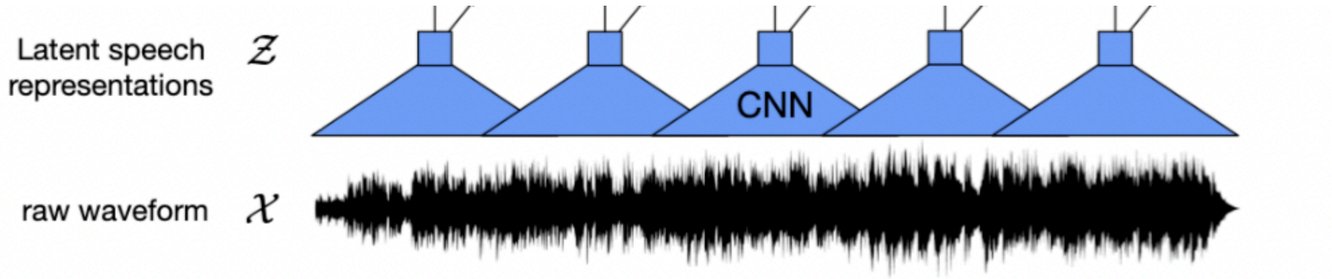
## Training objective




- Code book penalty to encourage more codes to be used

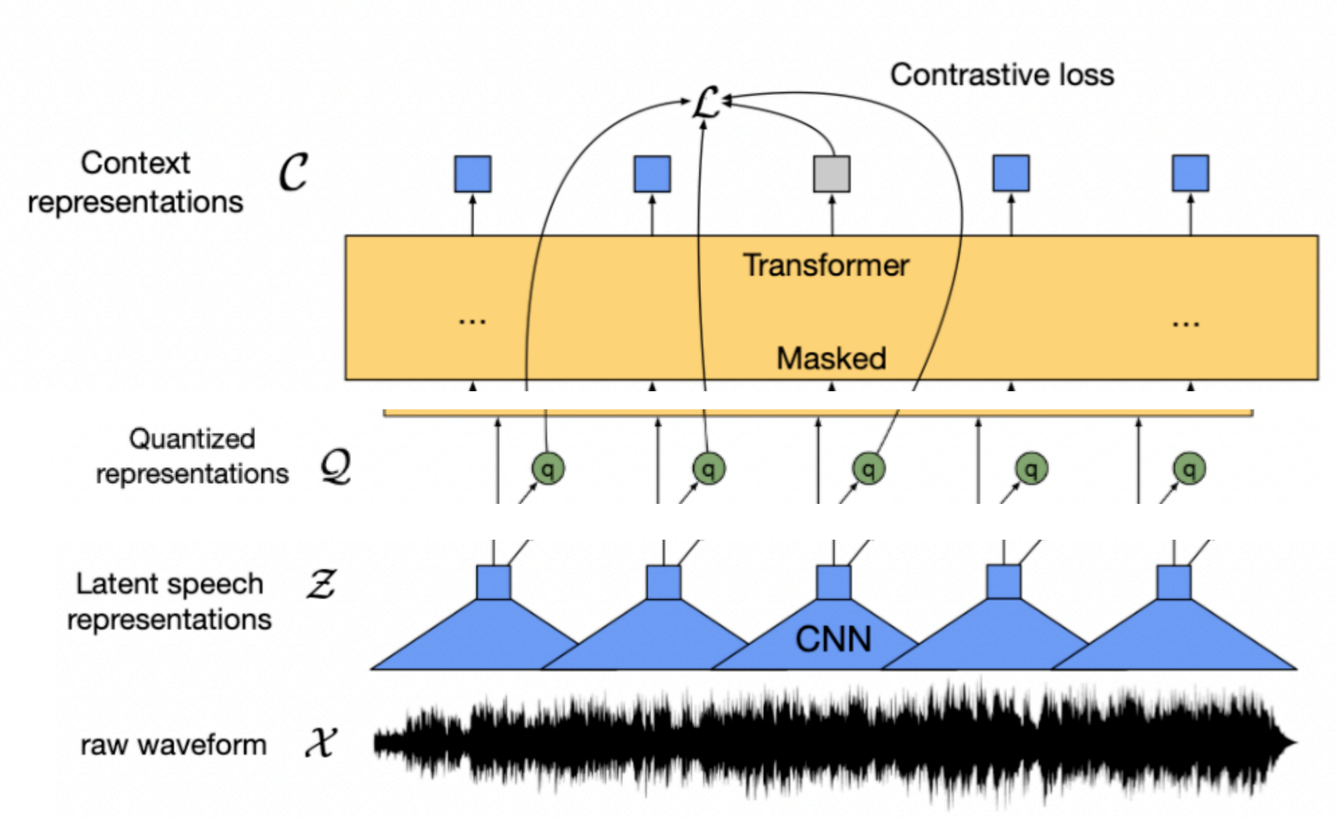
# wav2vec 2.0

- input raw waveform
- multi-layer convolutional feature encoder



# wav2vec 2.0

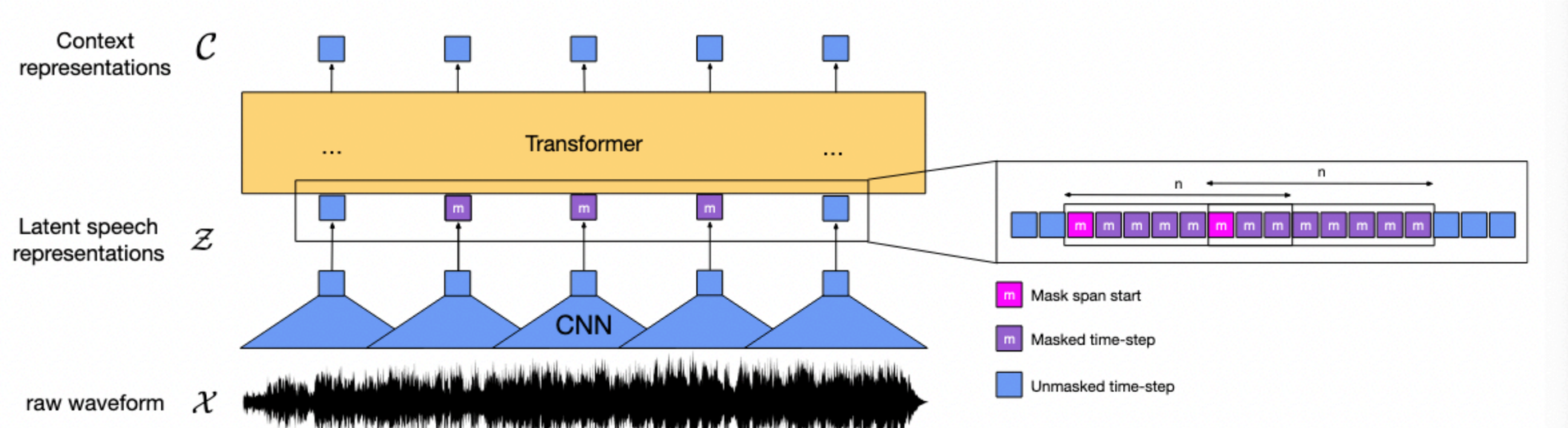
- input raw waveform
- multi-layer convolutional feature encoder
- quantization
- masking of quantized sequence (similar to MLM)
- Transformer Network to build contextualized representations
- contrastive loss —> true latent vs distractors



# wav2vec 2.0

## Masking

- Sample starting points for masks without replacement, then expand to 10 time steps
- spans can overlap
- For a 15s sample, ~49% of time steps masked with an average span length of ~300ms (BERT: ~15% of tokens)
-



# wav2vec 2.0

## Fine-tuning for ASR

- Add a single linear projection on top into target vocab and train with CTC loss with a low learning rate (CNN encoder is not trained/ frozen).
- Use modified SpecAugment in latent space to prevent early overfitting
- Uses wav2letter decoder with the official 4gram LM and Transformer LM

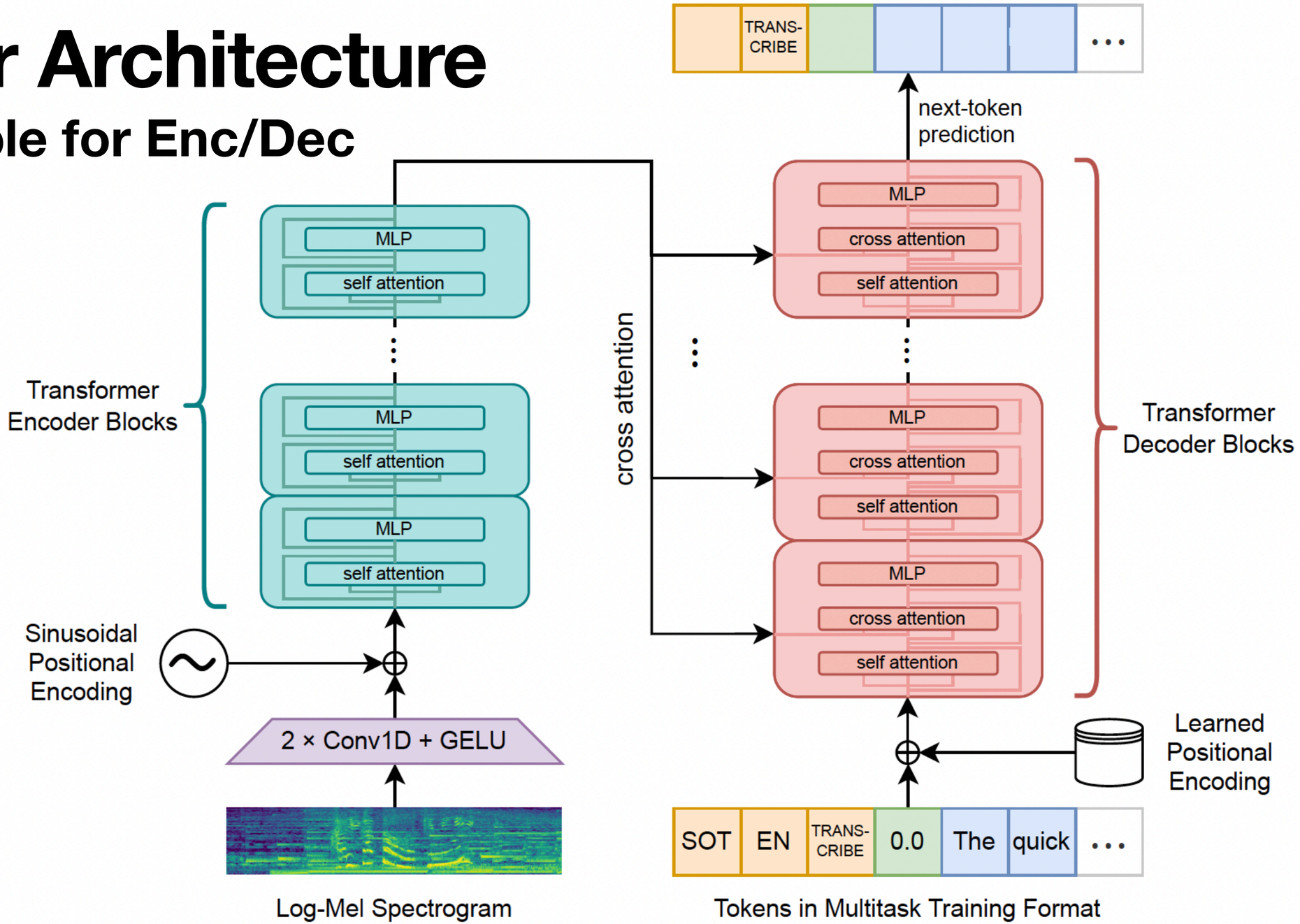
# Outlook

## Encoders, Decoders and Encoder-Decoder

- Encoders (eg. BERT)
  - Typically trained as contrastive (unsupervised) or CTC (sequence)
  - Work great as representation
- Decoders (eg. GPT)
  - Typically trained as autoregressive models
  - Work great as (constrained) generators of new sequences
- Encoder/Decoders (eg. Whisper)
  - Combine best of both worlds :-)

# Whisper Architecture

## ...as example for Enc/Dec



# Sources

- http://michaelauli.github.io/talks/wav2vec-ssl.pdf
- wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations
  Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
  Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova