

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

Terminology

Self supervised learning

- Artificial Neural Network Method
- between supervised and unsupervised learning
- First solve task with ‘pseudo-labels’ (intialization)
- Then supervised / or unsupervised training
- contrastive SSL
 - contrastive loss, show positive examples and unlabeled negative samples
- non-contrastive SSL
 - only positive examples
 - needs a non-trainable predictor on the target side

Self training



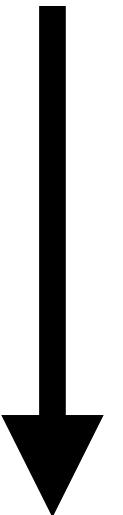
I like cats

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

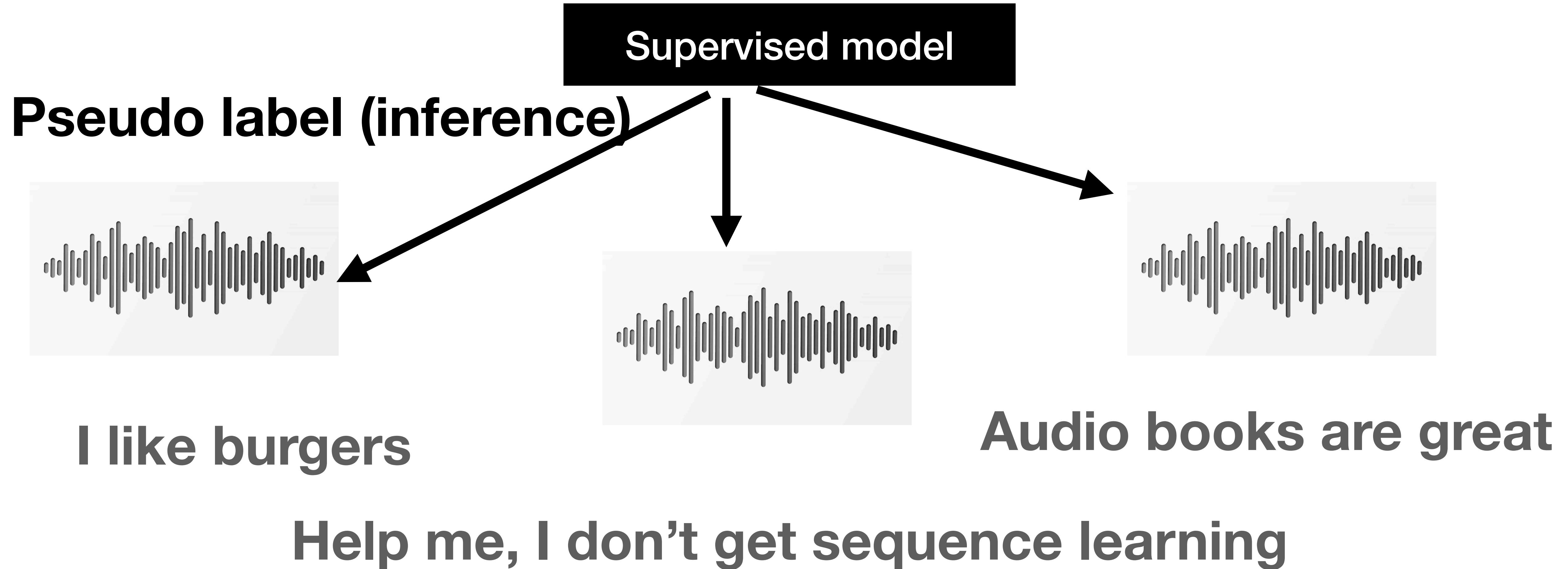


Help me, I don't get sequence learning



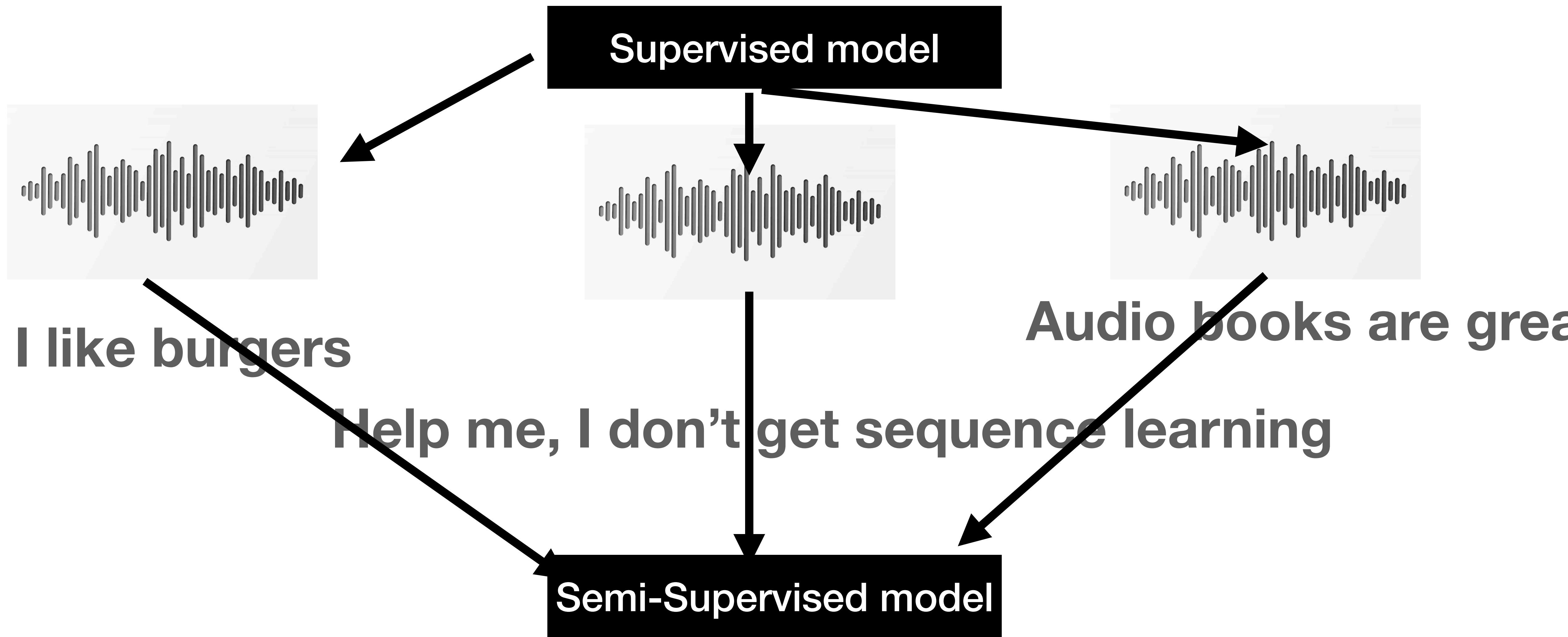
Audio books are great

Self training



Get quality unlabeled
audio data

Self training



Pretraining

Sebastian Bayerl, Korbinian Riedhammer



TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Image processing

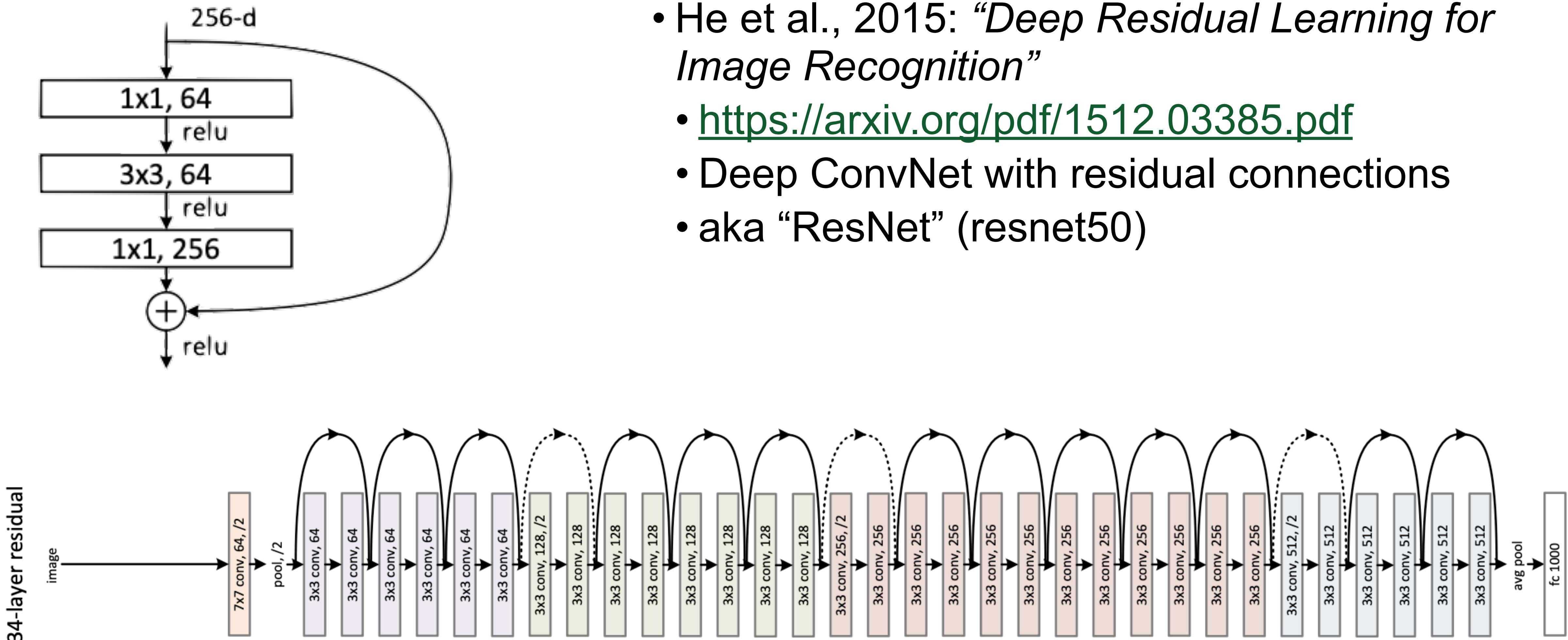
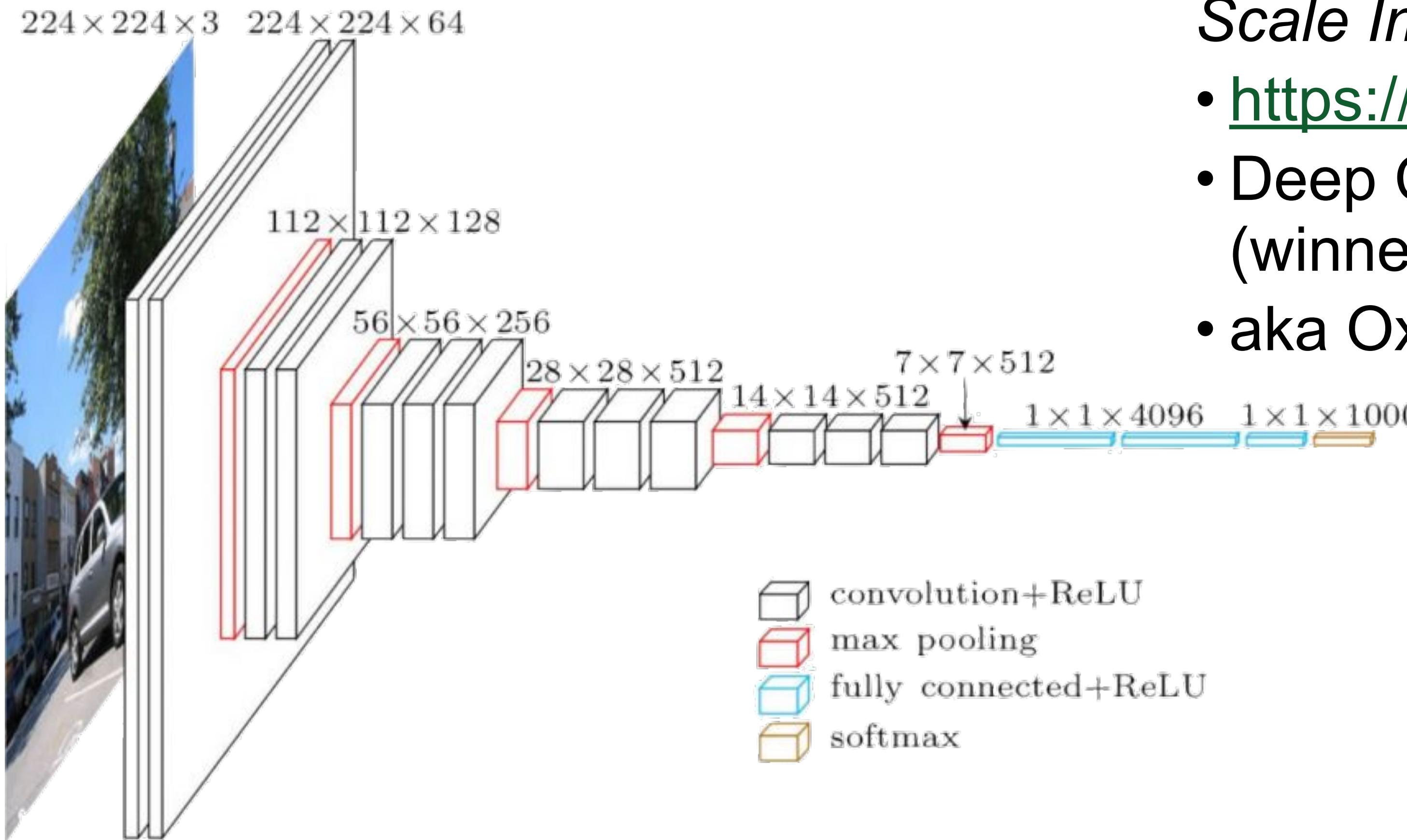


Image processing



- 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

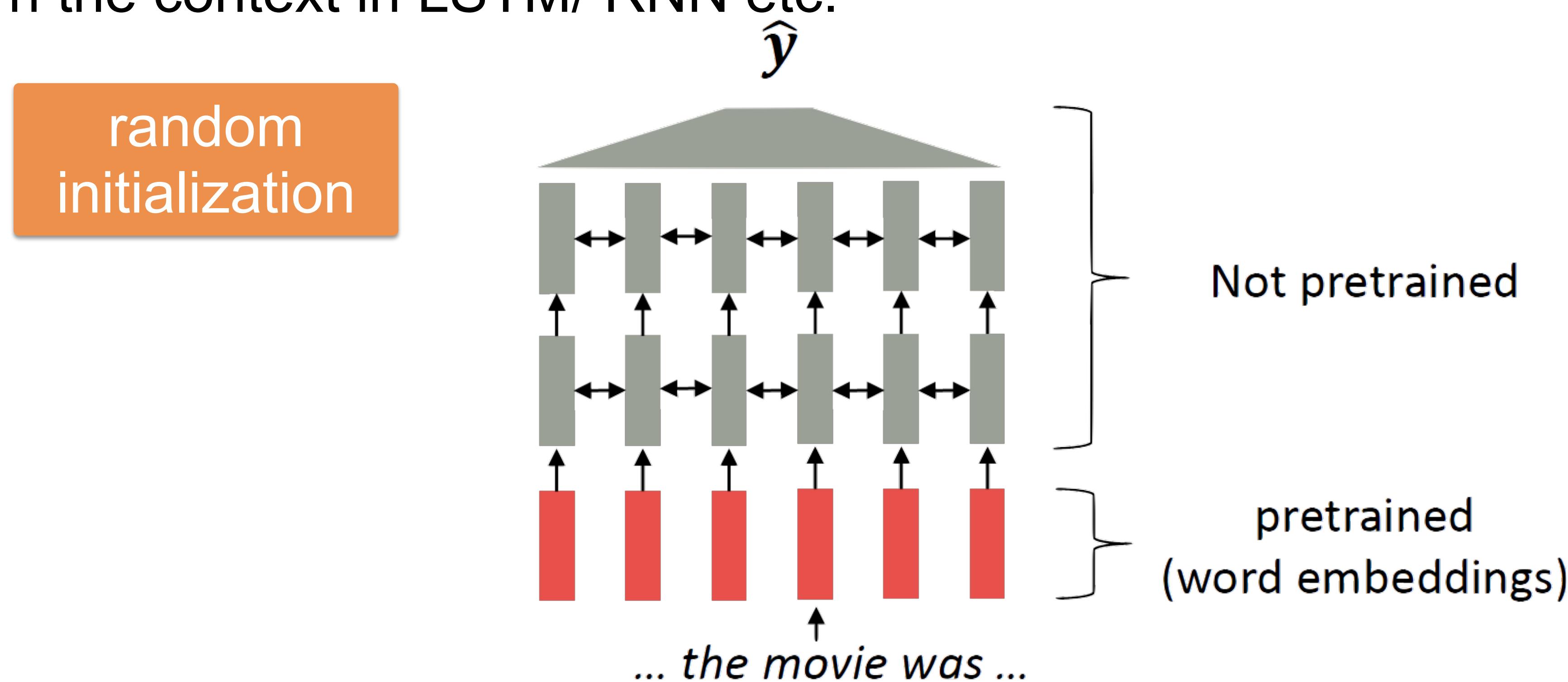
Audio/Speech Processing

Representation Learning

- Snyder et al., 2018: “X-Vectors: Robust 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

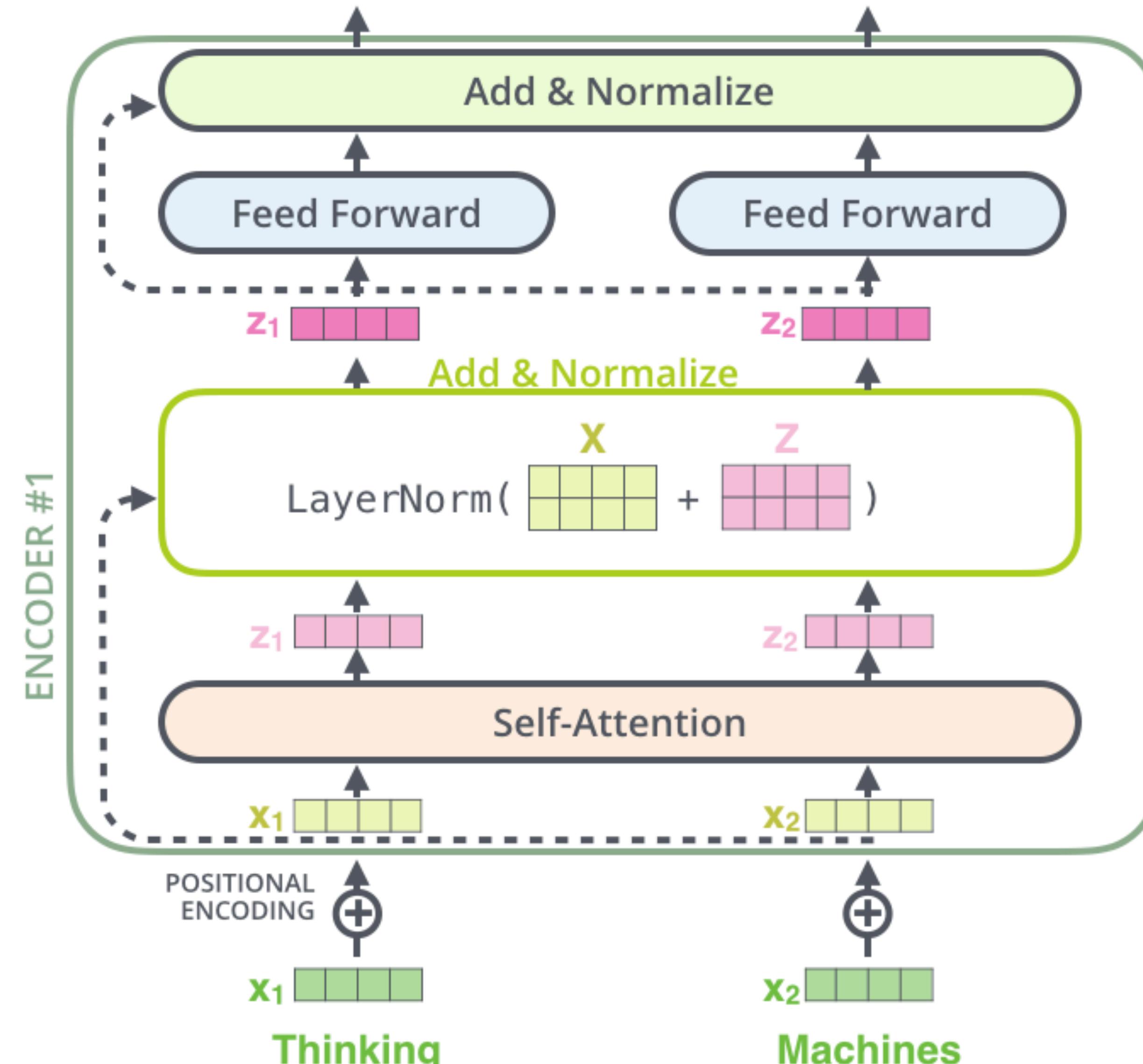
- Up to 2017: start with pre-trained word embeddings (no context!, but somewhat semantic relationships, e.g., word2vec)
- Learn the context in LSTM/ RNN etc.



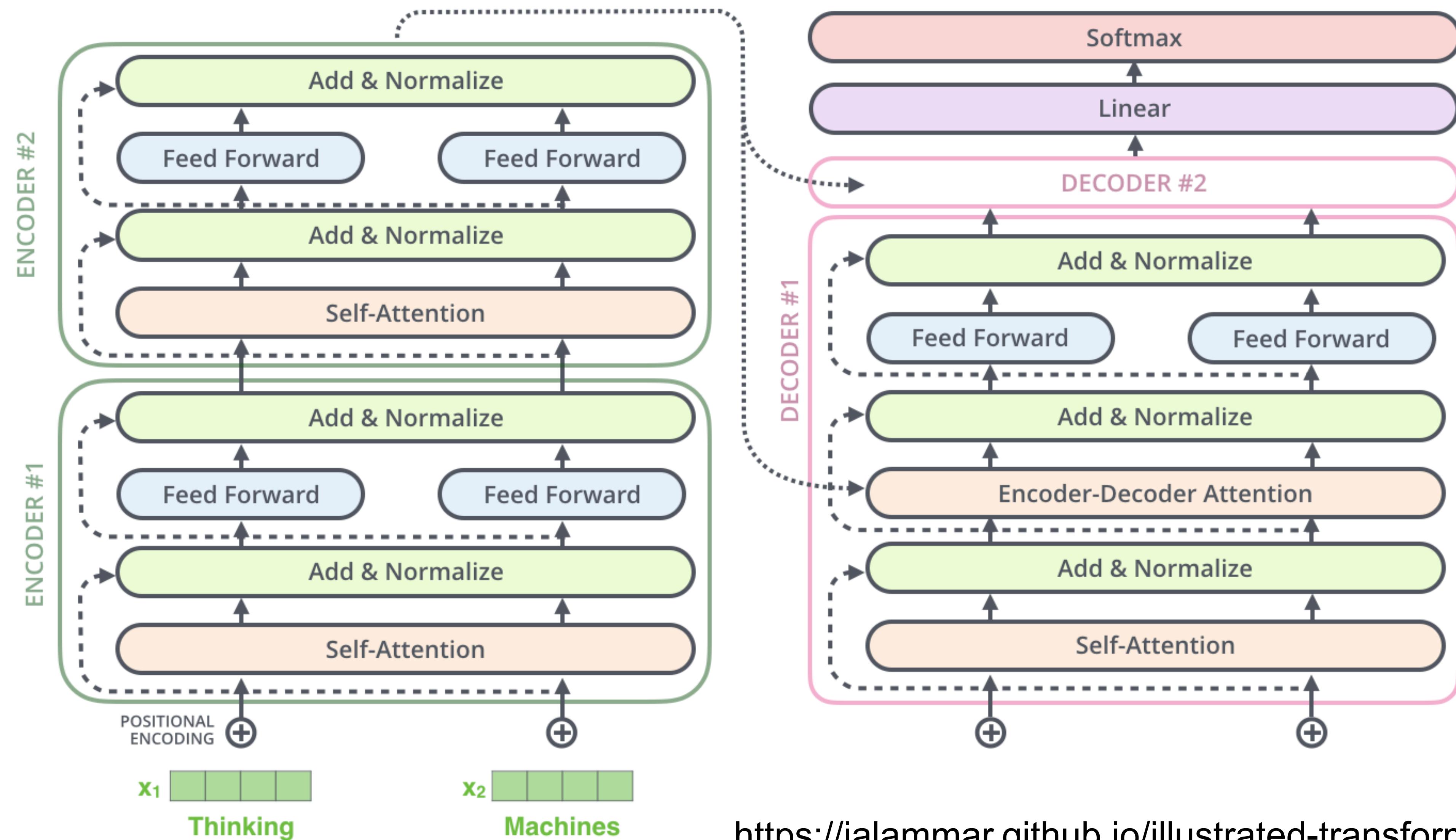
Transformer architectures

Sebastian Bayerl, Korbinian Riedhammer

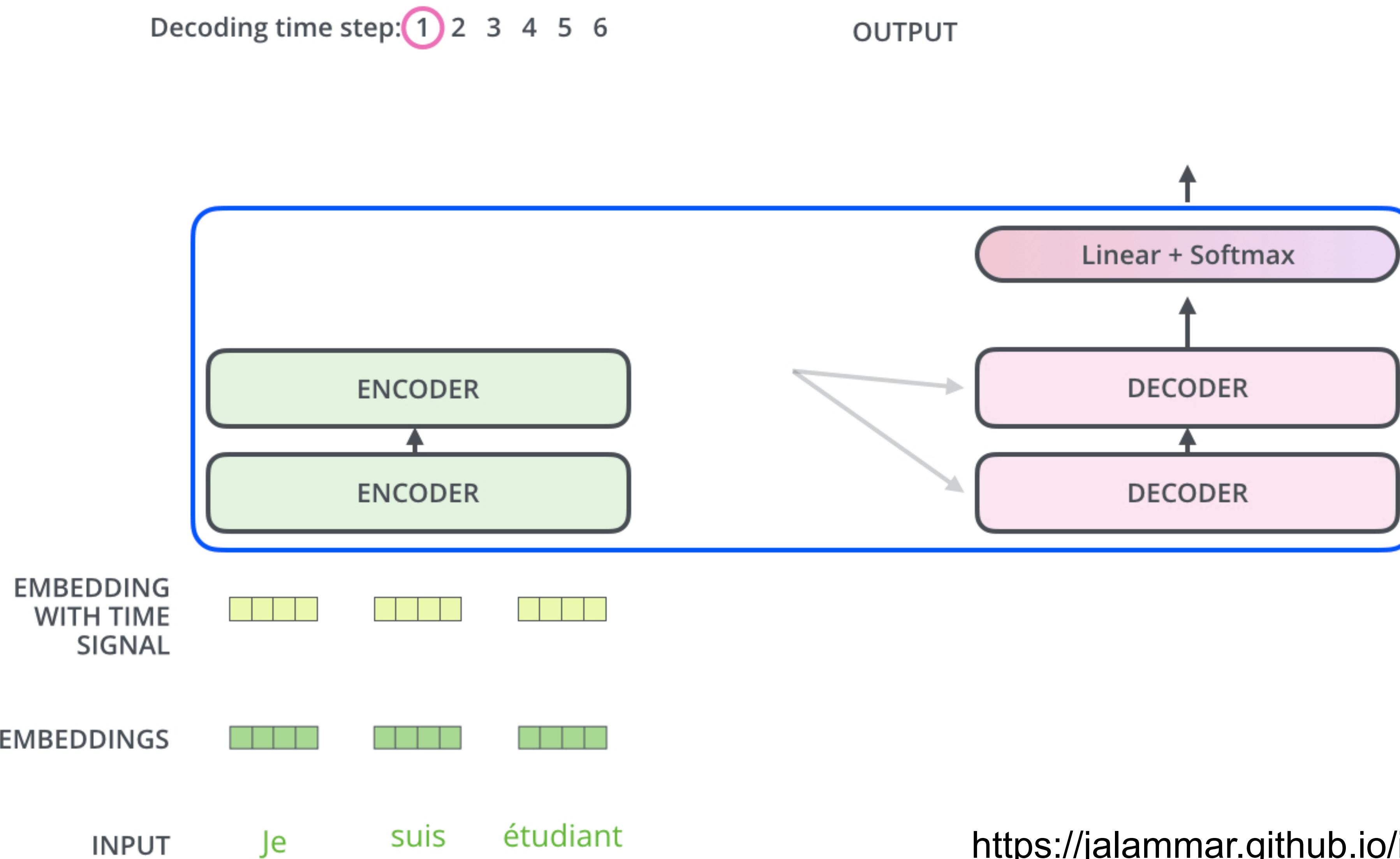
Residuals and LayerNorm



Decoder



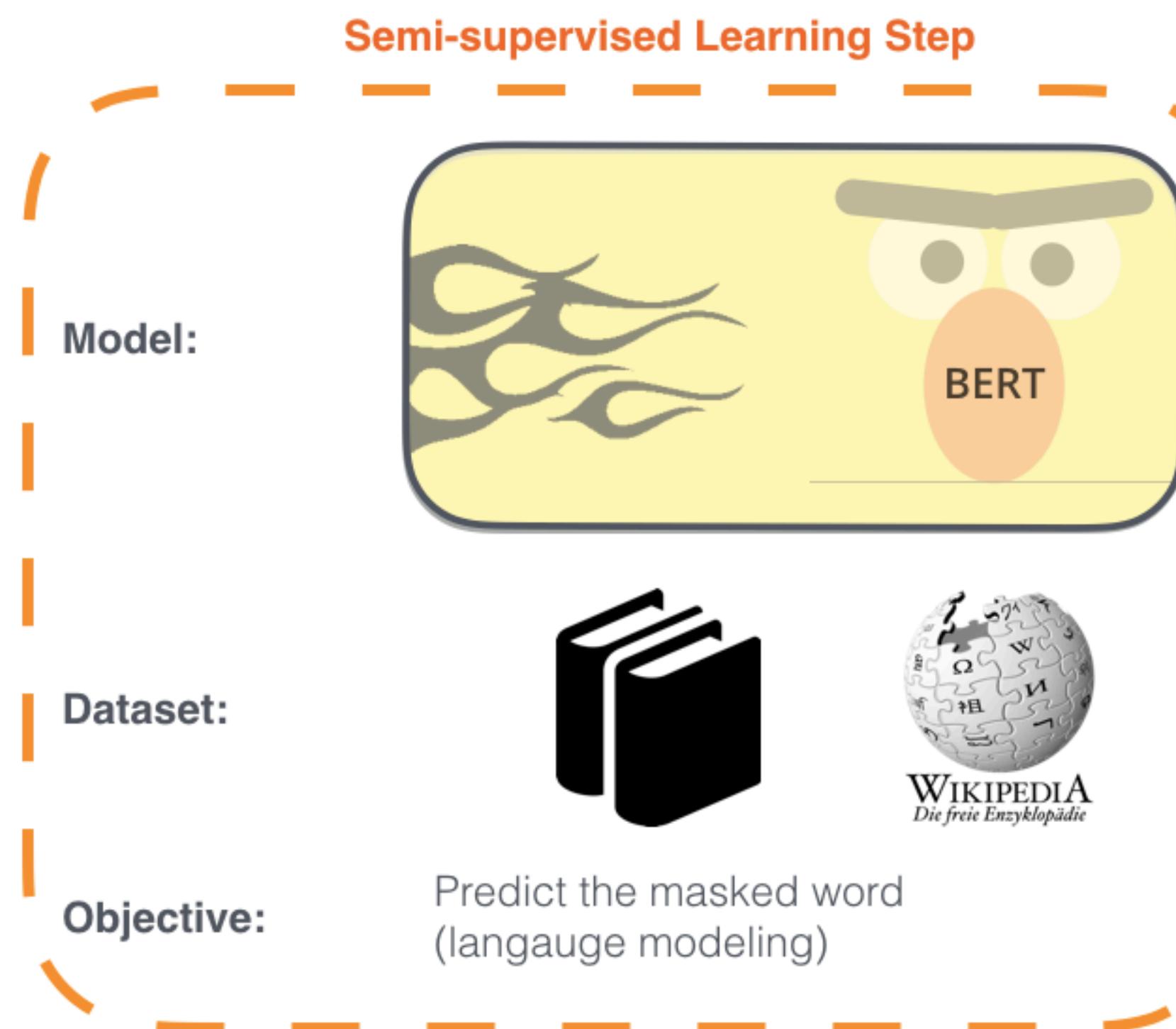
Transformer: Animated



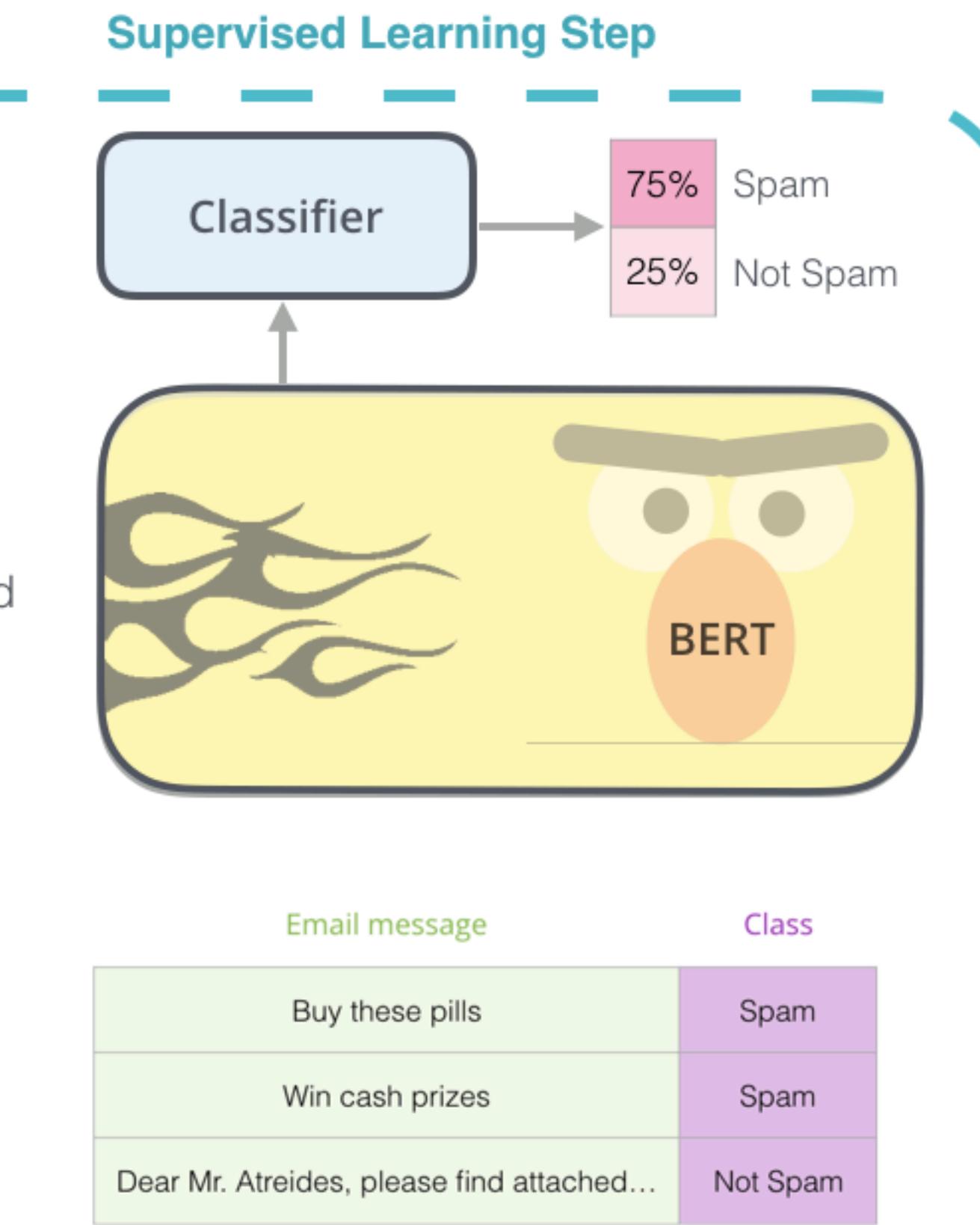
BERT

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.



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



BERT

1

 <https://skywell.software> › blog › ho...

How Does Siri Work: Technology and Algorithm - Skywell Software

Rab. 14, 1440 AH — Siri uses two main technologies: speech recognition and natural language processing (NLP). The first technology is taking the words that a human ...

 <https://itchronicles.com> › is-siri-al

Is Siri AI? Find Out Now - Brought To You By ITChronicles

Rab. II 8, 1442 AH — Siri is an integrated, voice- controlled personal assistant available for users of the Apple computing and telecommunications platform. Siri ...

 <https://www.pocket-lint.com> › 1...

What is Siri and how does Siri work? - Pocket-lint

Muh. 26, 1442 AH — Siri is based on the based on the fields of Artificial Intelligence and Natural Language Processing, and it is comprised of three ...

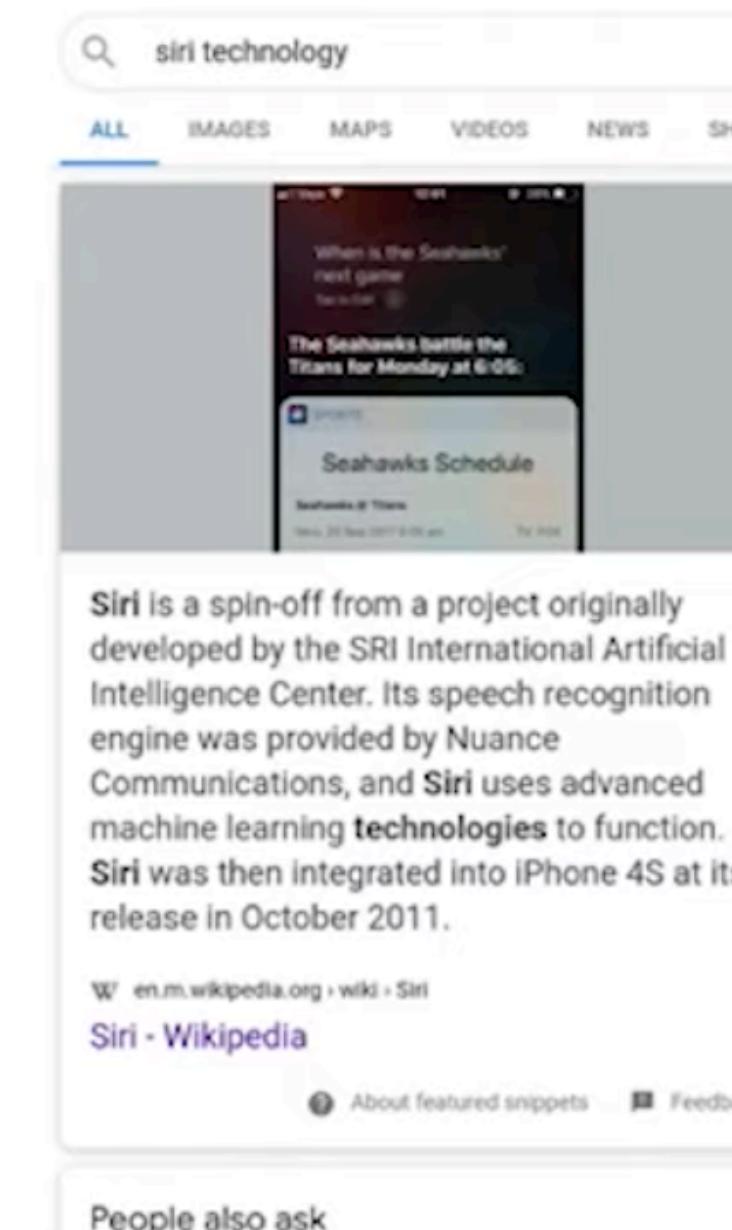
 <https://medium.com> › dish › 75-yea...

75 Years of Innovation: Siri. How Siri entered the world and gave... | by SRI ...

Ram. 7, 1441 AH — Siri, the first virtual assistant with a voice, which responded by way of technologies based on Artificial Intelligence (AI), was born of this desire to ...

**Text encoding
Similarity retrieval**

2



**Text
Summarization**

3

Hebrew	Italian
Japanese	Korean
Malay	Norwegian
Portuguese	Russian
Spanish	Swedish
Thai	Turkish
Type	Intelligent personal assistant
Website	www.apple.com/siri/

Siri is a spin-off from a project originally developed by the SRI International Artificial Intelligence Center. Its speech recognition engine was provided by Nuance Communications, and Siri uses advanced machine learning **technologies** to function. ... Siri was then integrated into iPhone 4S at its release in October 2011.

W en.m.wikipedia.org › wiki › Siri
[Siri - Wikipedia](#)

About featured snippets Feedback

People also ask

**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

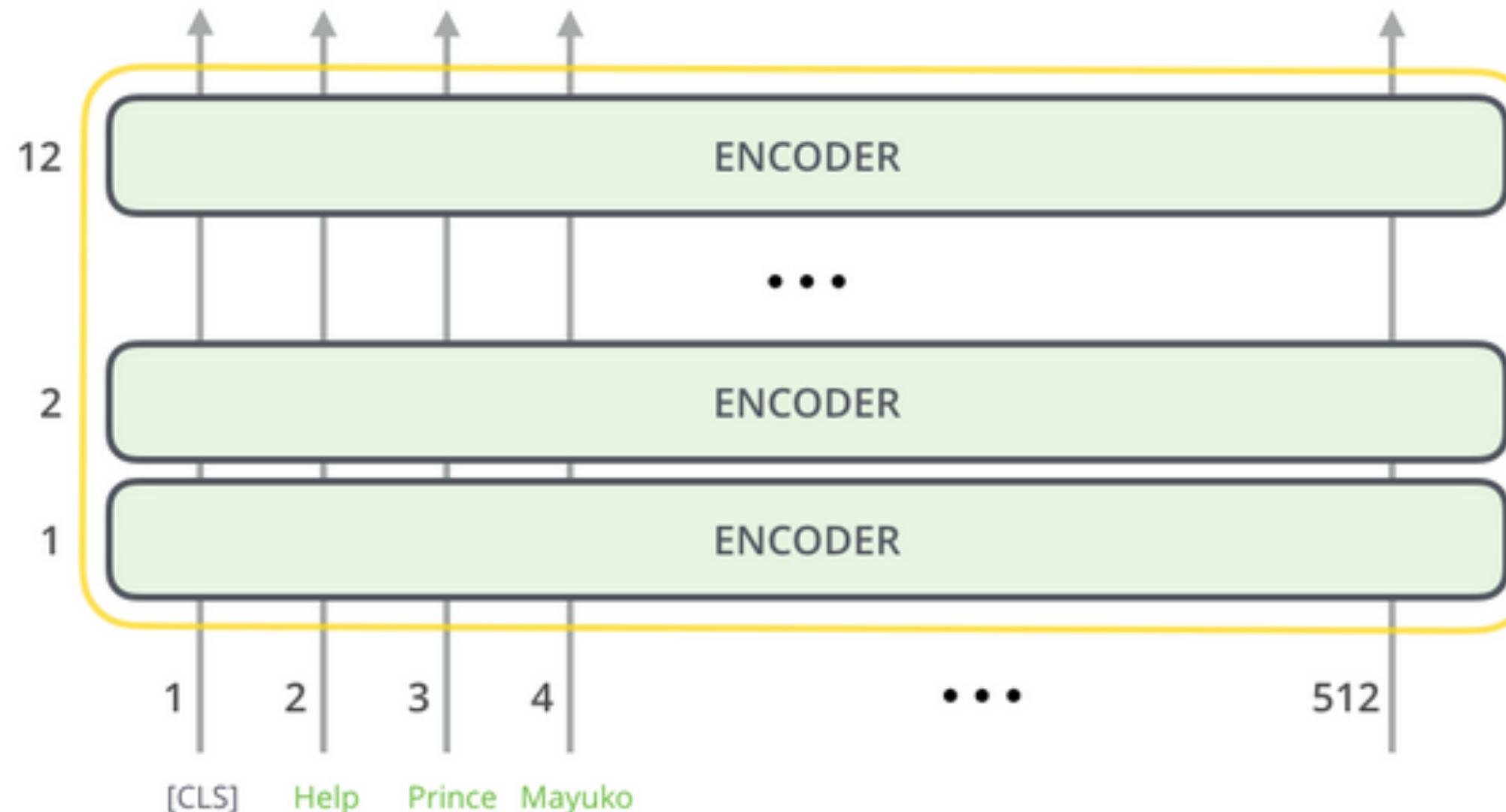
prerequisites

- 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

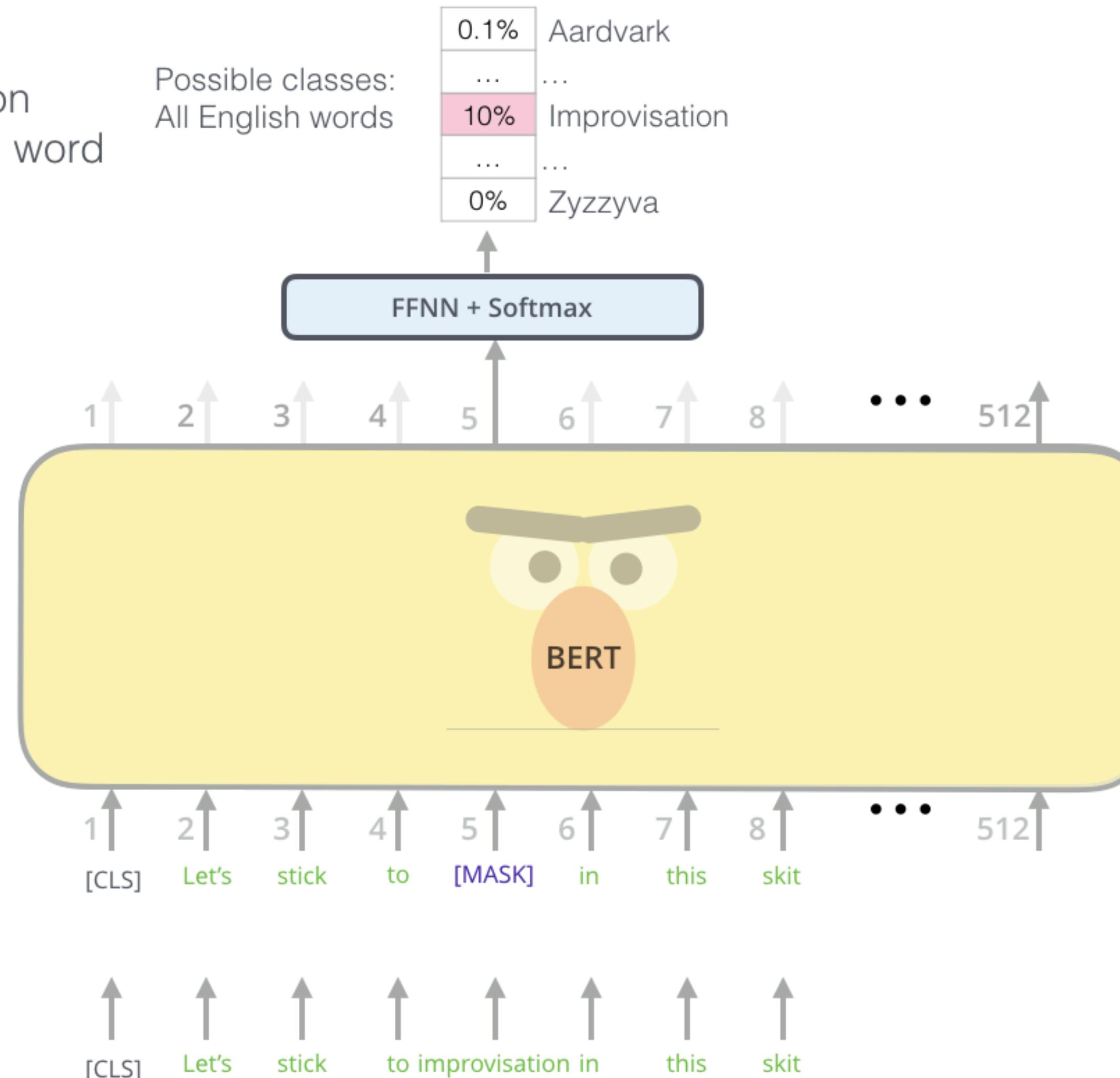
Use the output of the masked word's position to predict the masked word

Possible classes:
All English words

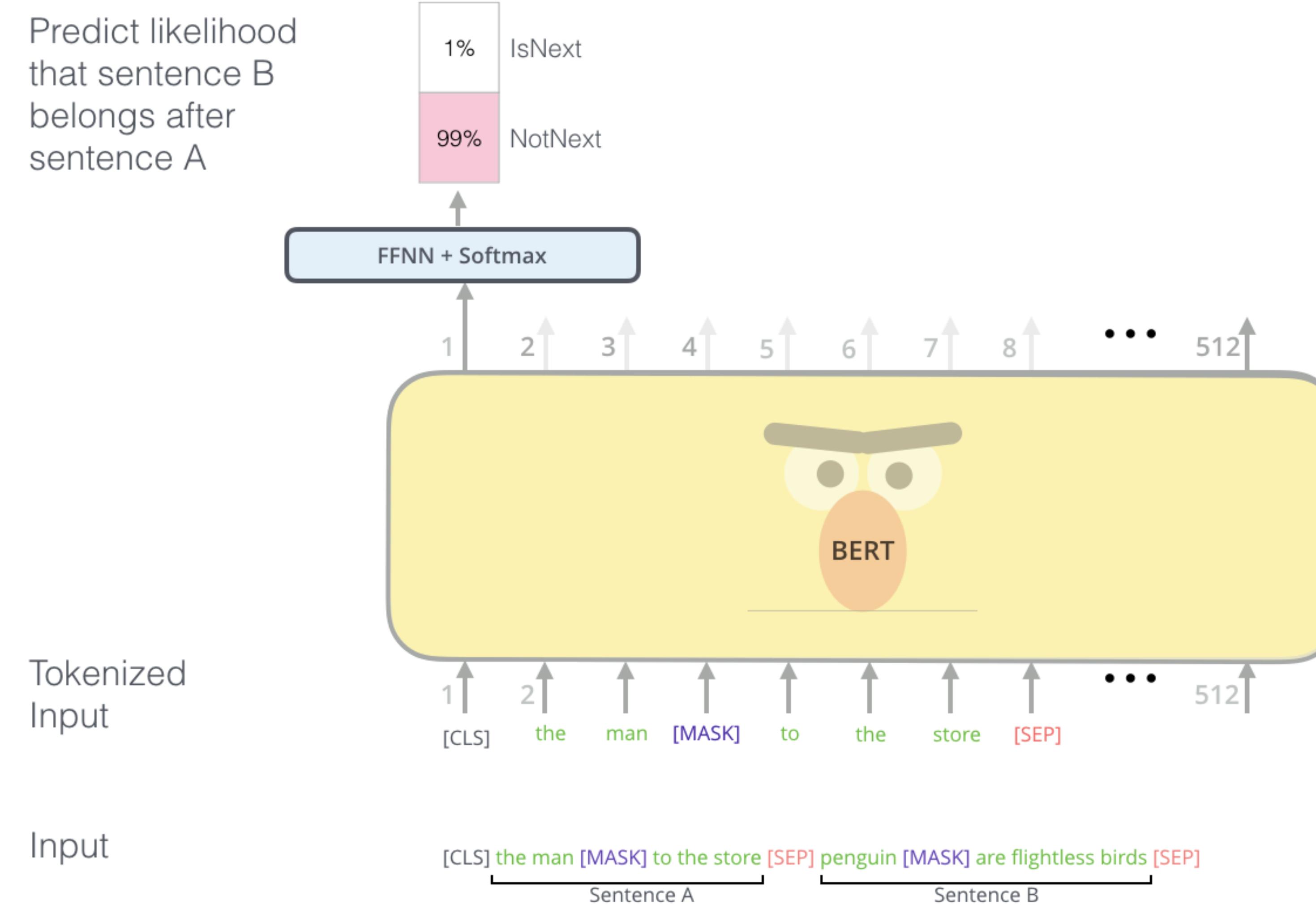
0.1% Aardvark
...
10% Improvisation
...
0% Zzyzyva

Randomly mask 15% of tokens

Input

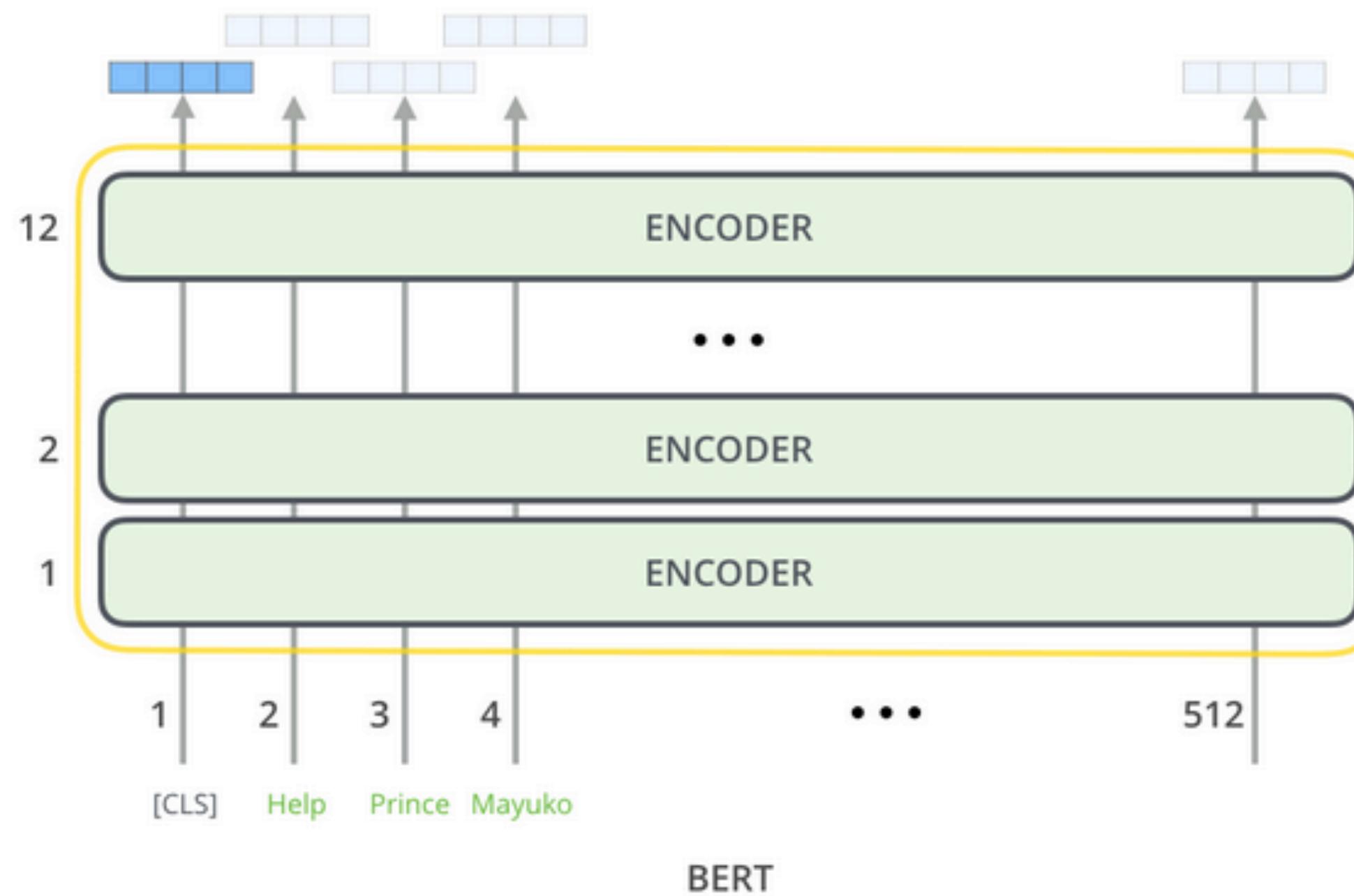


BERT- two sentence task



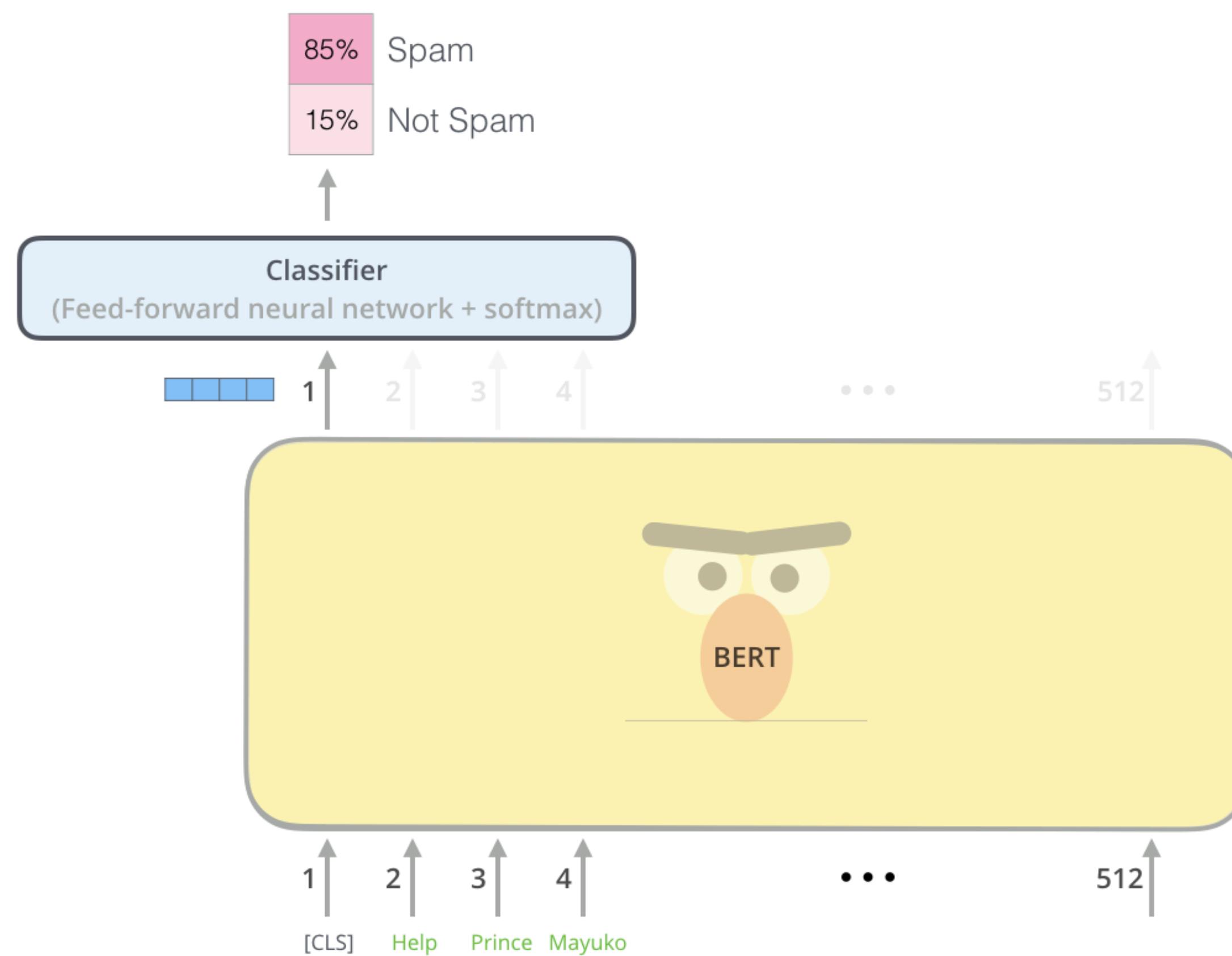
BERT – Model architecture

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



BERT – Model architecture

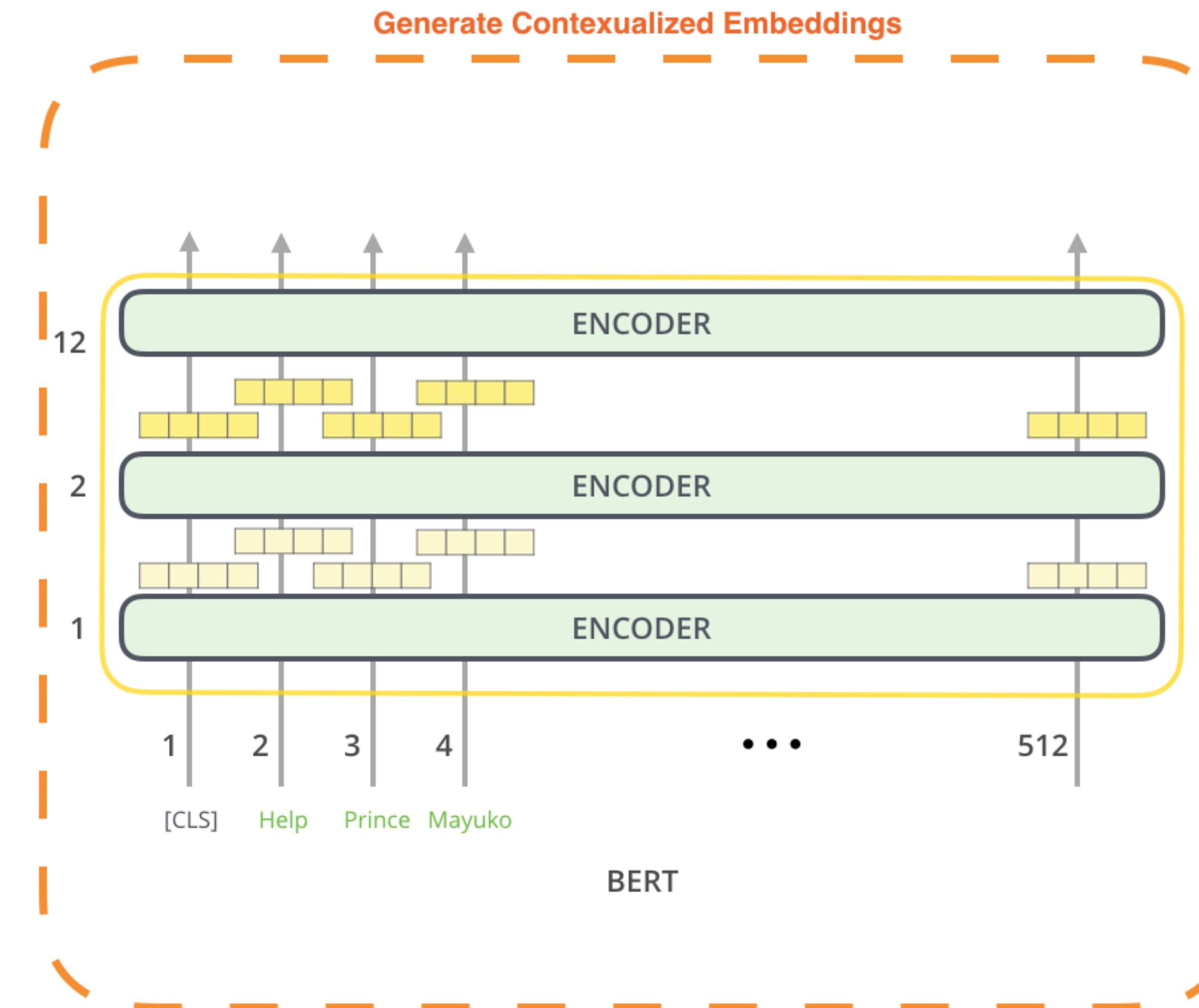
- Vectors can now be used for downstream tasks, e.g. SPAM classification



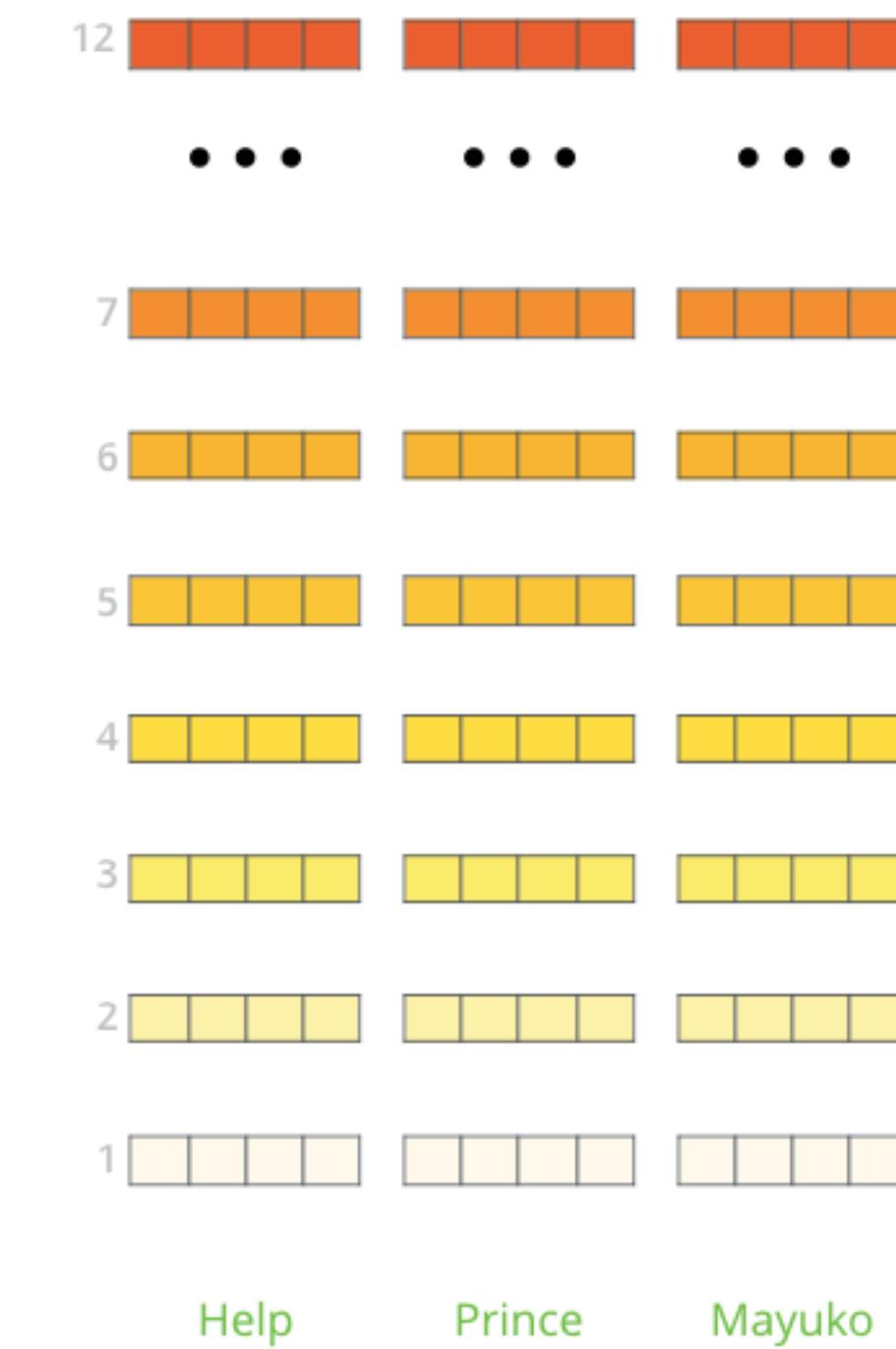
BERT – Model architecture

- 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 Extractor



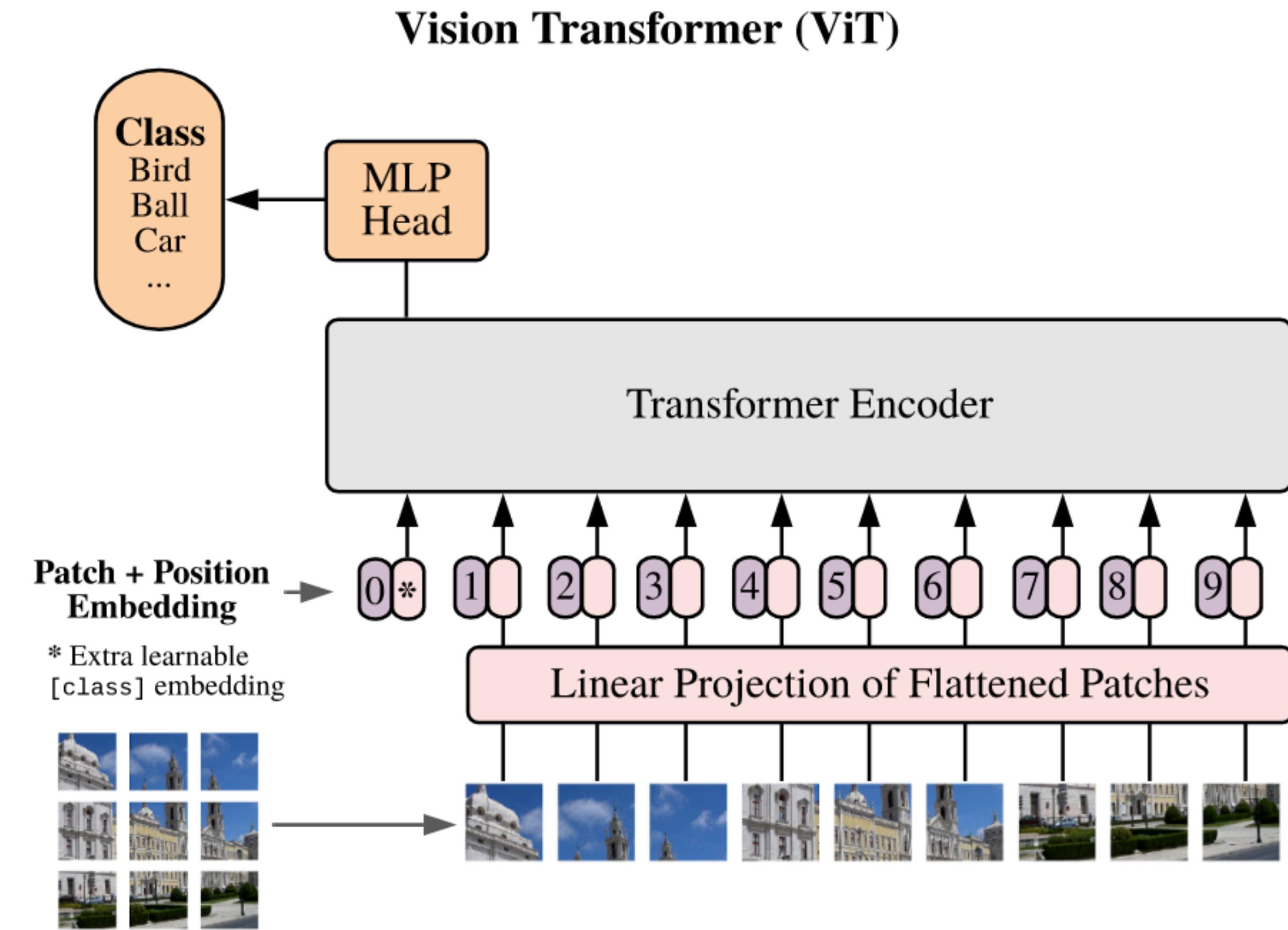
The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

Vision Transformer

- Why not think of images as sequences?
- images as a sequence of patches (16x16)
- self supervised pretraining
- gap to large scale supervised pretraining
- other task than image recognition?

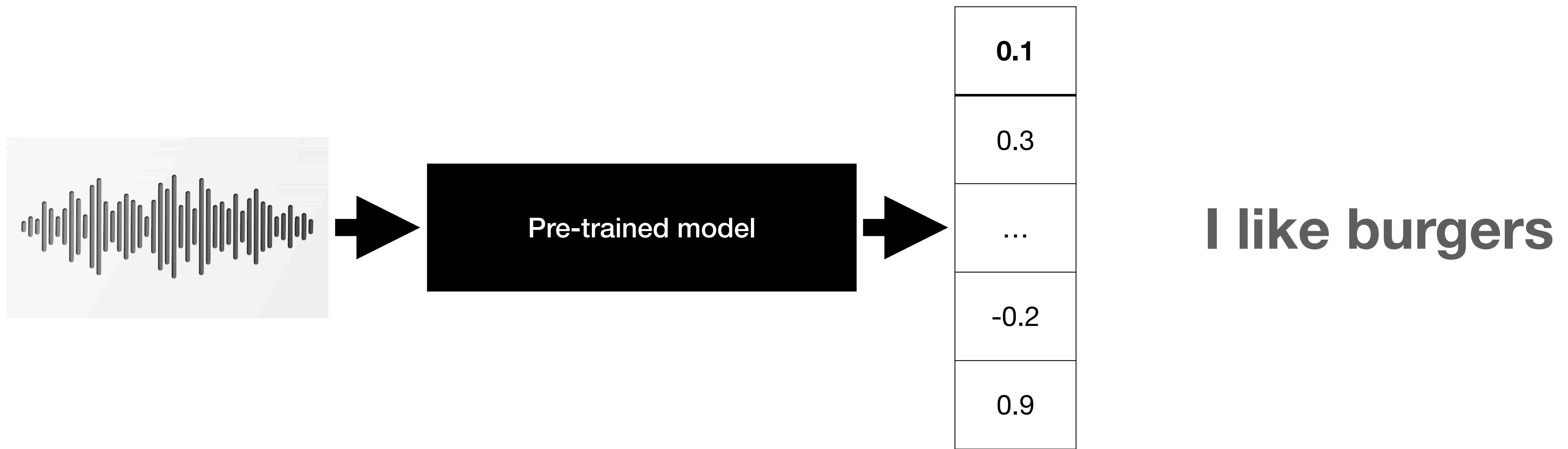


What about audio?

How to learn good representations of
audio data?

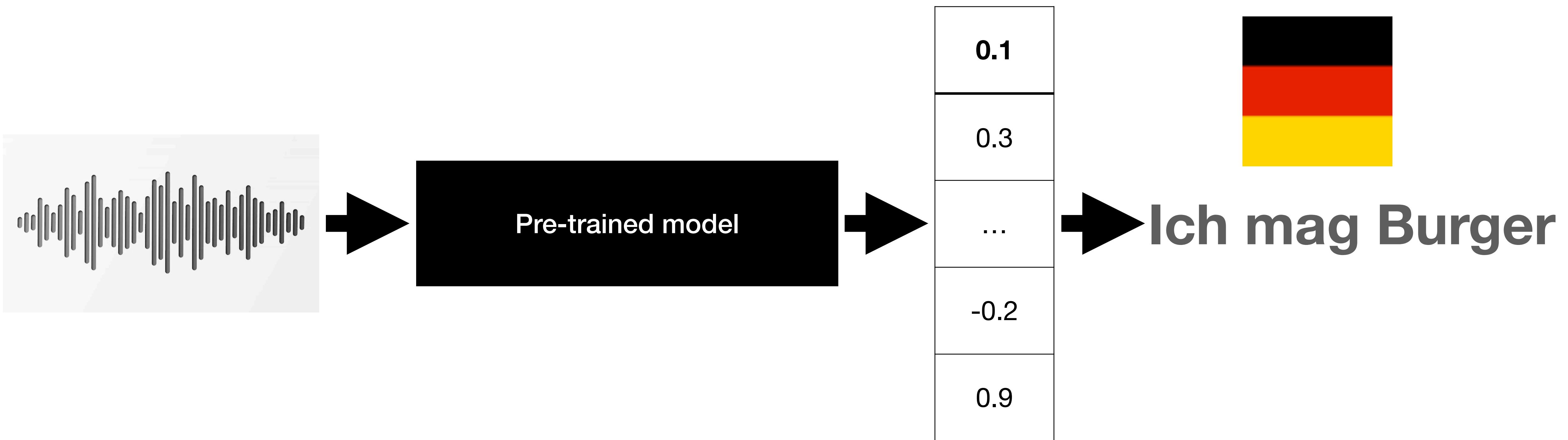
What would be a good representation?

Speech recognition



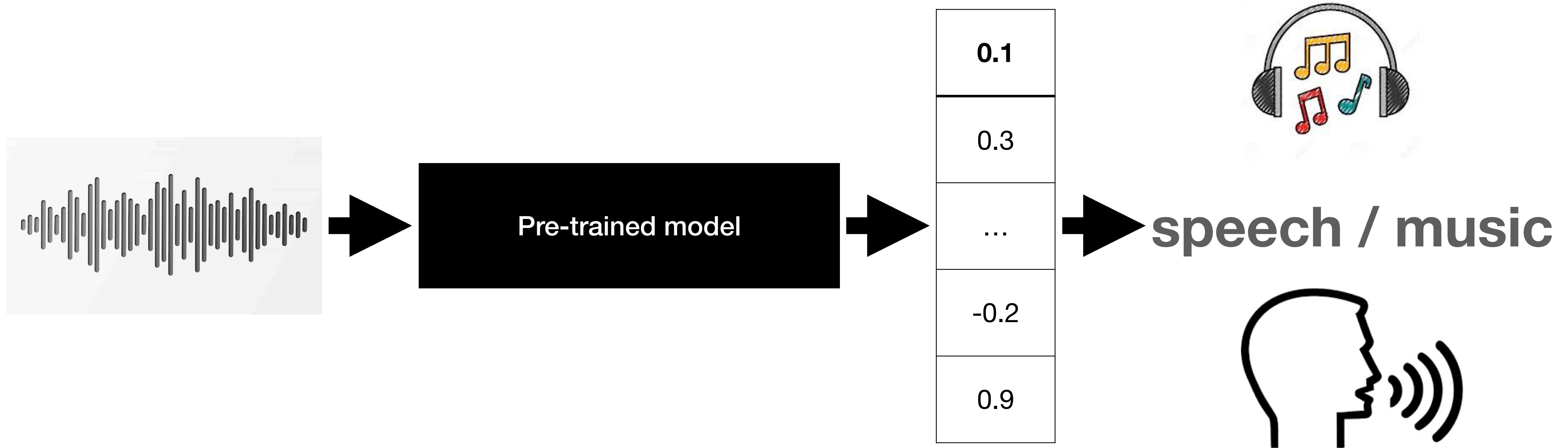
What would be a good representation?

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

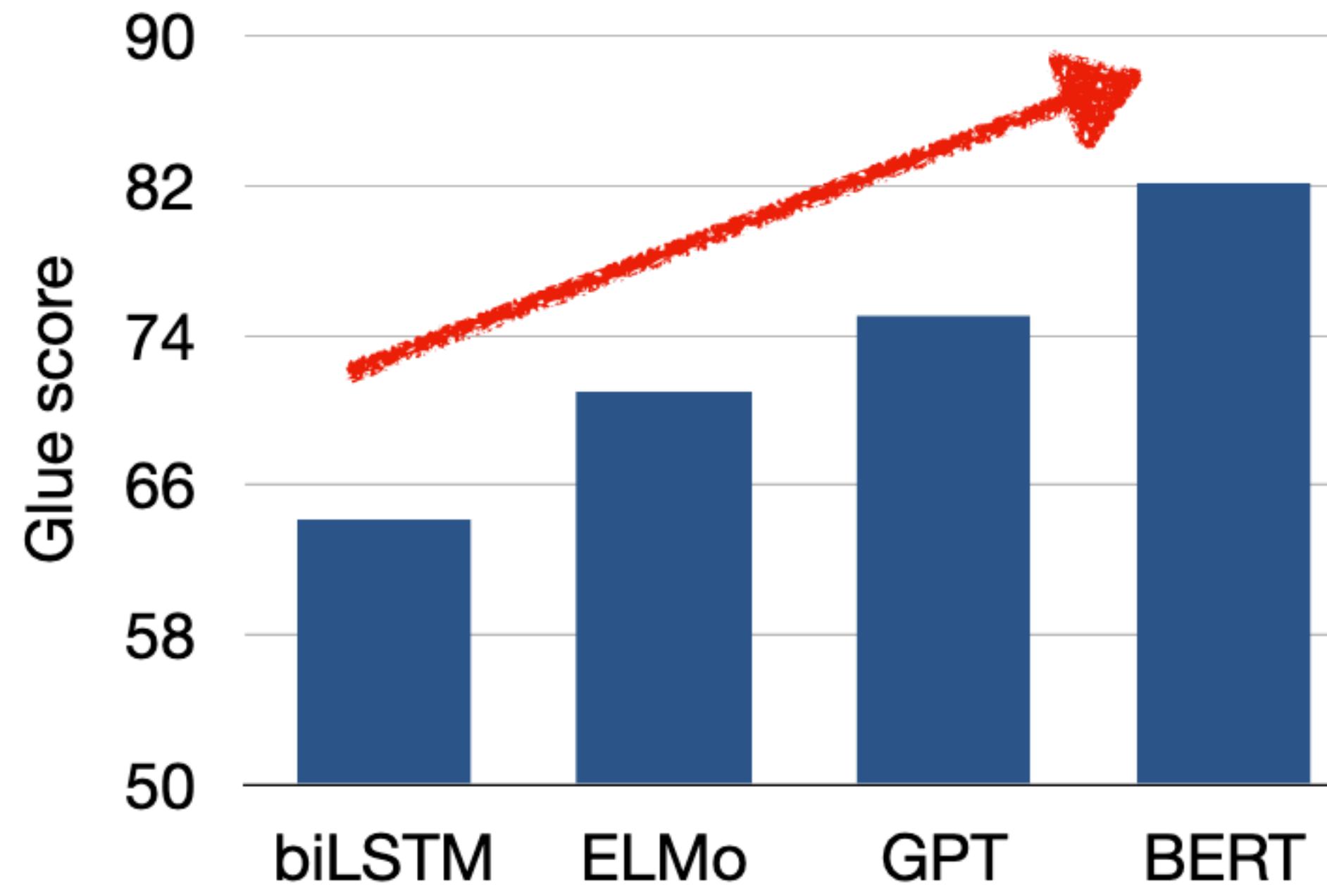


I like burgers

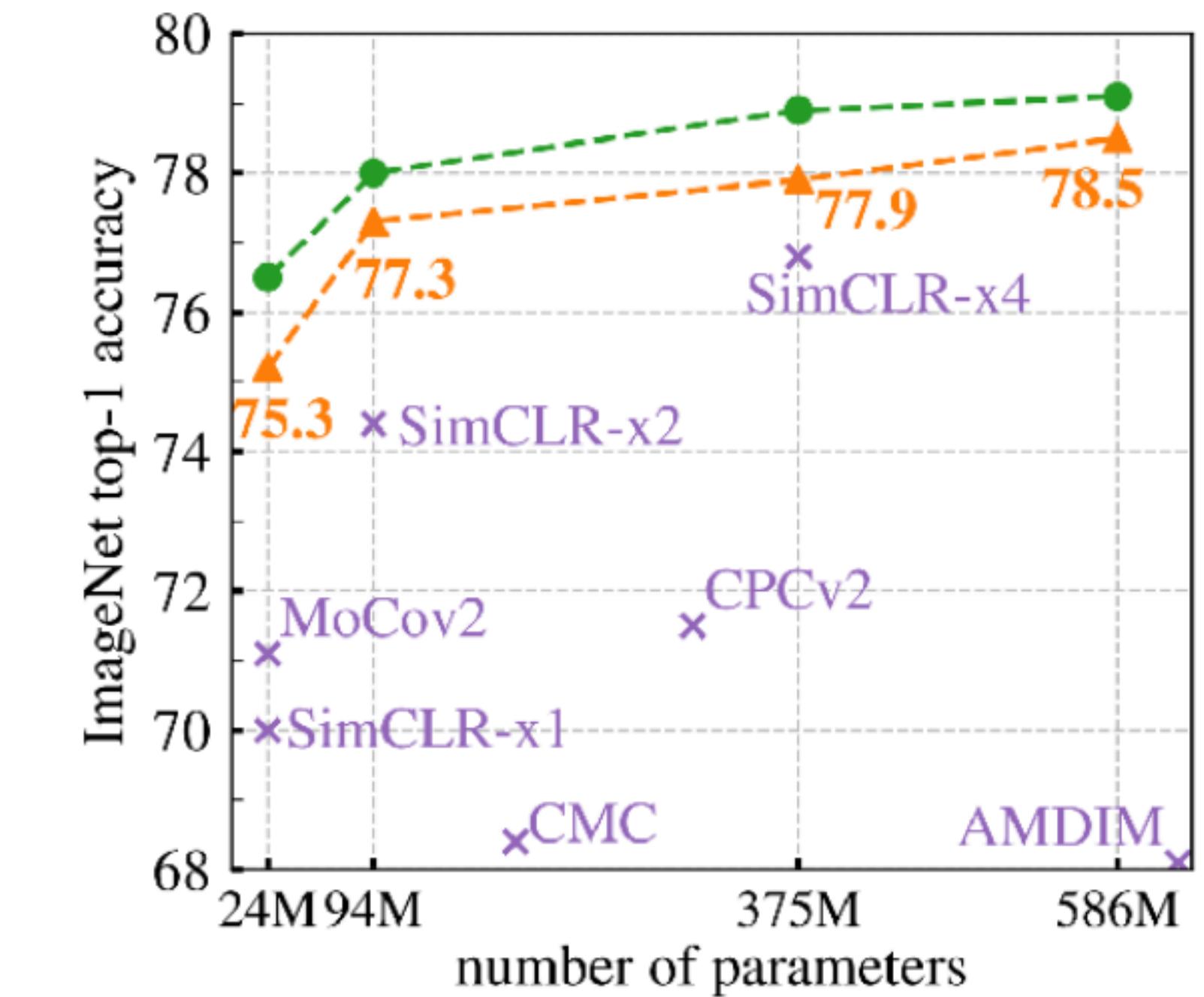
- Training on 1000s of hours of data for good systems
- Many languages, dialects, domains etc.

Pretraining In other fields...

Pre-training in NLP



Pre-training in Computer Vision



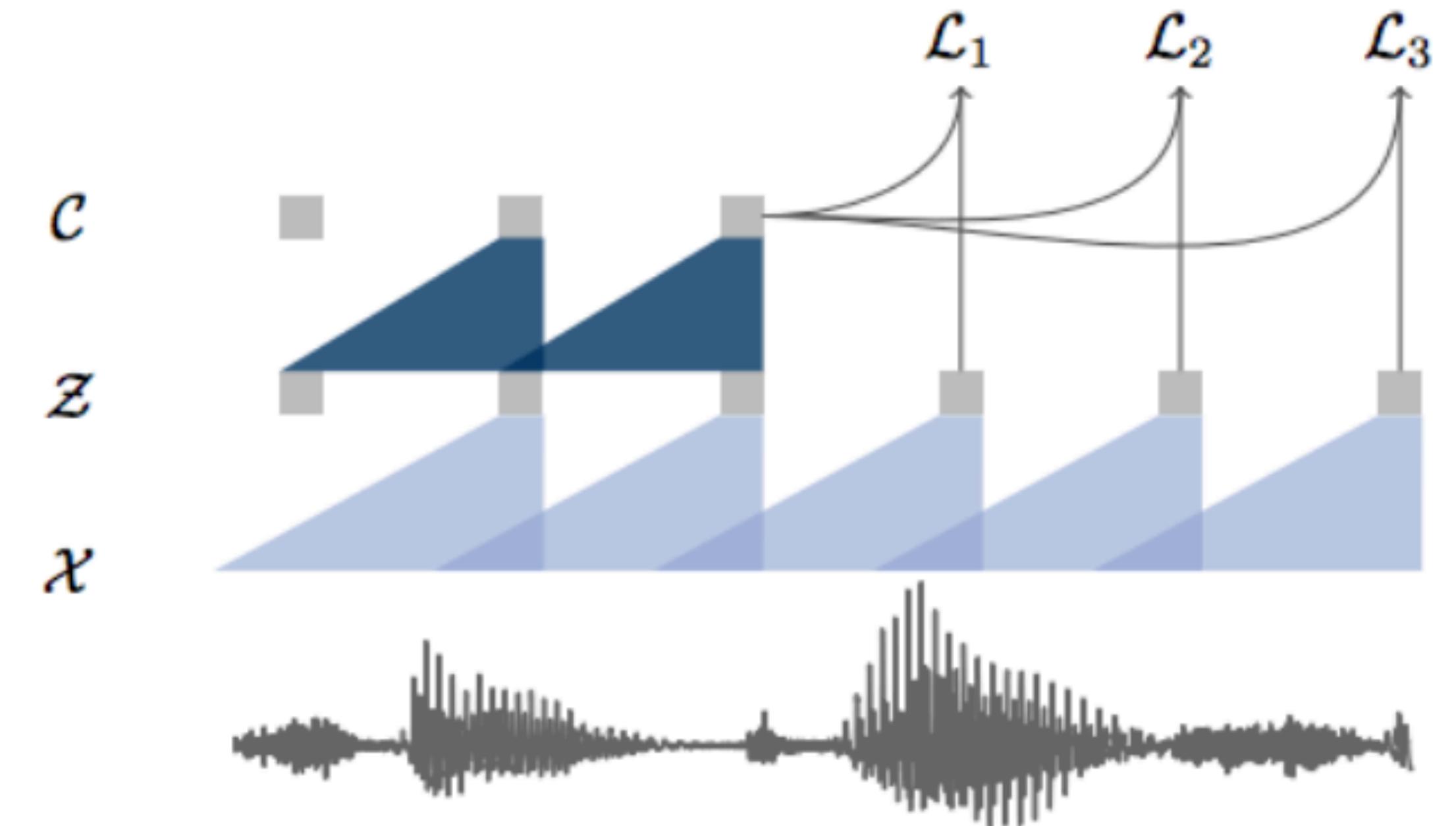
wav2vec 2.0

Learning good representations of
audio data
from unlabeled audio!

wav2vec 2.0

Gradual development wav2vec

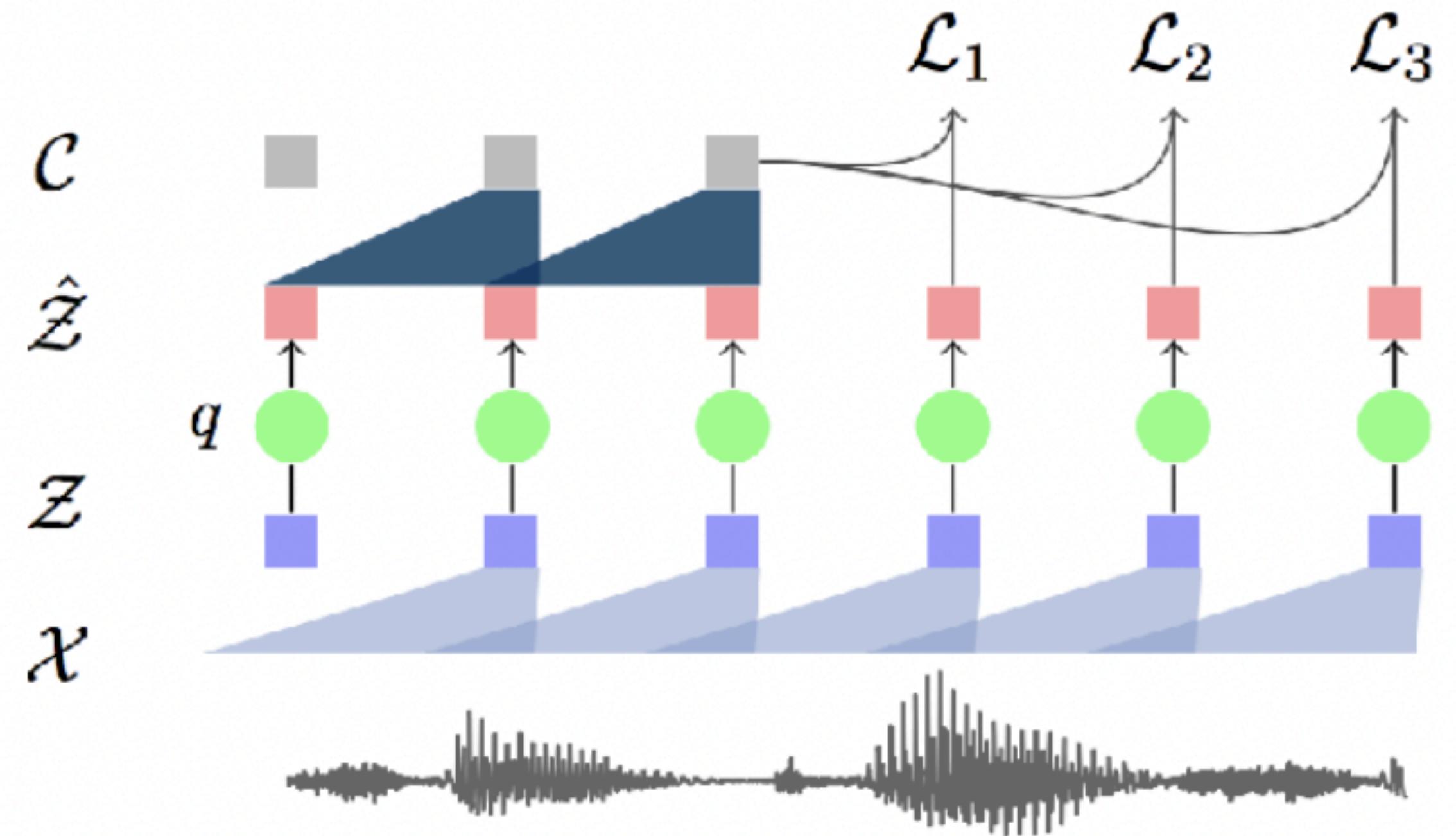
- fully convolutional
- binary cross entropy loss
- representations for ASR tasks



wav2vec 2.0

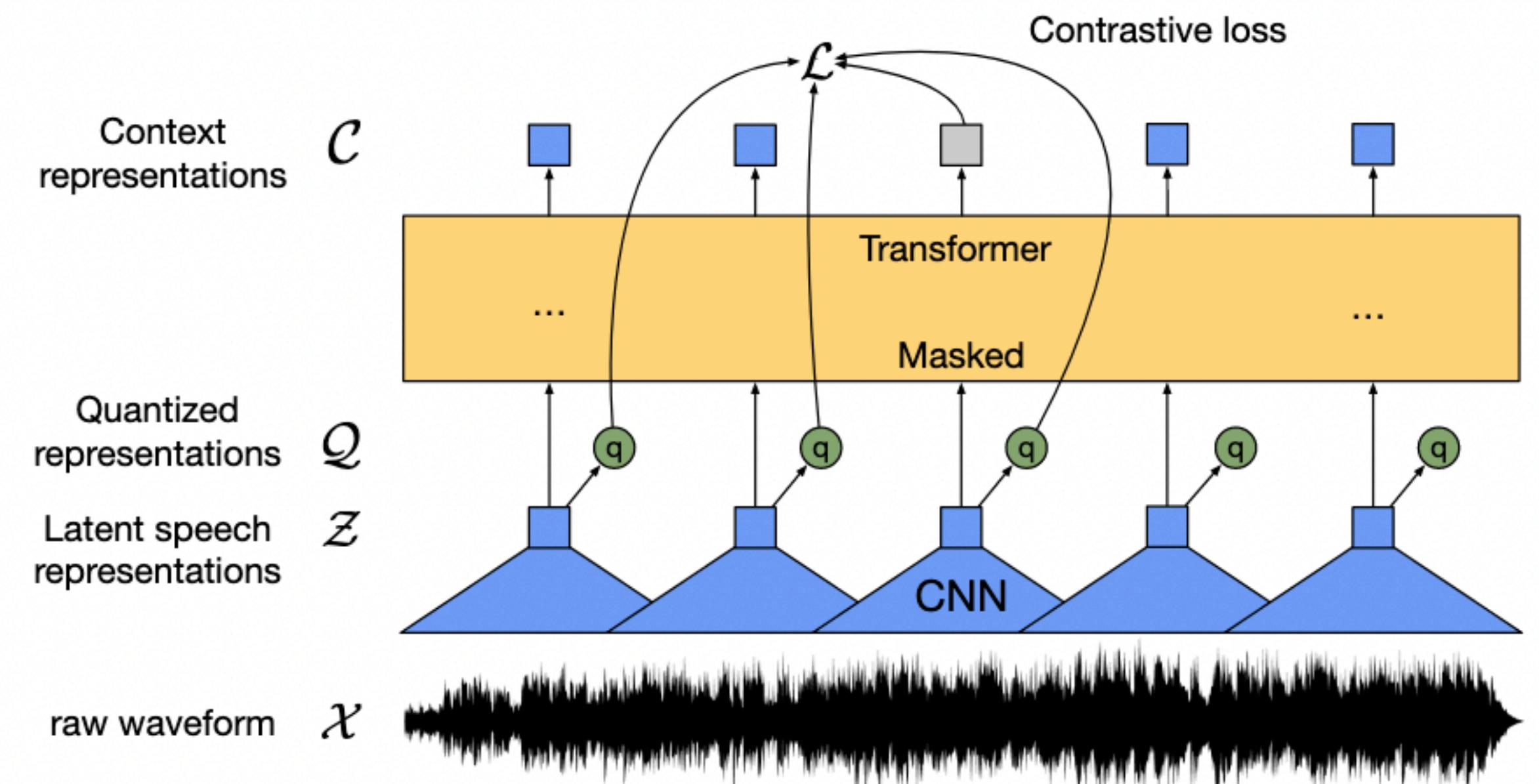
gradual step: vq-wav2vec

- 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

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

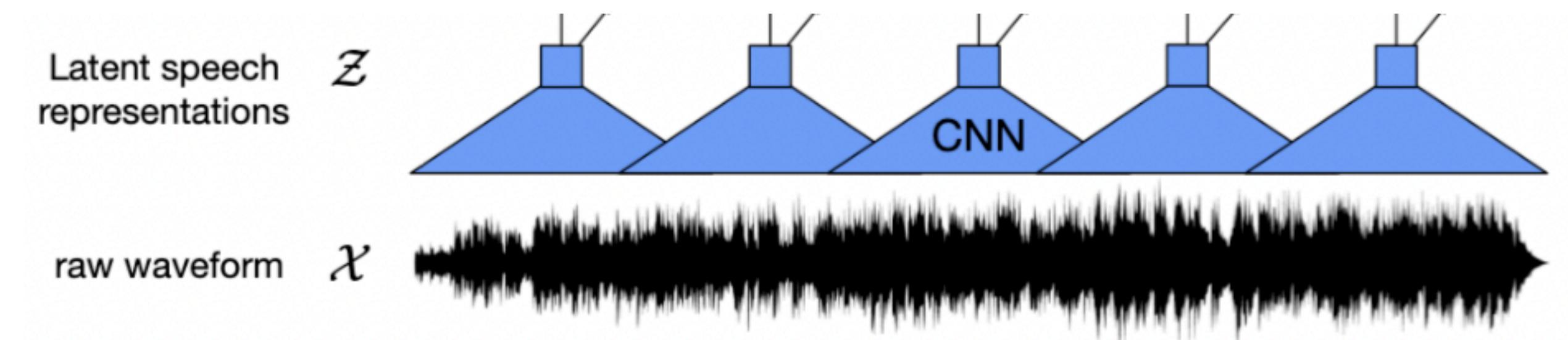
Diagram illustrating the components of the training objective:

- Cosine similarity**: Points to the term $sim(\mathbf{c}_t, \mathbf{q}_t)$.
- Context representation**: Points to the variable \mathbf{c}_t .
- Discrete latent speech representation**: Points to the variable \mathbf{q}_t .
- Negative samples**: Points to the term $\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t}$.
- Temperature**: Points to the term κ .

- Code book penalty to encourage more codes to be used

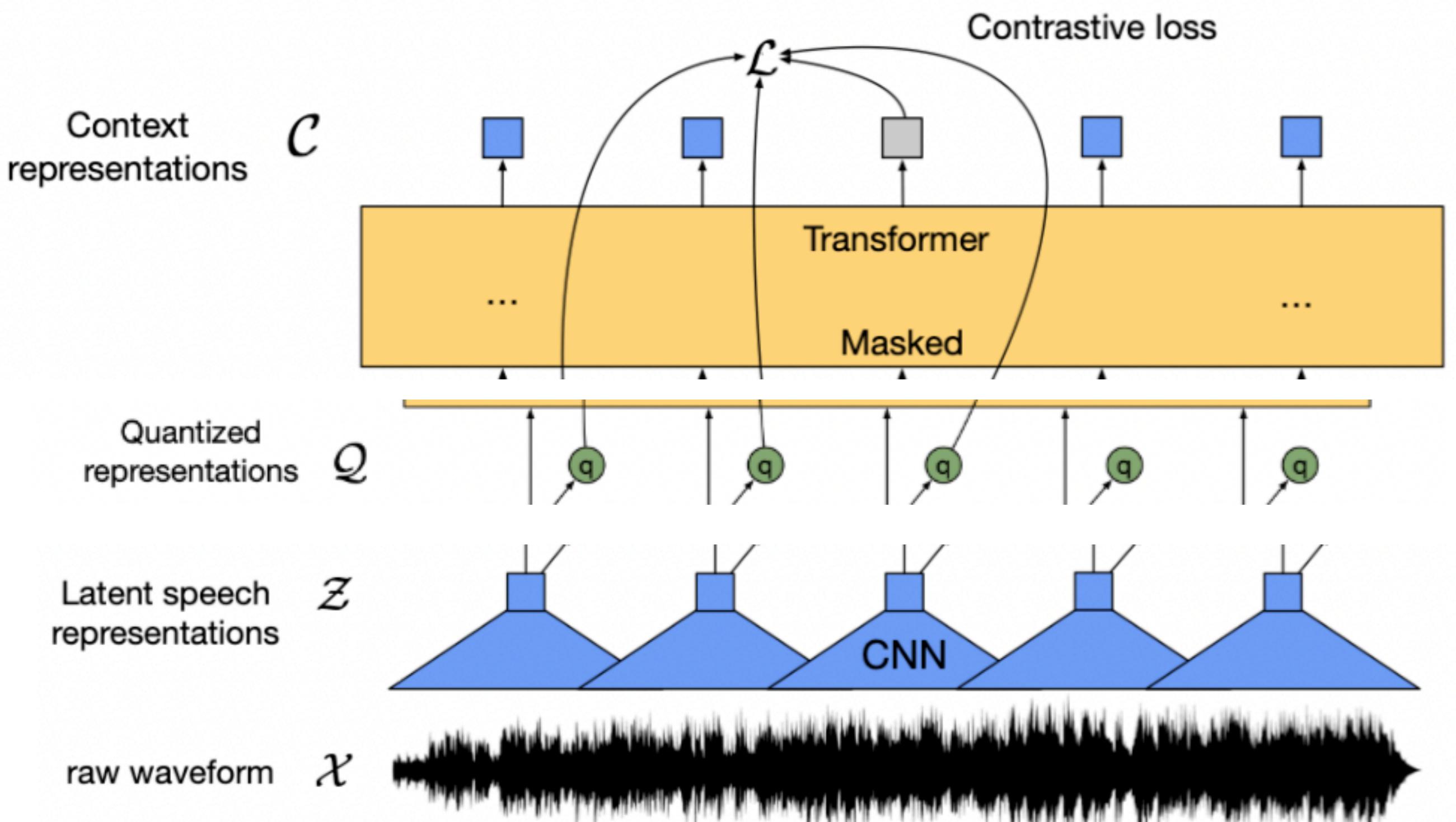
wav2vec 2.0

- input raw waveform
- multi-layer convolutional feature encoder



wav2vec 2.0

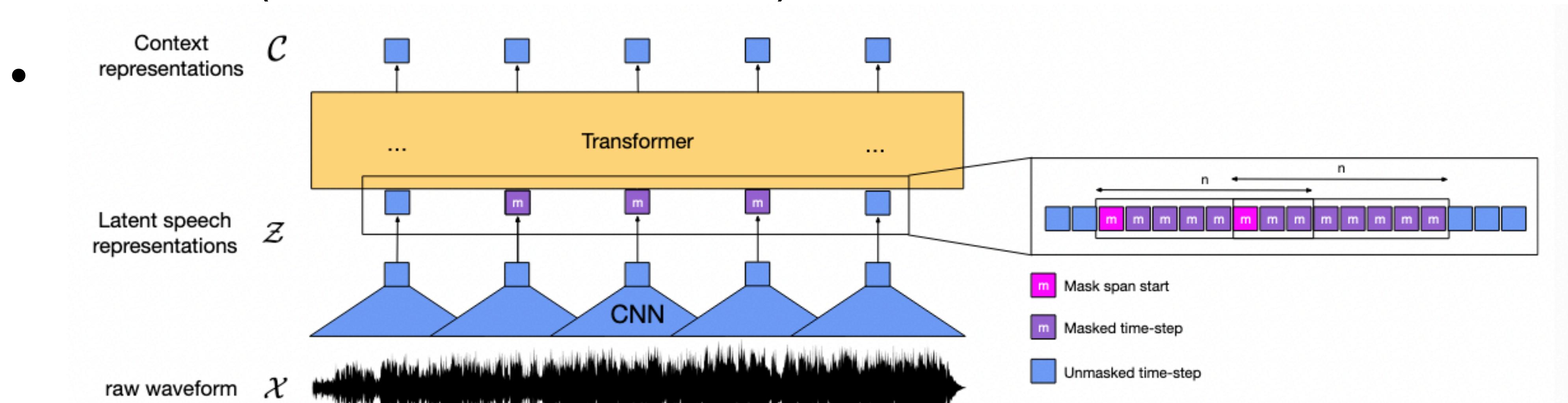
- input raw waveform
- multi-layer convolutional feature encoder
- quantization
- masking of quantized sequence (similar to MLM)
- Transformer Network to build contextualized representations
- contrastive loss —> true latent vs distractors



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

Transformers

- We saw that Transformers work well on...
 - Text
 - Audio
 - Images
- Modality-specific targets
 - Speech: human speech units
 - Natural Language Processing: words, subwords
 - Computer Vision: visual tokens, pixels
- This is all very different w.r.t. the units, training therefore works quite different
- Could we unify the way we look at the data?

data2vec

- Idea: Transformers work basically for everything, why not unify the training procedure?
- Same self-supervised learning method for different modalities
 - Speech
 - Text
 - Images
- Contextualized representations
 - Self-distillation setup using Transformers
 - Predicting latent representations of full input (teacher) instead of predicting modality-specific targets

data2vec

- Modality-specific targets
 - Speech: human speech units
 - Natural Language Processing: words, subwords
 - Computer Vision: visual tokens, pixels
- State of the art/competitive on benchmarks
 - Speech: Librispeech (test-other)
 - Natural Language Processing: GLUE
 - Image Classification: ImageNet-1K
- **data2vec is not a new model architecture** or multimodal pretraining, it focuses on **a unified self-supervised learning strategy**

data2vec

Image	Audio	NLP
<p>Vision Transformer (ViT)</p> <p>Class Bird Ball Car ...</p> <p>MLP Head</p> <p>Transformer Encoder</p> <p>Patch + Position Embedding * Extra learnable [class] embedding</p> <p>Linear Projection of Flattened Patches</p>	<p>Context representations C</p> <p>Quantized representations Q</p> <p>Latent speech representations Z</p> <p>raw waveform x</p> <p>Transformer</p> <p>Masked</p> <p>Contrastive loss L^*</p>	
<ul style="list-style-type: none"> • BEiT: Multiple adjacent patches • Block ≥ 16 patches • Mask 60% of the patches 	<ul style="list-style-type: none"> • Wav2vec2.0: 16 kHz waveform input • Encoder: temporal convolutions • Sample $p=0.065$ of all timesteps as starting indices • Mask subsequent 10 timesteps <ul style="list-style-type: none"> • ~49% of timesteps masked for typical training sequence 	<ul style="list-style-type: none"> • Masked LM Task without Next Sentence Prediction (NSP) • Mask 15% of tokens (80% mask token, 10% unchanged, 10% replaced random token)

data2vec

- Do not predict this:
 - Speech: human speech units
 - Natural Language Processing: words, subwords
 - Computer Vision: visual tokens, pixels
- Rather predict representations
- Different inputs, but same prediction target, independent of modality
- —> **data2vec is not a new model architecture or multimodal pretraining, it focuses on a unified self-supervised learning strategy**

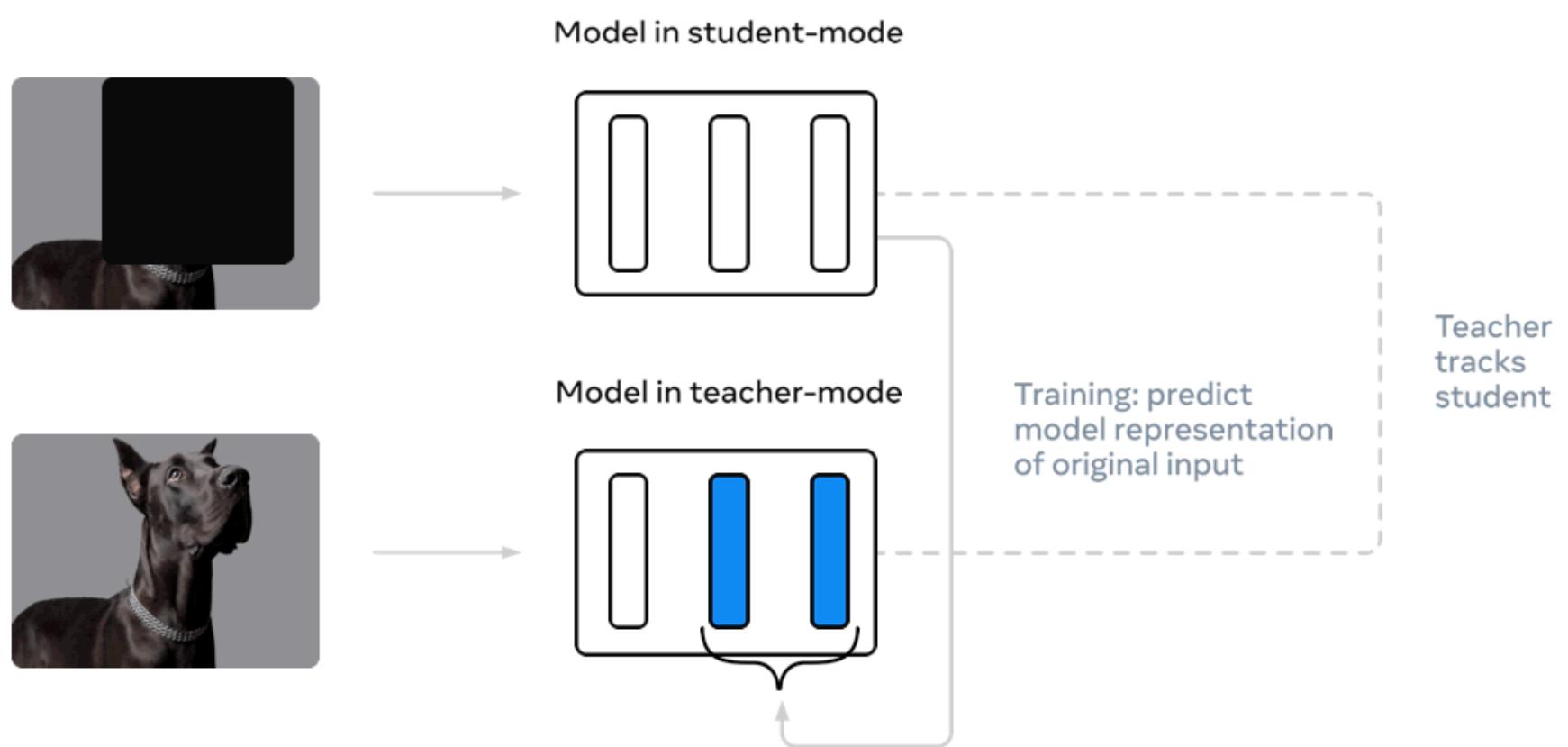
data2vec

training objective

- Smoothed L1 Loss

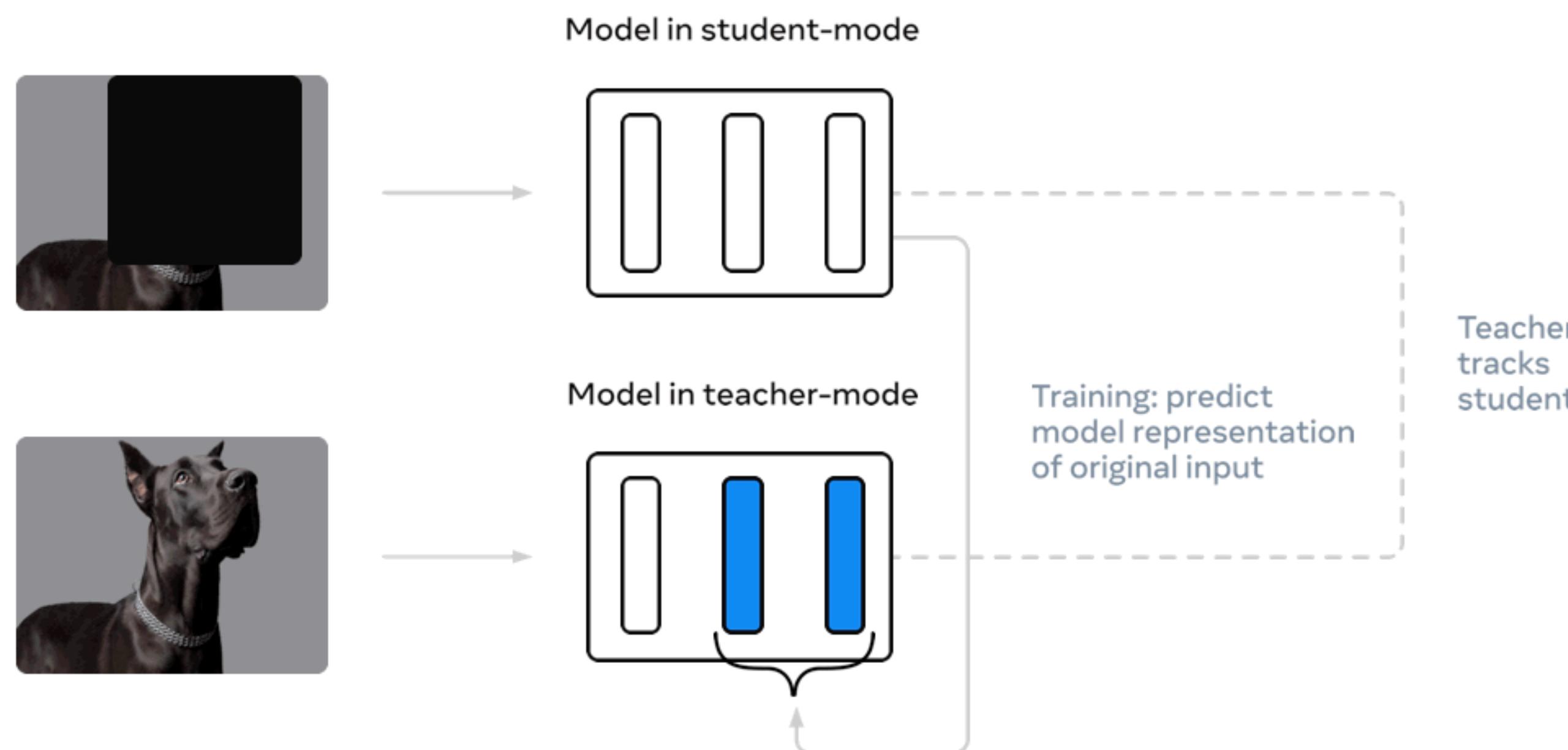
$$\mathcal{L}(y_t, f_t(x)) = \begin{cases} \frac{1}{2}(y_t - f_t(x))^2 / \beta & |y_t - f_t(x)| \leq \beta \\ (|y_t - f_t(x)| - \frac{1}{2}\beta) & \text{otherwise} \end{cases}$$

Our method for images



data2vec

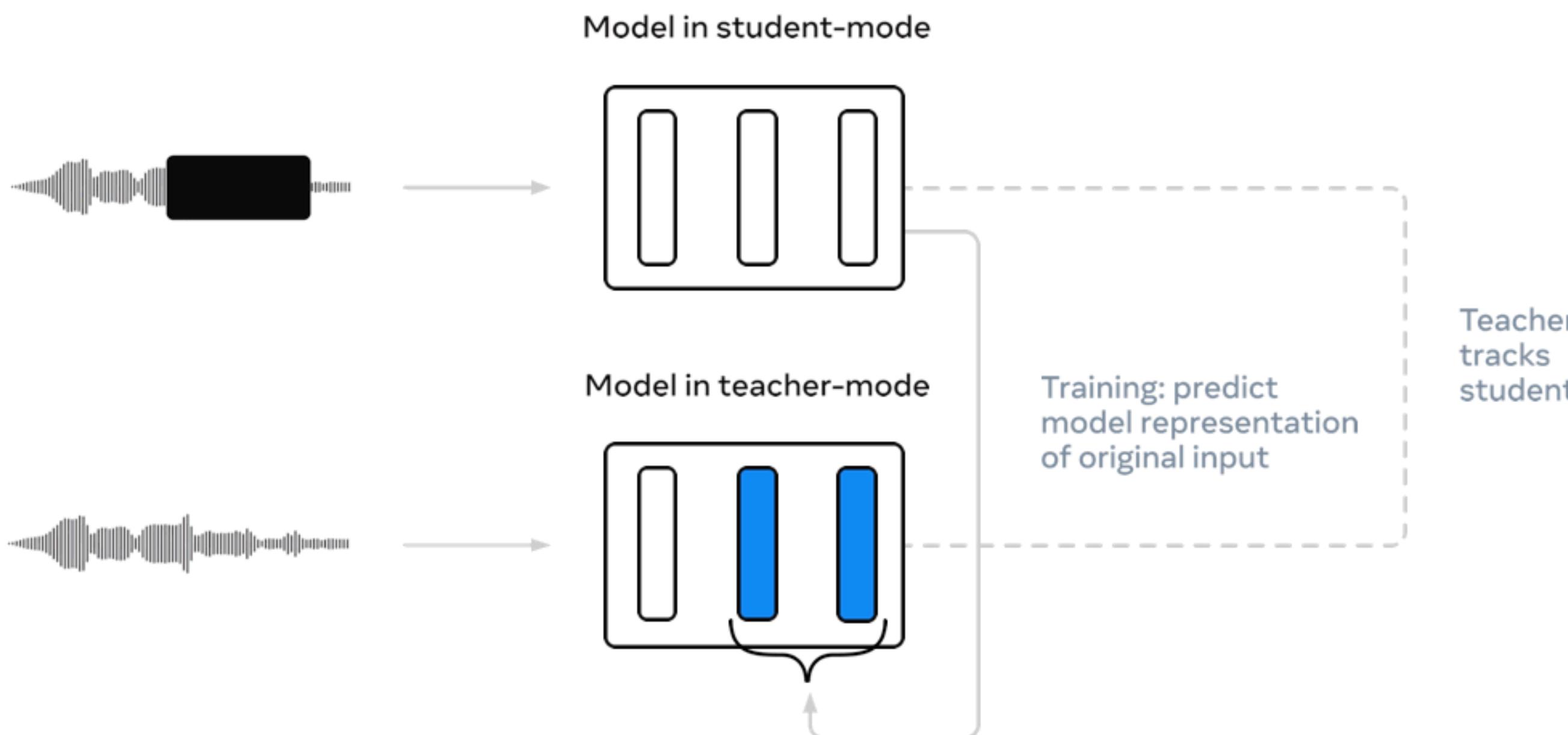
Our method for images



•

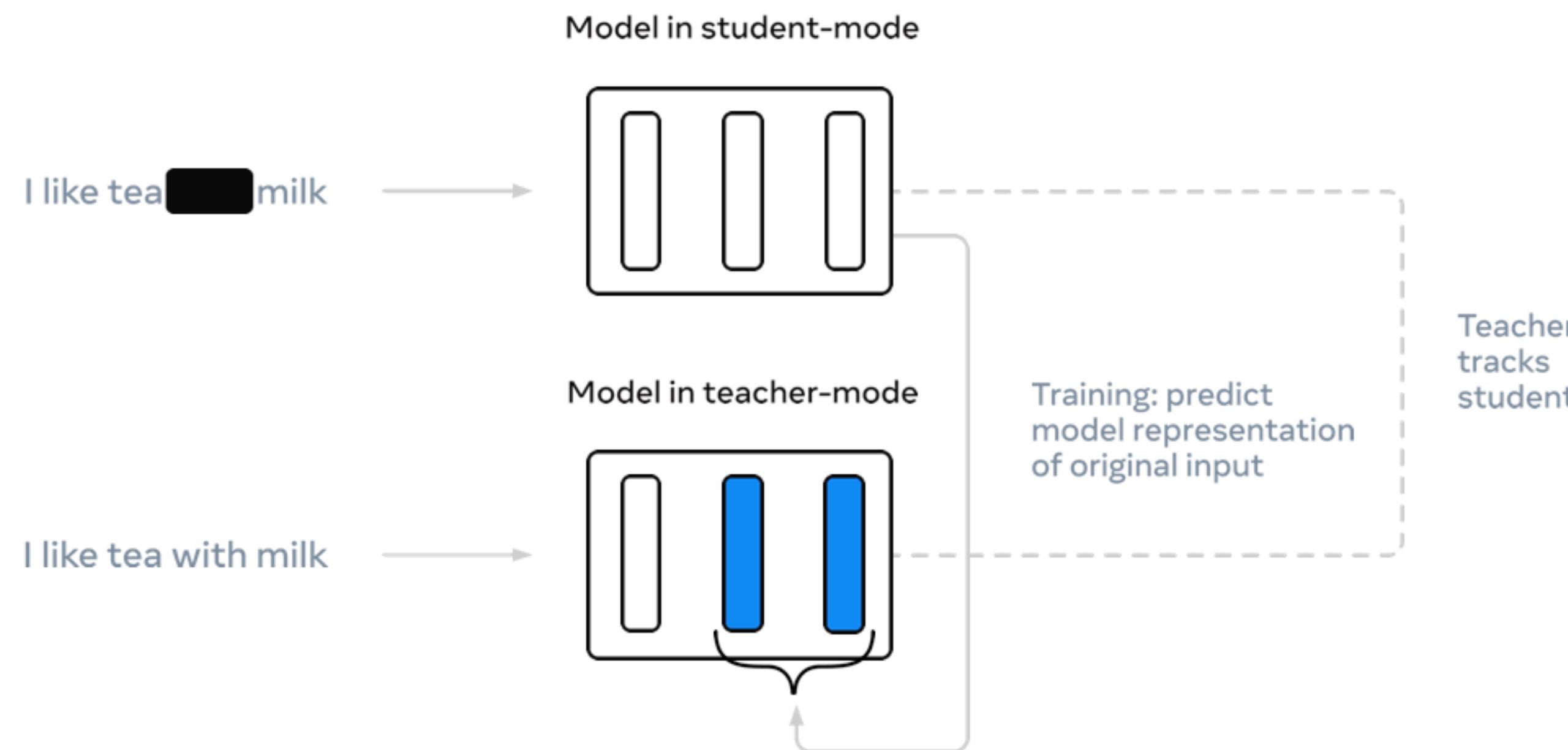
data2vec

Our method for speech



data2vec

Our method for text



•

data2vec

pre-training teacher

Image	Audio	NLP
<p>Vision Transformer (ViT)</p> <p>Class Bird Ball Car ...</p> <p>MLP Head</p> <p>Transformer Encoder</p> <p>Patch + Position Embedding * Extra learnable [class] embedding</p> <p>Linear Projection of Flattened Patches</p>	<p>Context representations C</p> <p>Quantized representations Q</p> <p>Latent speech representations Z</p> <p>raw waveform X</p> <p>Transformer</p> <p>Masked</p> <p>Contrastive loss L^*</p>	
<ul style="list-style-type: none"> Pretrained 800 epochs Batchsize 2048 (8192) Cosine schedule/warmup Finetuning + softmax-normalized projection 	<ul style="list-style-type: none"> Pretrained 400K updates Batchsize 63mins (61M frames) Tri-stage scheduler/warmup Finetuning regime of wav2vec2.0 	<ul style="list-style-type: none"> Pretrained 1M updates Batchsize 256 (512 tokens) Tri-stage scheduler/warmup Finetuned w.r.t. to General Language Understanding Evaluation (GLUE) benchmark

data2vec

- Modality-specific feature extractors & masking
- Unified learning regime across modalities
- Modality-specific feature extractors & masking strategies
- Future work: self-supervised + Perceiver (operates on raw data)

Sources

- <http://michaelauli.github.io/talks/wav2vec-ssl.pdf>
- <https://jalammar.github.io/illustrated-transformer/>
- <https://ai.facebook.com/blog/the-first-high-performance-self-supervised-algorithm-that-works-for-speech-vision-and-text/>
- 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
- data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language
Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, Michael Auli