Современное распознавание речи

21.11.2024

Не забудьте отметиться и оставить отзыв!

О чем поговорим (в рамках 2ух занятий)

- Введение в ЦОС, как представлен звук в памяти компьютера
- Распознавание речи ASR (какую задачу решаем, как оцениваем качество решения)
- Архитектуры моделей для решения задачи ASR (рассмотрим основные подходы, подробнее поговорим про SOTA решения)
- SSL в задаче ASR
- Применение LM для улучшения качества распознавания

Чего не будет на занятиях:

- TTS (text-to-speech)
- KWS (keyword spotting)
- VQE (voice quality enhancement)

План занятия

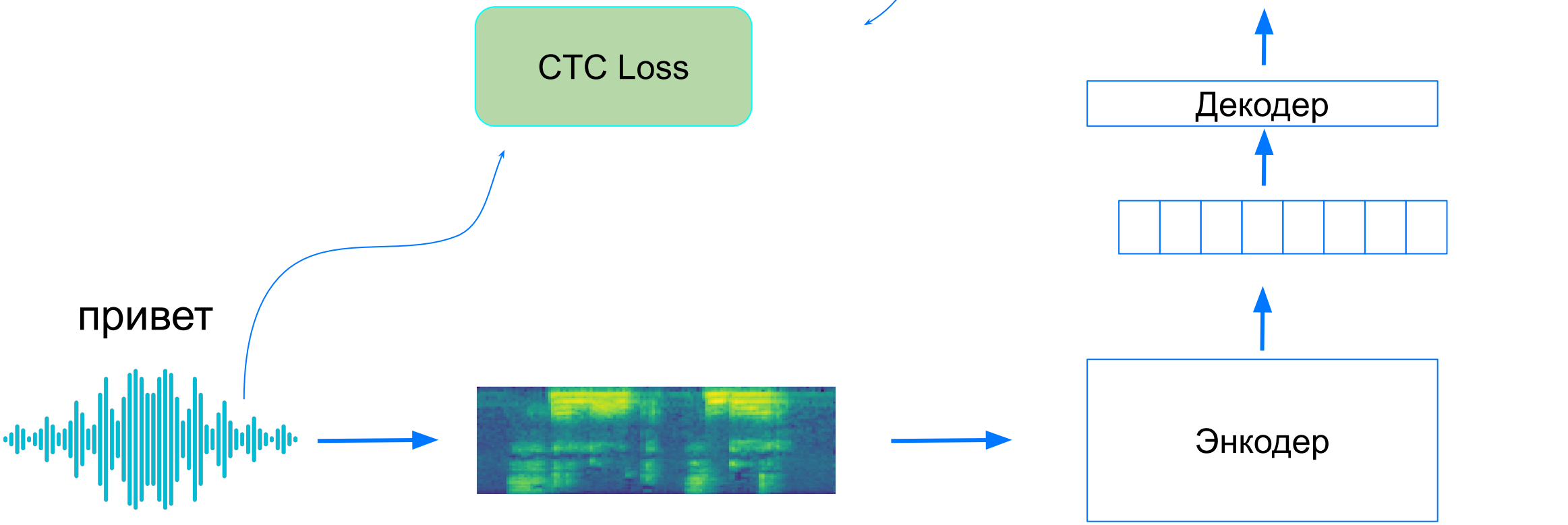
- Вспоминаем CTC
- Seq2seq для задачи ASR
- Listen Attend and Spell
- Whisper
- LM rescoring
- SSL in ASR
- Further reading

Вспомним СТС

CTC

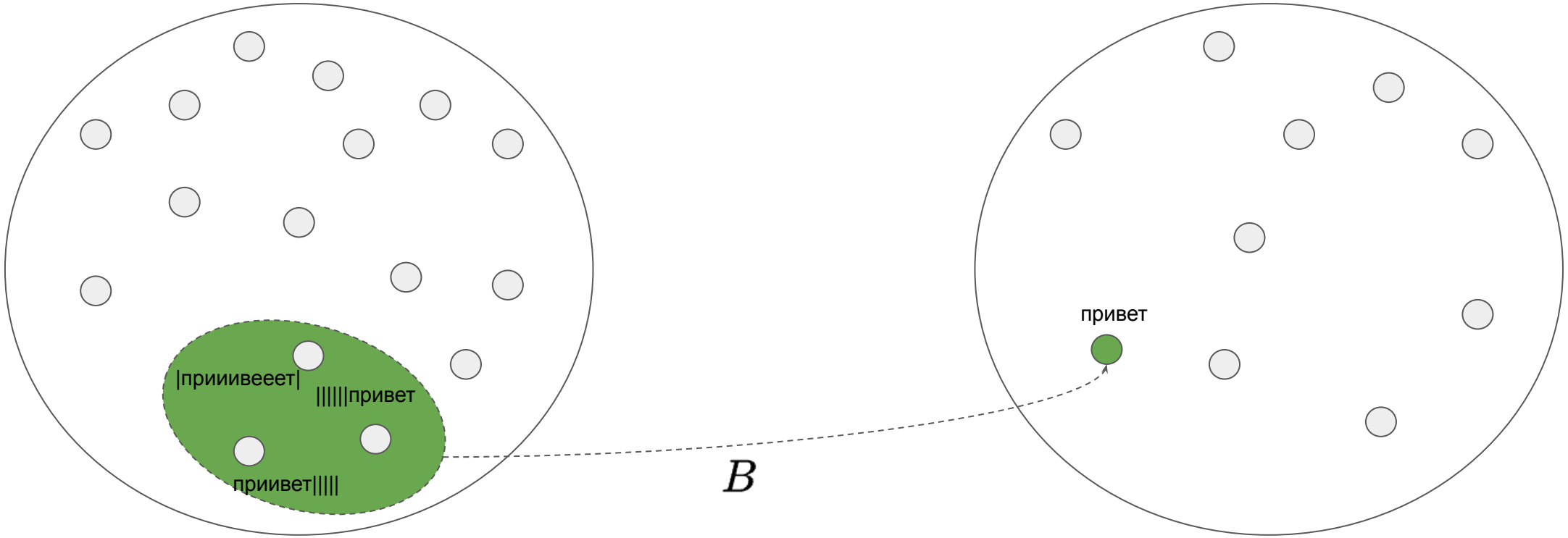
$L = \min -\log(P(Y|X))$

Иными словами - максимизируем
P(гипотезы|спектрограммы)



CTC Loss

$$P(\text{привет}) = P(|\text{прииивееет}|) + P(|||||\text{привет}) + P(\text{приивет}|||||) + \ldots$$



Множество всех возможных путей

Множество всех возможных гипотез

CTC Loss

$$\mathcal{L}(X, R) = -\log \sum_{C \in B^{-1}(R)} P(C|X) = -\log \sum_{C \in B^{-1}(R)} \prod_{t=1}^{T} p(c_t|X)$$

спека

GT гипотеза

гипотезы для которых R = B(C)

$$|V| = 7, seq\ len = 12, possible\ paths = 7^{12} \sim 14B$$

[Как считать CTC лосс эффективно]

CTC Loss summary

- контекстно независимые предсказания
- можем предсказывать параллельно для каждого hidden
- не склонен к бредогенерации
- len(text) > len(encoder hiddens) - никогда не предскажем корректную гипотезу

Вспомним постановку задачи ASR

Постановка задачи ASR

Model

Привет

Sequence
(spec frames)

Sequence
(tokens)

Постановка задачи ASR

Привет

Model

Sequence
(spec frames)

Sequence
(tokens)

NMT

Я есть грут

I am Groot

Постановка задачи ASR

Привет

Model

Sequence
(spec frames)

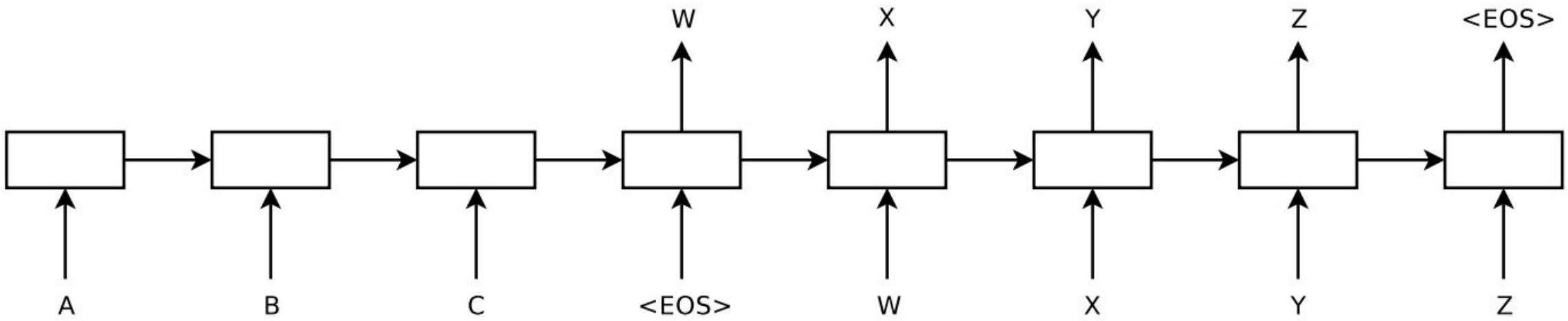
Sequence
(tokens)

NMT

Я есть грут → I am Groot

sequence to sequence modelling

Sequence to sequence learning with NN

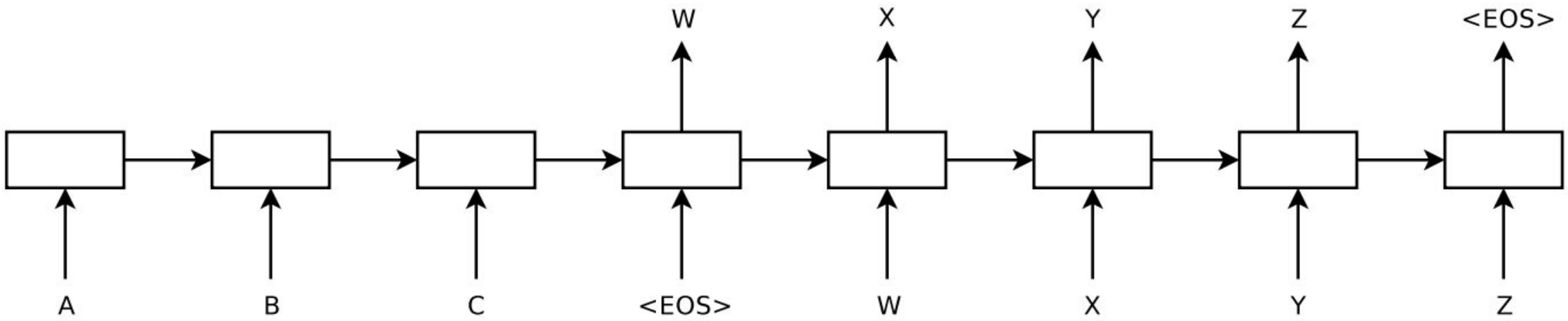


encoder и decoder - разные модели

encoder

decoder

Sequence to sequence learning with NN



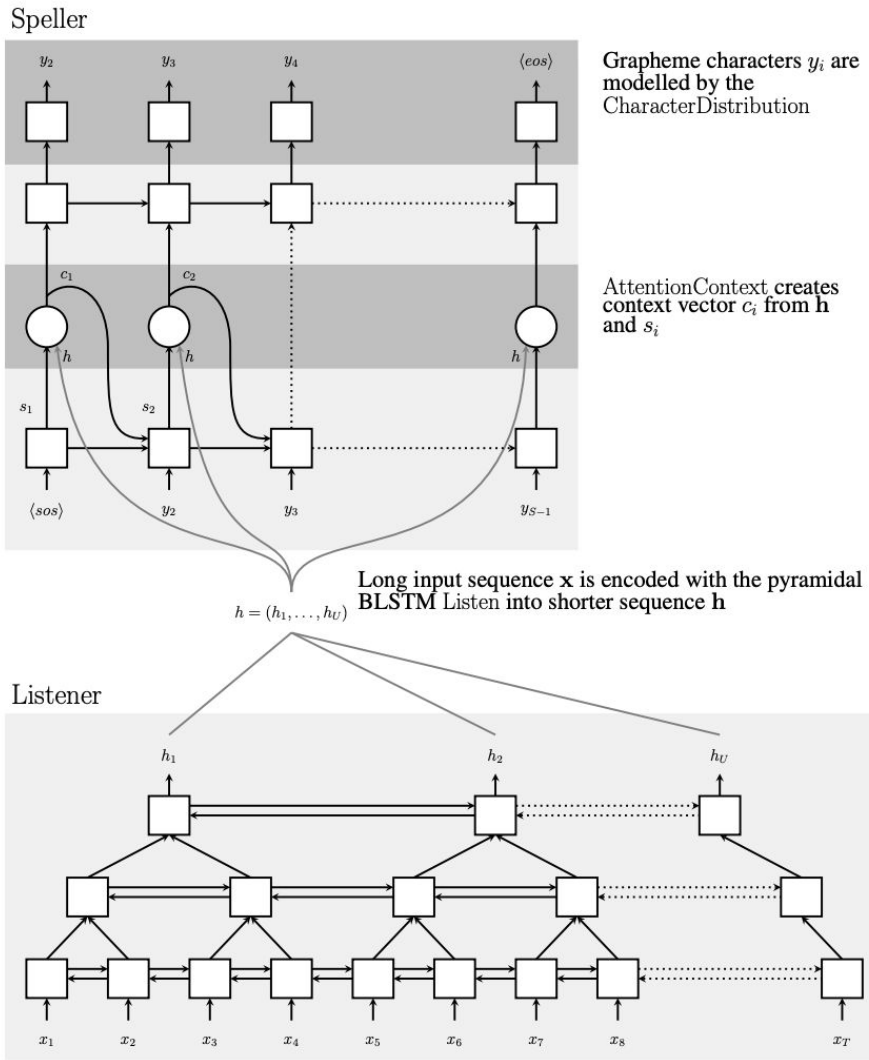
encoder

как можно улучшить подход?

decoder

LAS

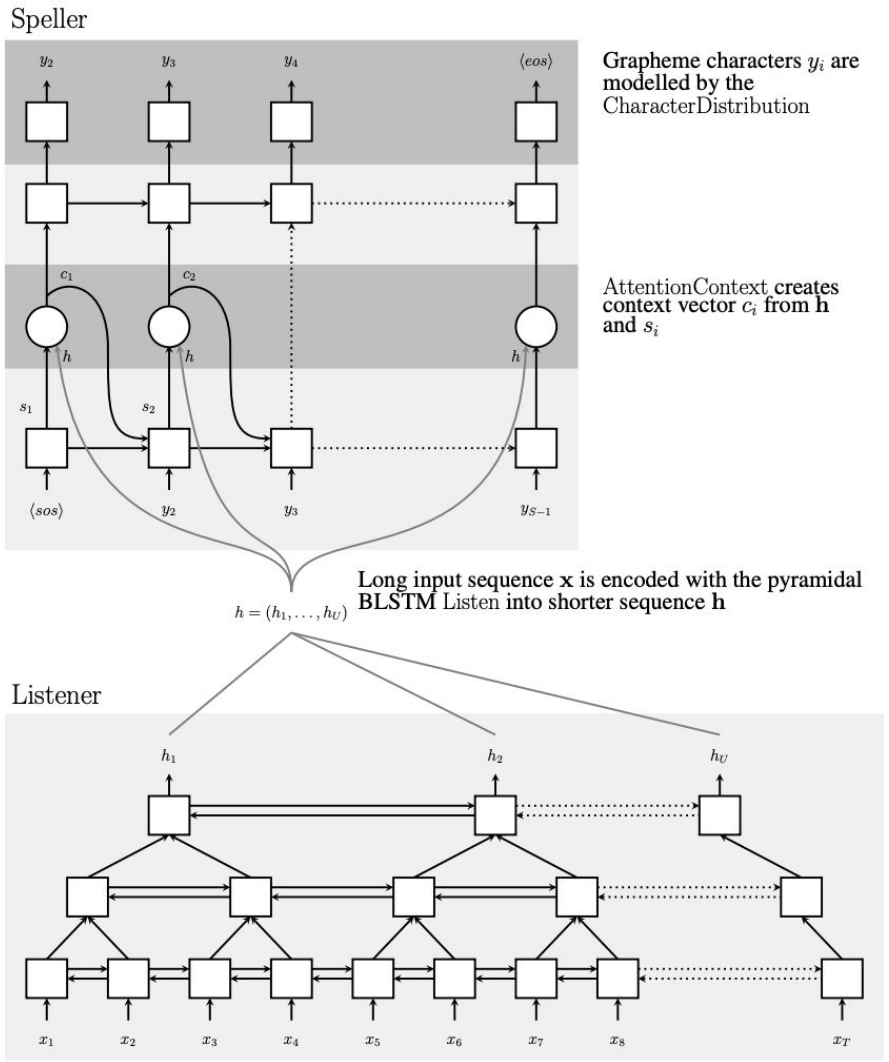
Listen Attend and Spell



декодер

энкодер

Listen Attend and Spell

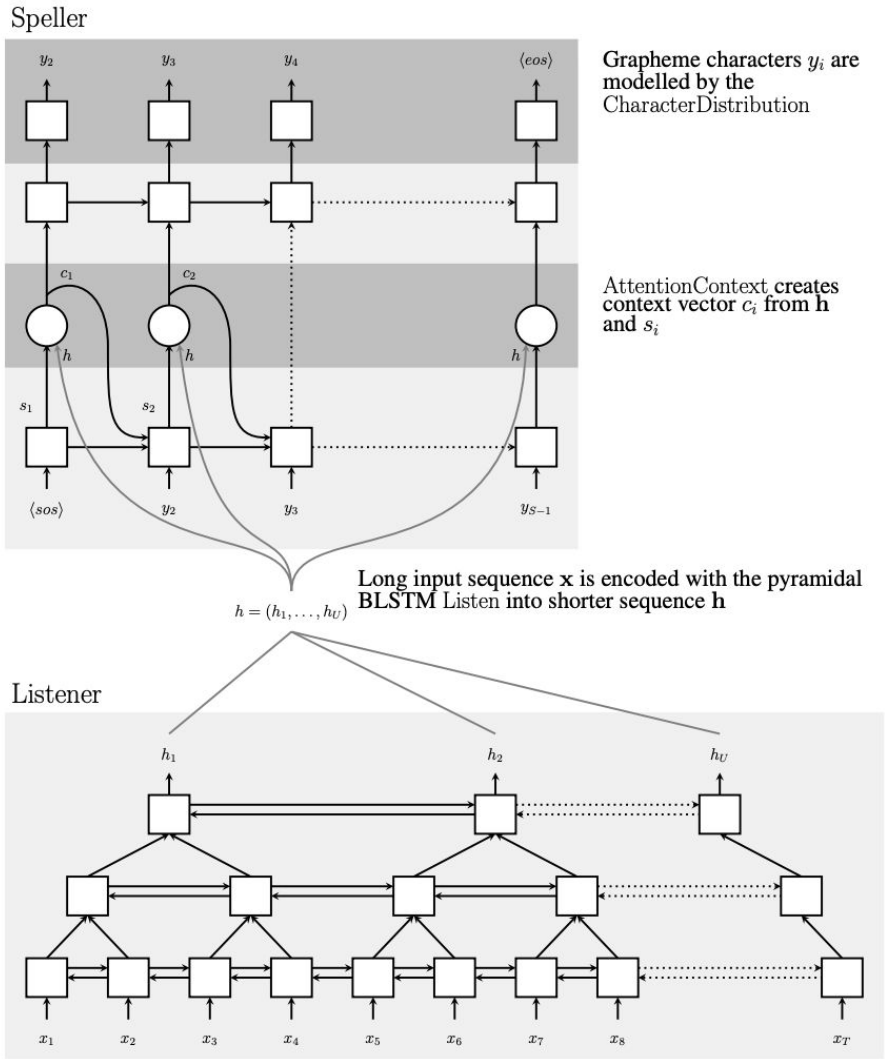


$$P(\mathbf{y}|\mathbf{x}) = \prod_i P(y_i|\mathbf{x}, y_{<i})$$

Теперь декодер контекстно зависимый

Какой здесь лосс?

Listen Attend and Spell

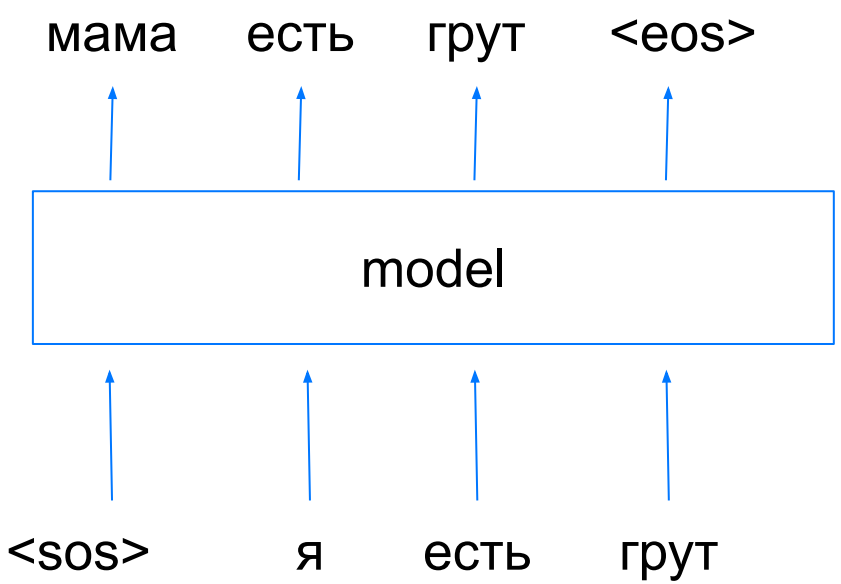


$$P(\mathbf{y}|\mathbf{x}) = \prod_i P(y_i|\mathbf{x}, y_{<i})$$

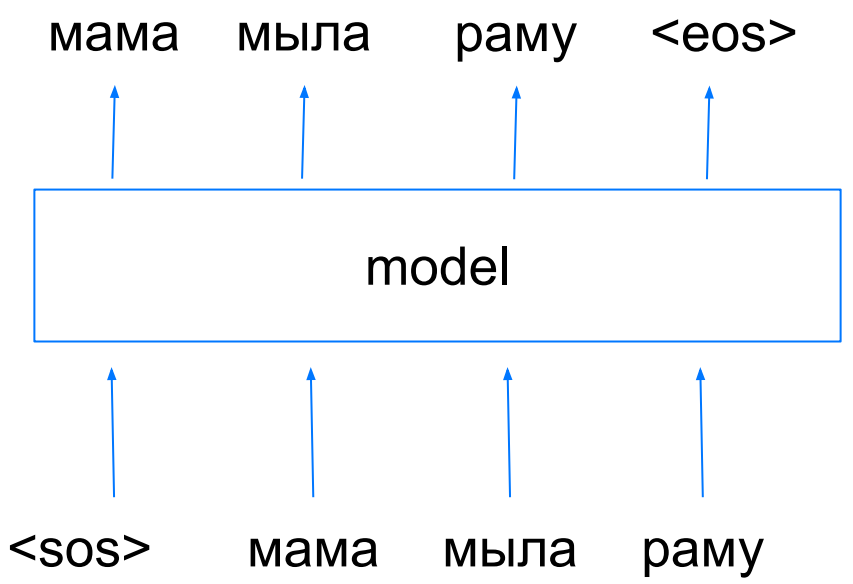
Оптимизируется кросс энтропия

$$L_{CE} = \sum CE_i$$

Teacher forcing VS Autoregressive Predictions



GT: <sos> я есть грут



GT: <sos> я есть грут

Хорошо, а где трансформеры?

Whisper

Основные идеи:

- Хотим получить супер классную модель, которая работает в zero-shot режиме одинаково хорошо на разных доменах
- Чем больше данных из разных доменов, тем будет лучше генерализация модели
- Сложно собрать большой сет данных, давайте собирать weakly supervised data (680K hours up to 3M hours)
- Растим модель, растим датасет
- Обучаем модель в мультитаск режиме (сразу несколько задач), это помогает сделать ее устойчивой к доменному сдвигу

Whisper

Multitask training data (680k hours)

English transcription

"Ask not what your country can do for ···"

Ask not what your country can do for ···

Any-to-English speech translation

"El rápido zorro marrón salta sobre ···"

The quick brown fox jumps over ···

Non-English transcription

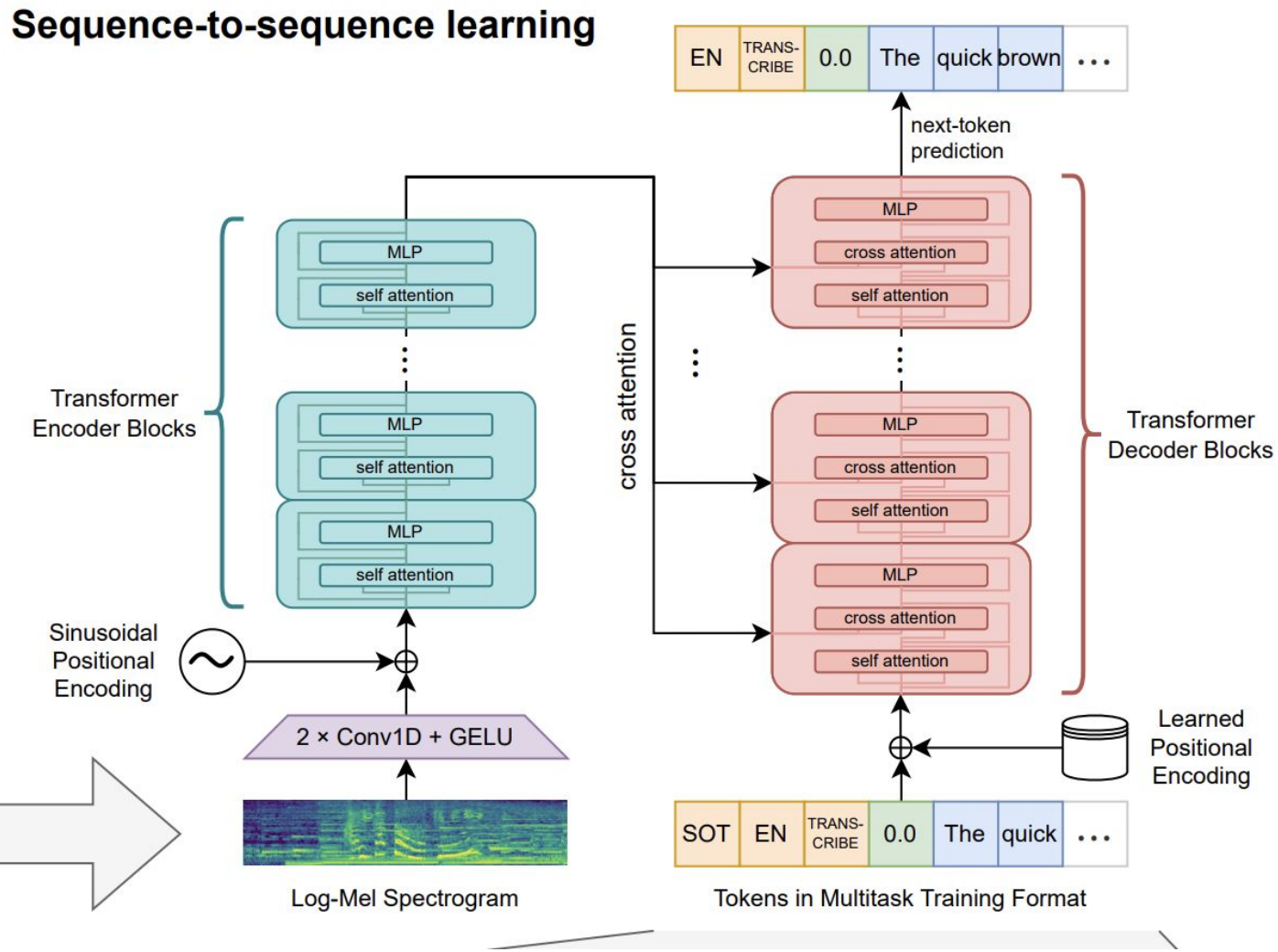
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···"

언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···

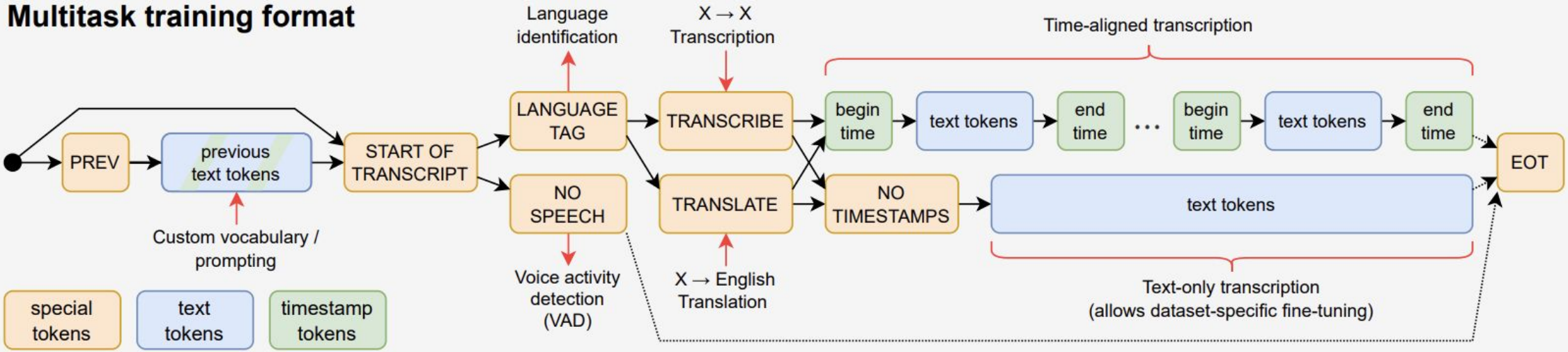
No speech

(background music playing)

Ø



Whisper Token format



Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Table 1. Architecture details of the Whisper model family.

Whisper Results

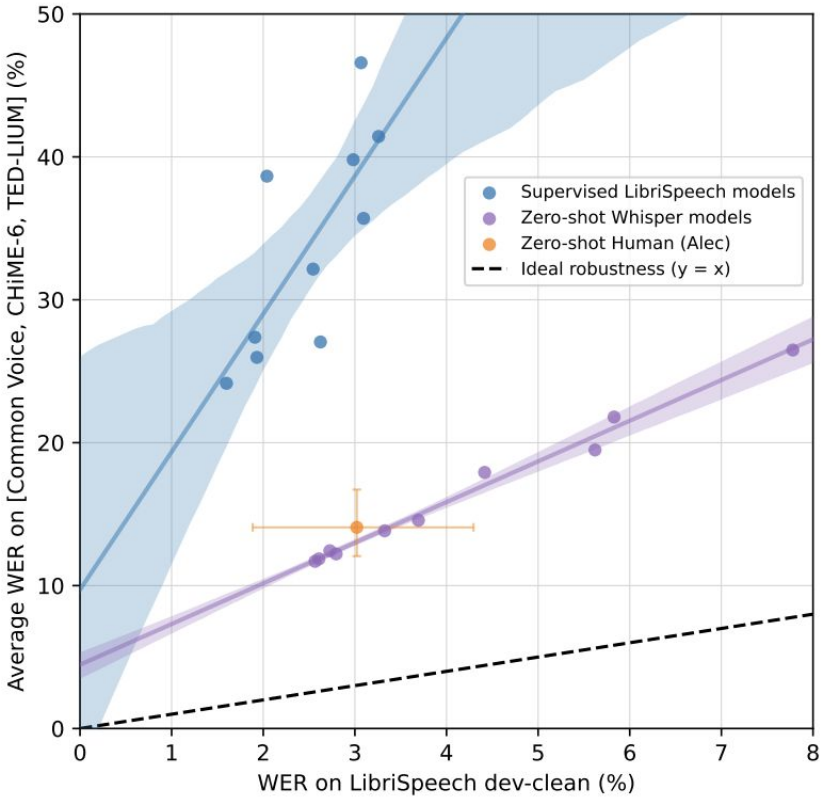


Figure 2. **Zero-shot Whisper models close the gap to human robustness.** Despite matching or outperforming a human on LibriSpeech dev-clean, supervised LibriSpeech models make roughly twice as many errors as a human on other datasets demonstrating their brittleness and lack of robustness. The estimated robustness frontier of zero-shot Whisper models, however, includes the 95% confidence interval for this particular human.

Dataset	wav2vec 2.0 Large (no LM)	Whisper Large V2	RER (%)
LibriSpeech Clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.2**	74.7
Common Voice	29.9	**9.0**	69.9
Fleurs En	14.6	**4.4**	69.9
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.5**	61.2
VoxPopuli En	17.9	**7.3**	59.2
CORAAL	35.6	**16.2**	54.5
AMI IHM	37.0	**16.9**	54.3
Switchboard	28.3	**13.8**	51.2
CallHome	34.8	**17.6**	49.4
WSJ	7.7	**3.9**	49.4
AMI SDM1	67.6	**36.4**	46.2
LibriSpeech Other	6.2	**5.2**	16.1
Average	29.3	**12.8**	55.2

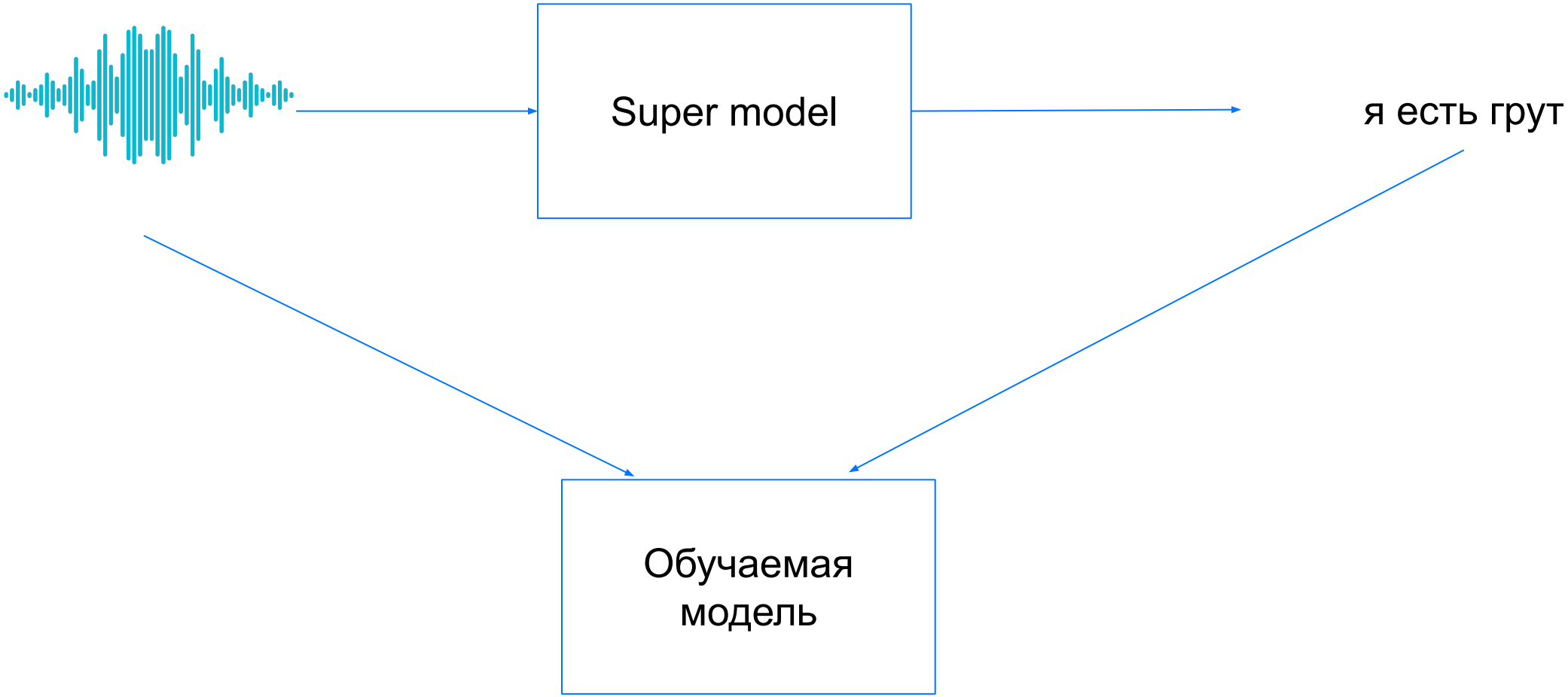
Table 2. **Detailed comparison of effective robustness across various datasets.** Although both models perform within 0.1% of each other on LibriSpeech, a zero-shot Whisper model performs much better on other datasets than expected for its LibriSpeech performance and makes 55.2% less errors on average. Results reported in word error rate (WER) for both models after applying our text normalizer.

the smallest zero-shot Whisper model, which has only 39

А как улучшить ASR без такого огромного количества данных?

Как насчет дистилляции моделей?

Hard distillation



Distillation by embeddings

teacher hiddens

student hiddens

Loss = CE + L2(student hiddens - teacher hiddens)

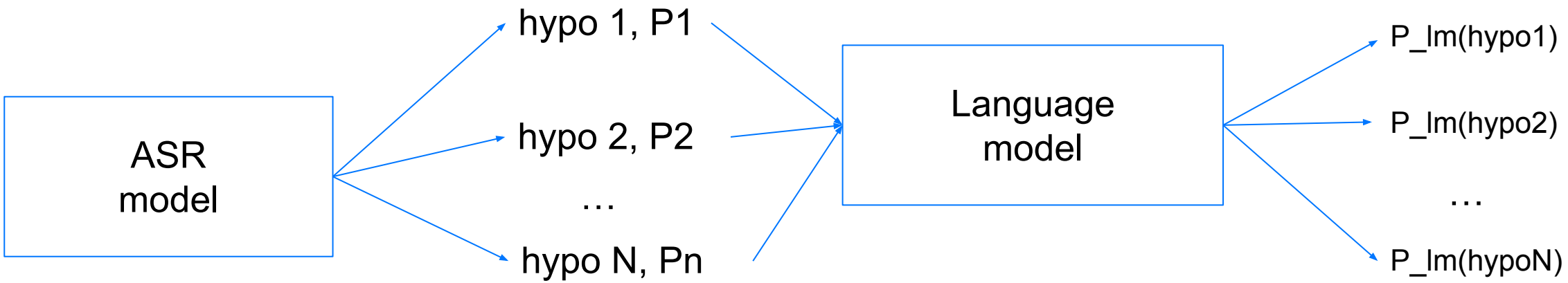
student hiddens - teacher hiddens = error vector = zero vector

student << teacher

Можем ли как то использовать языковую модель (LM)?

Second pass rescoring

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \log P(\mathbf{y}|\mathbf{x}) + \lambda \log P_{LM}(\mathbf{y}) + \gamma \text{len}(\mathbf{y})$$

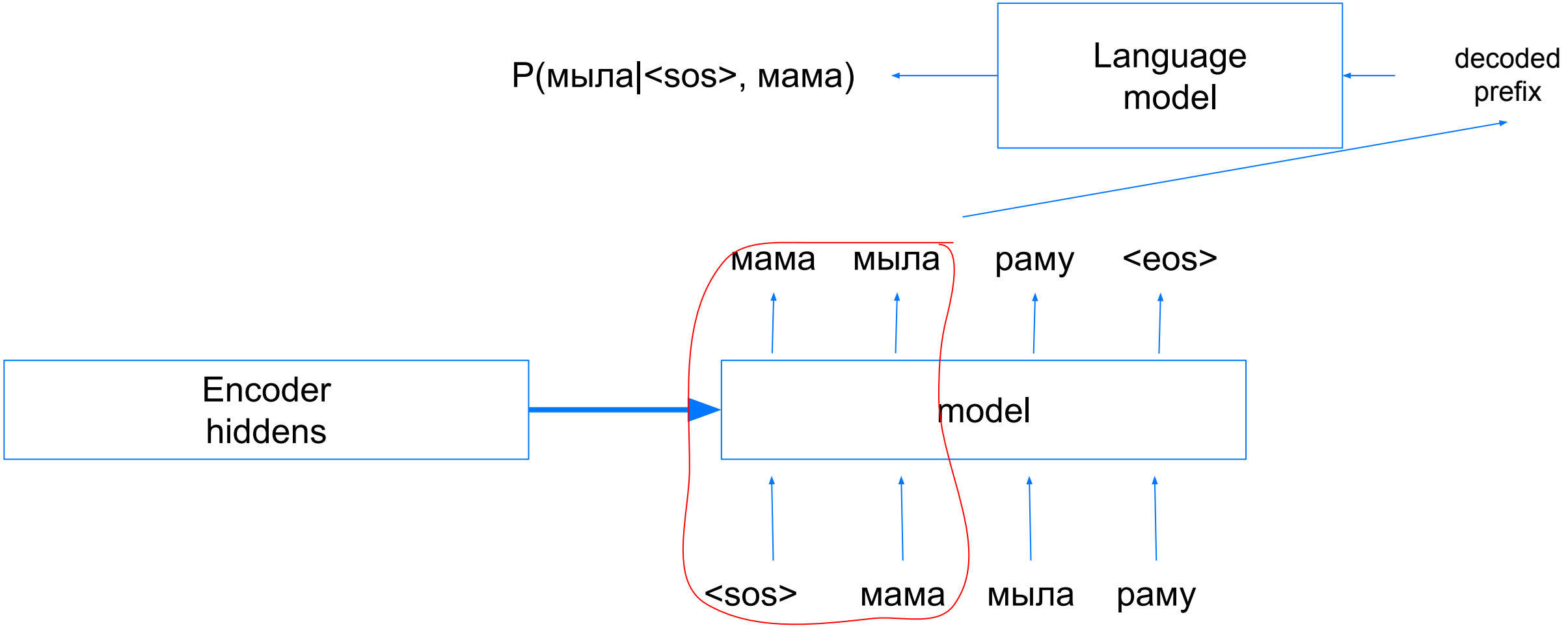


P_lm("мама мыла раму") = 0.25
P_lm("мама ыла рму") = 0.0001

hypos:
мама мыла раму
мама ыла рму
мамамамама
мыла рамума

GT: мама мыла раму

First Pass Rescoring

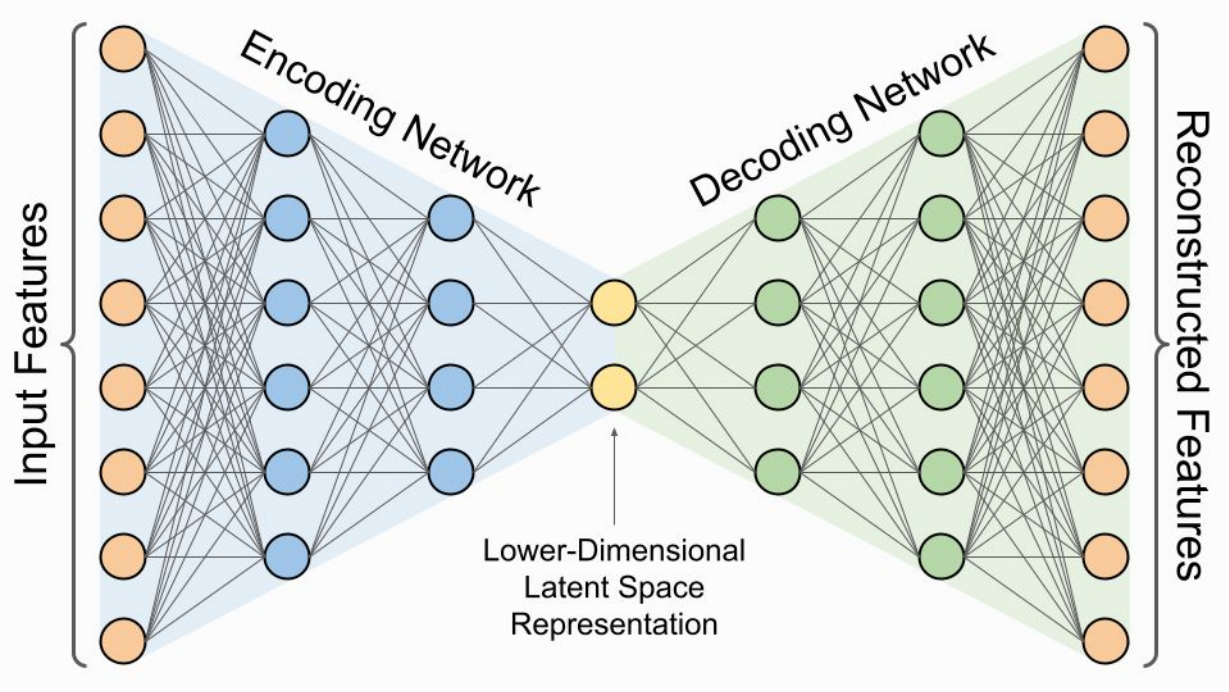


$$y_i^* = argmax_{y_i} P_{ASR}(y_i|x, y_{<i}) + \lambda P_{LM}(y_i|y_{<i})$$

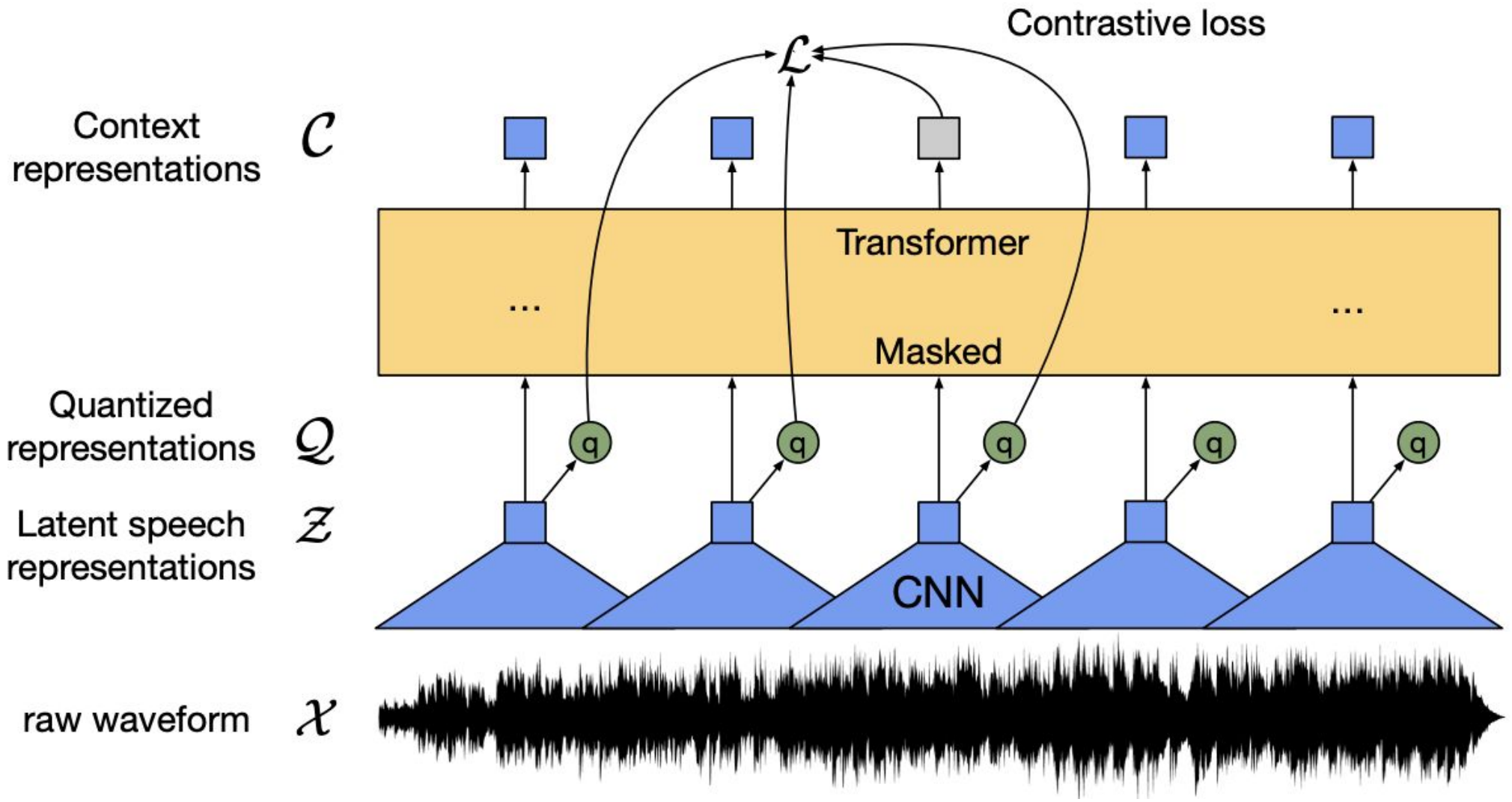
Self-supervised learning

Self-supervised Learning

Self-Supervised Learning (или unsupervised) - режим обучения моделей, при котором задача обучения не требует дополнительной разметки и формируется исходя из внутренней структуры самих объектов, либо из базовых знаний об объектах.



Wav2Vec2.0


Contrastive loss
$\mathcal{L}$
Context representations
$\mathcal{C}$
Transformer
...
...
Masked
Quantized representations
$\mathcal{Q}$
q
q
q
q
q
Latent speech representations
$\mathcal{Z}$
CNN
raw waveform
$\mathcal{X}$

Wav2Vec2.0

Picking codewords in differentiable way (Gumbel softmax)

nn.Linear

q t

Wav2Vec2.0

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d$$

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

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}$$

Дополнительно по теме

- Gumbel softmax
- MLM task (BERT)
- Product quantization
- Contrastive loss

Что сейчас в тренде?

Что еще посмотреть

- https://web.stanford.edu/~jurafsky/slp3/
- rnn-t
- librispeech leaderboard
- hf leader board
- decoder llm https://arxiv.org/pdf/2402.08846v1

Спасибо
за внимание!