LISTEN. ATTENTION. SPELL

William Chan, Navdeep Jaitly, Quoc V. Le, Oriol Vinyals
Google, 2015

Подготовила Гаврилова А. А.

ПЛАН

- ***Введение в ASR***
  - Препроцессинг звуковой дорожки
  - Акустическая модель
  - Языковая модель
  - Пунктуационная модель
  - Метрики
- ***Listen, Attend and Spell***
  - Архитектура
  - Функция потерь
  - Beam-Search decoding
  - Результаты

АУДИОФОРМАТЫ

- **Аналоговый сигнал**

  непрерывное множество значений

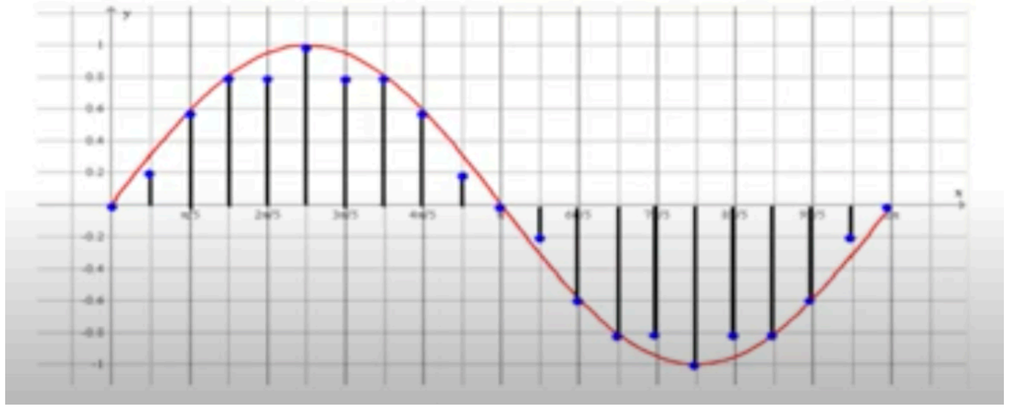
- **Цифровой сигнал**

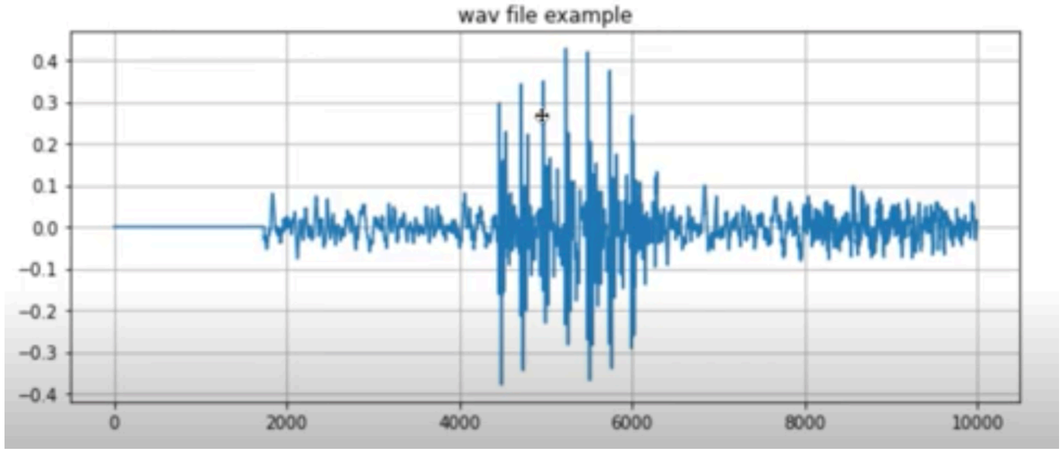
  последовательность дискретных значений

КАК ХРАНИТСЯ АУДИО?

- sample_rate - число отсчетов в секунду (16000, 22050)

- Librosa - библиотека для работы со звуком

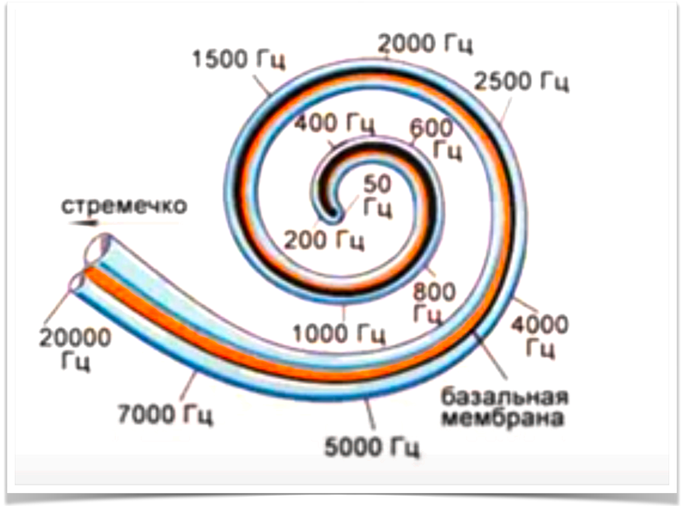




КАК ПЕРЕВЕСТИ АУДИО В ТЕНЗОР?

- Будем работать как с картинками (с ними умеем работать)

$$G(f) = \int_{-\infty}^{\infty} g(t) e^{-2\pi i f t} dt.$$



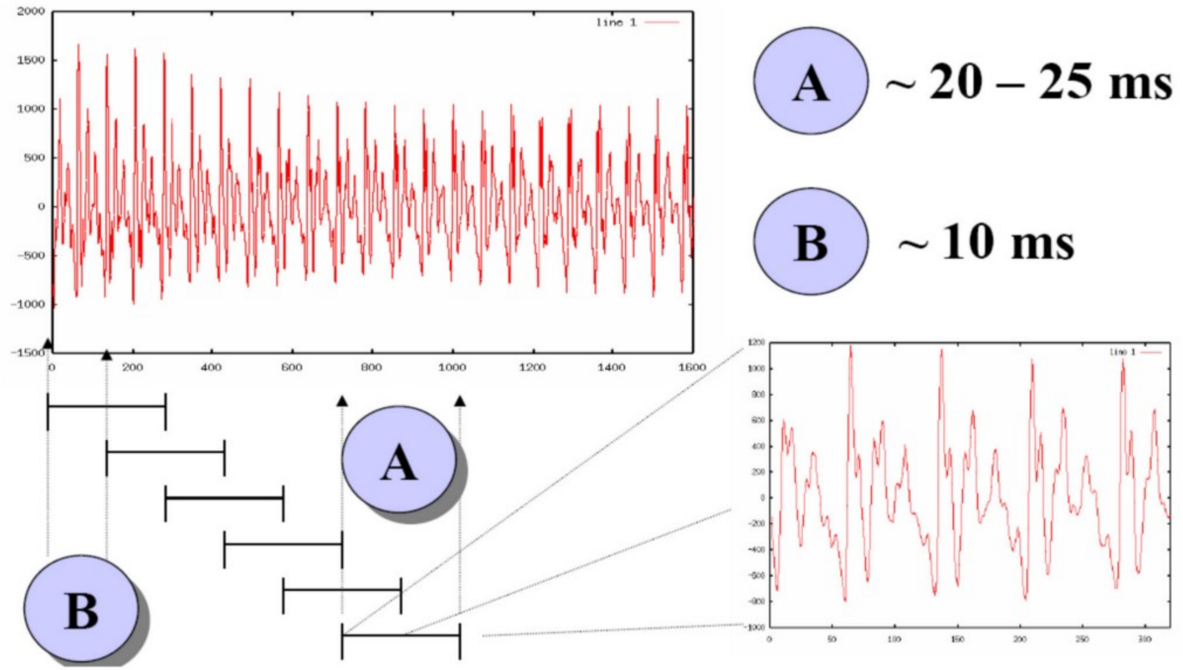
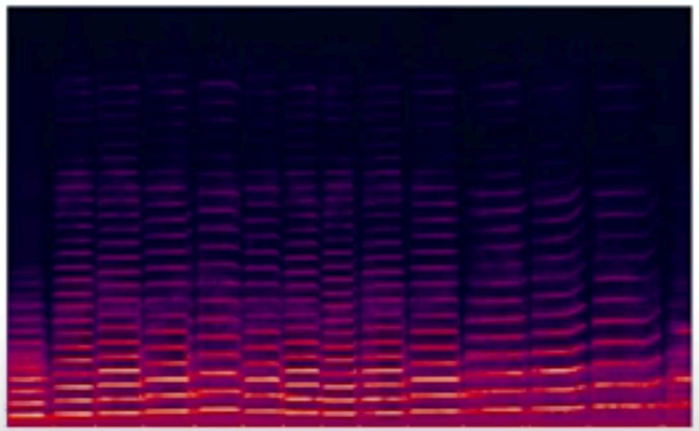
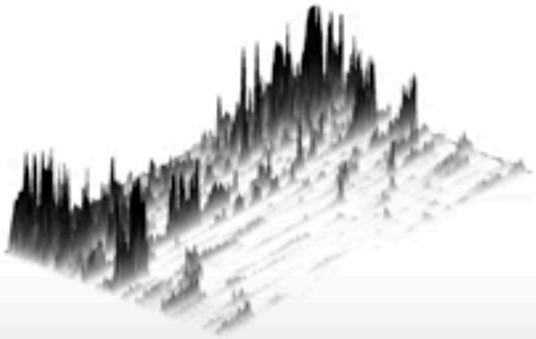


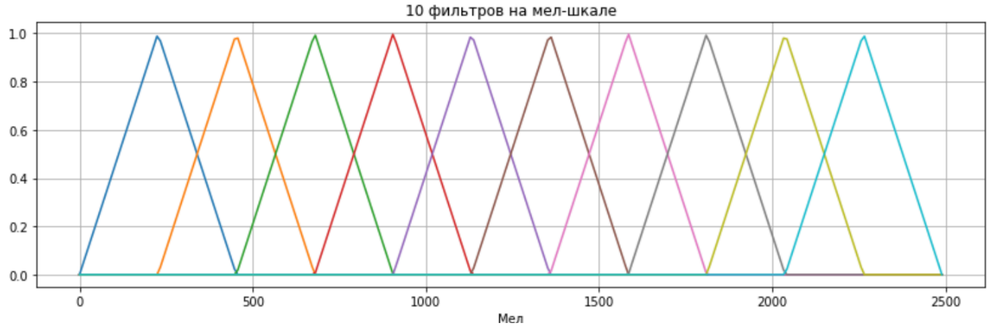
Image from Bryan Pellom

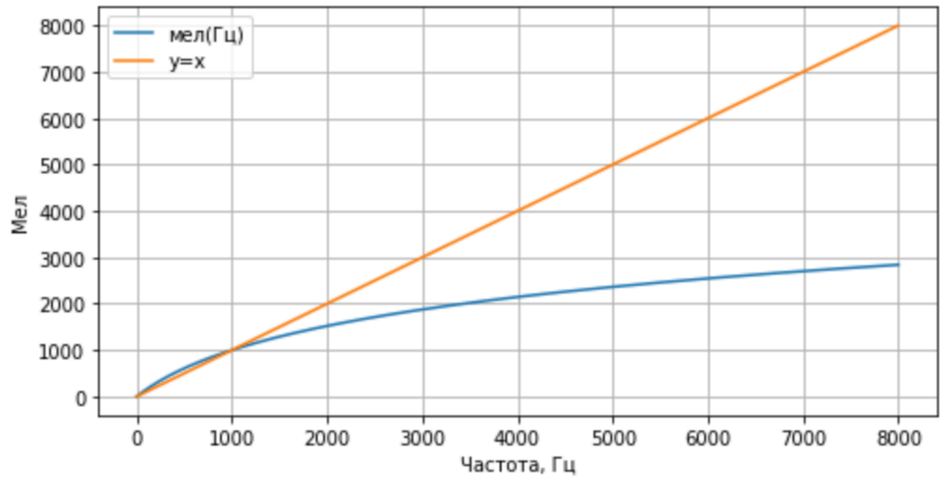
МЕЛ-СПЕКТРОГРАММА

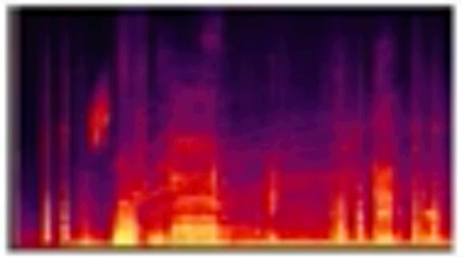
$$mel = 1127.01048 \ln(1 + \frac{freq}{700})$$











приве янадя

акустическая
модель

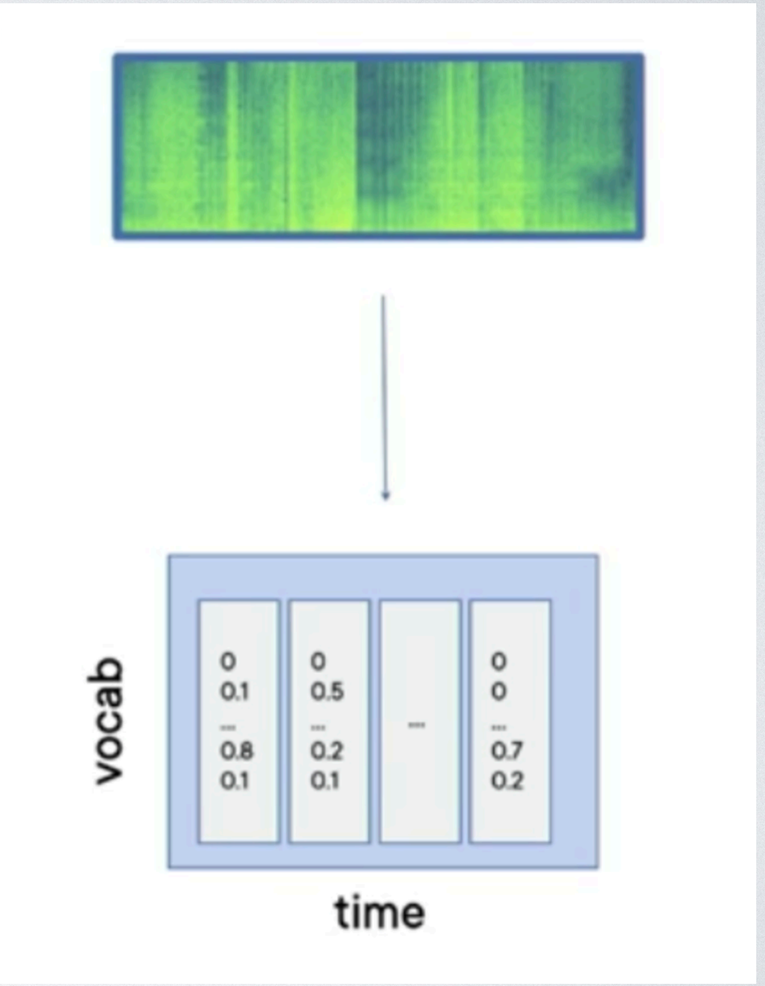
привет я надя

языковая
модель

profit

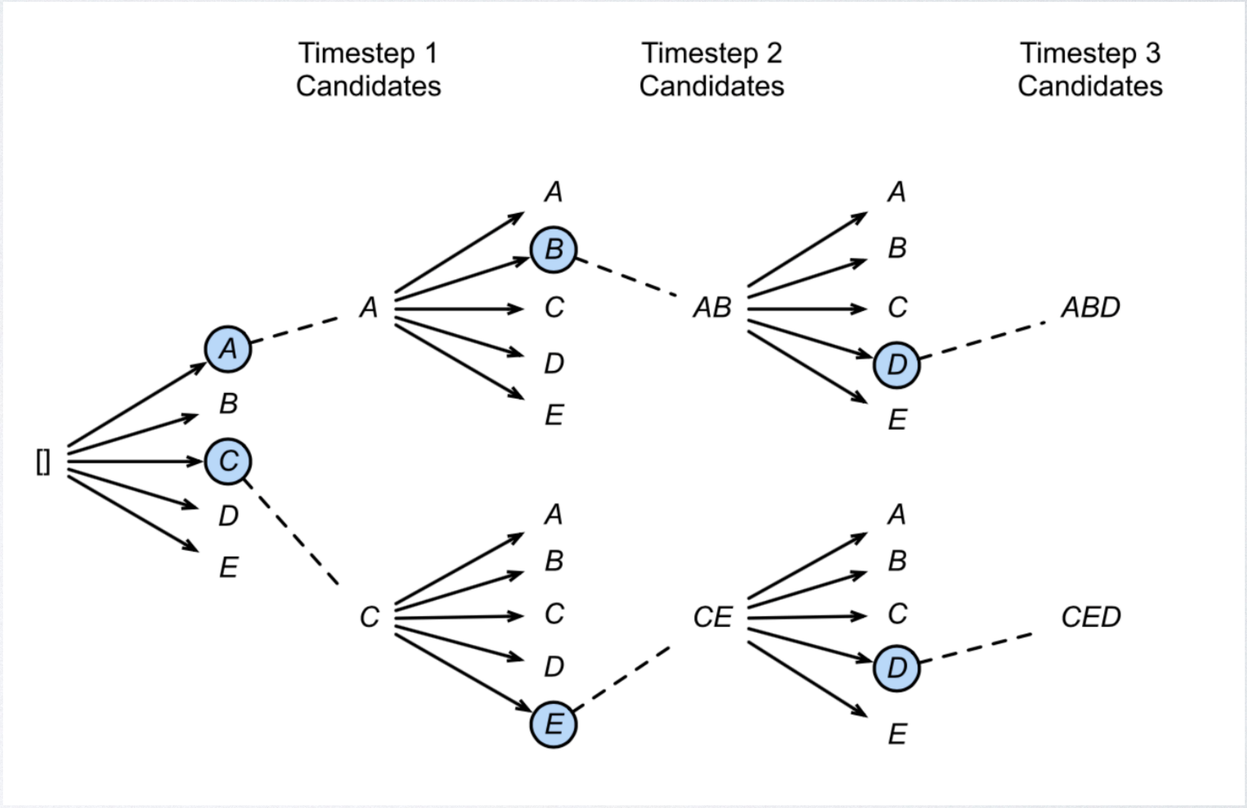
АКУСТИЧЕСКАЯ МОДЕЛЬ

- В качестве простой акустической модели в задаче распознавания речи использовались марковские модели, затем нейросети
- На выходе матрица распределения по времени вероятностей каждой фонемы

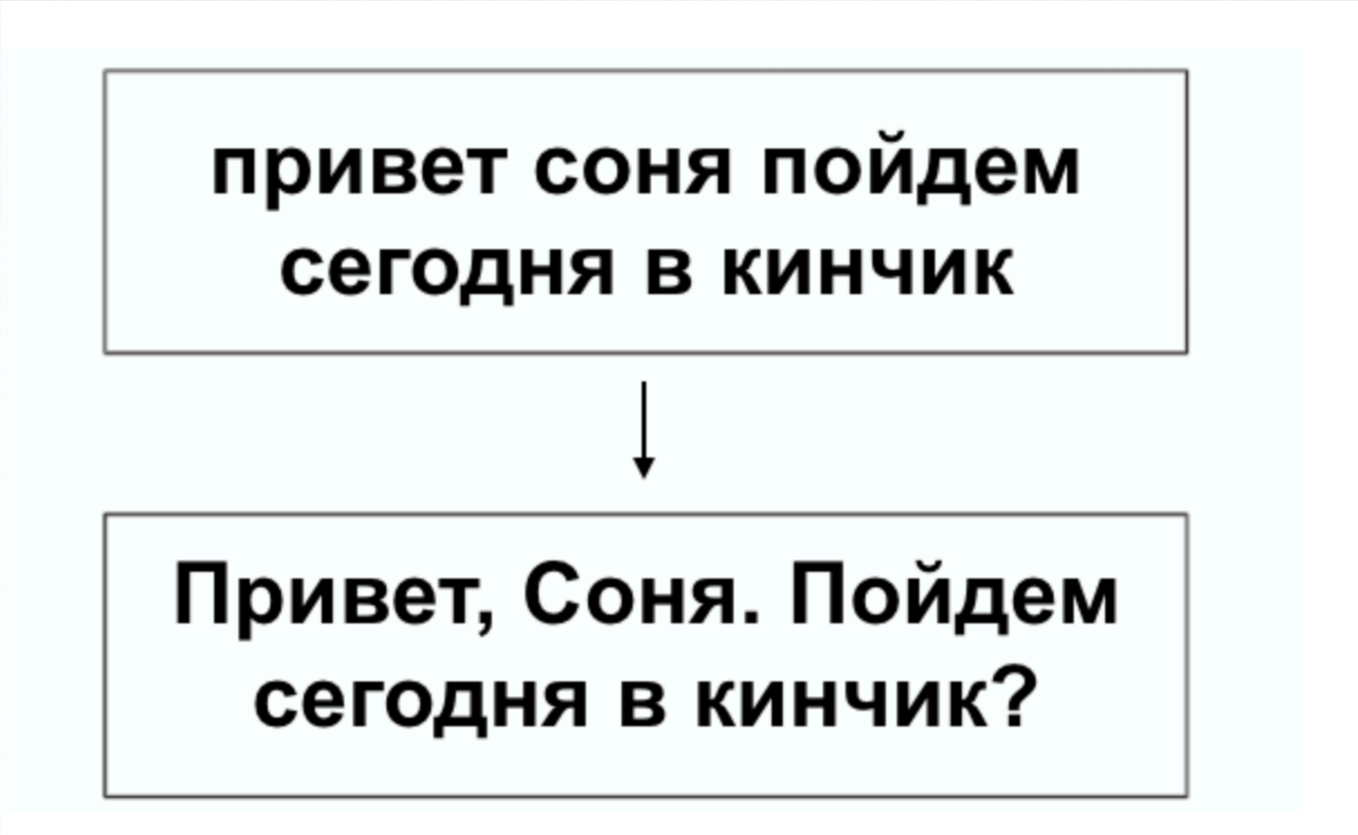


ЯЗЫКОВАЯ МОДЕЛЬ

Beam-search



ПУНКТУАЦИОННАЯ МОДЕЛЬ



МЕТРИКИ

WER

```
True: quick brown     fox jumped over a lazy dog
Pred: quick brow  an  fox jumped over   lazy dog
```

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

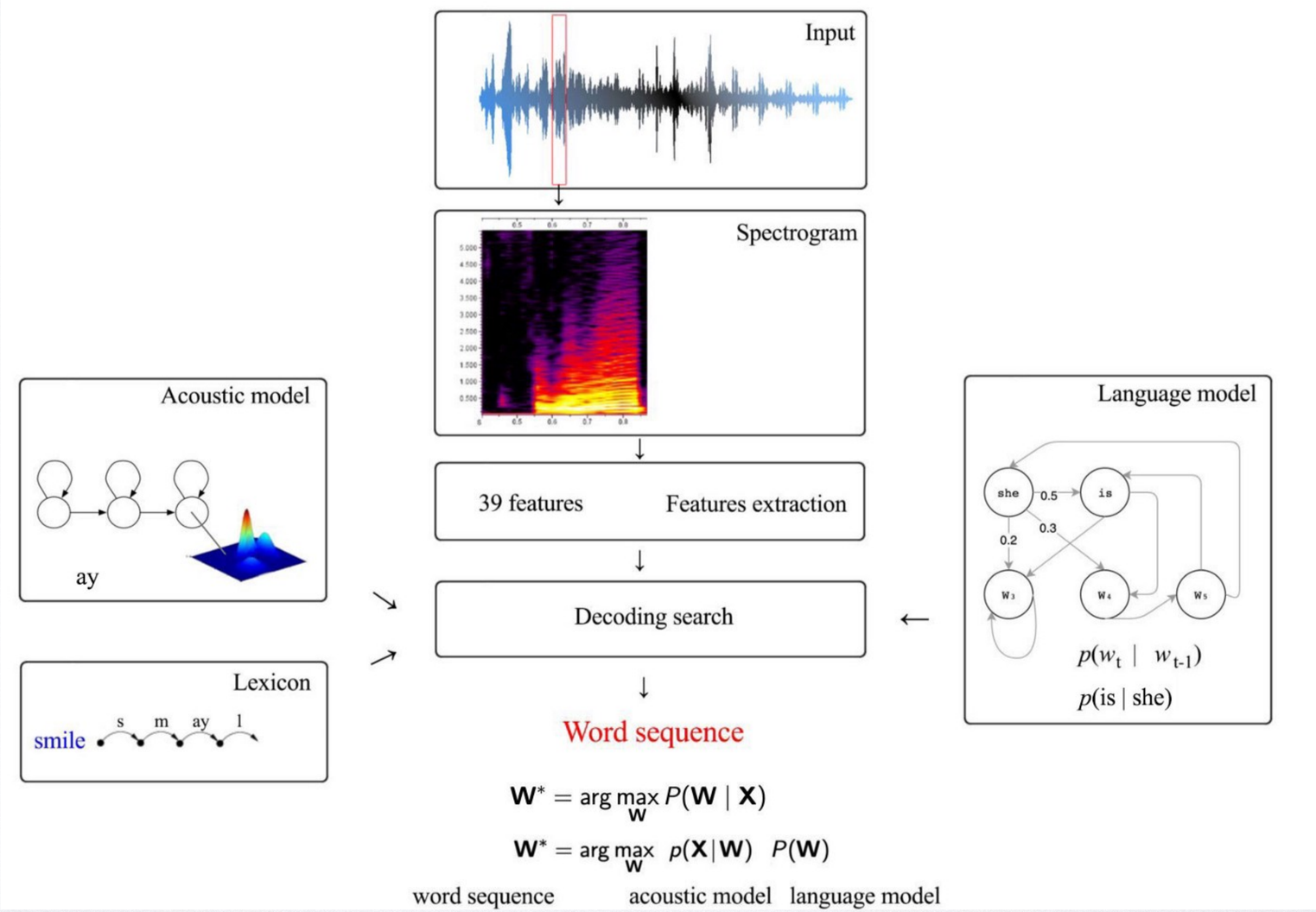
- S – число замен
- D – число удалений
- I – число вставок
- C – число корректных слов

N – **всего слов**

CER

То же, но с символами

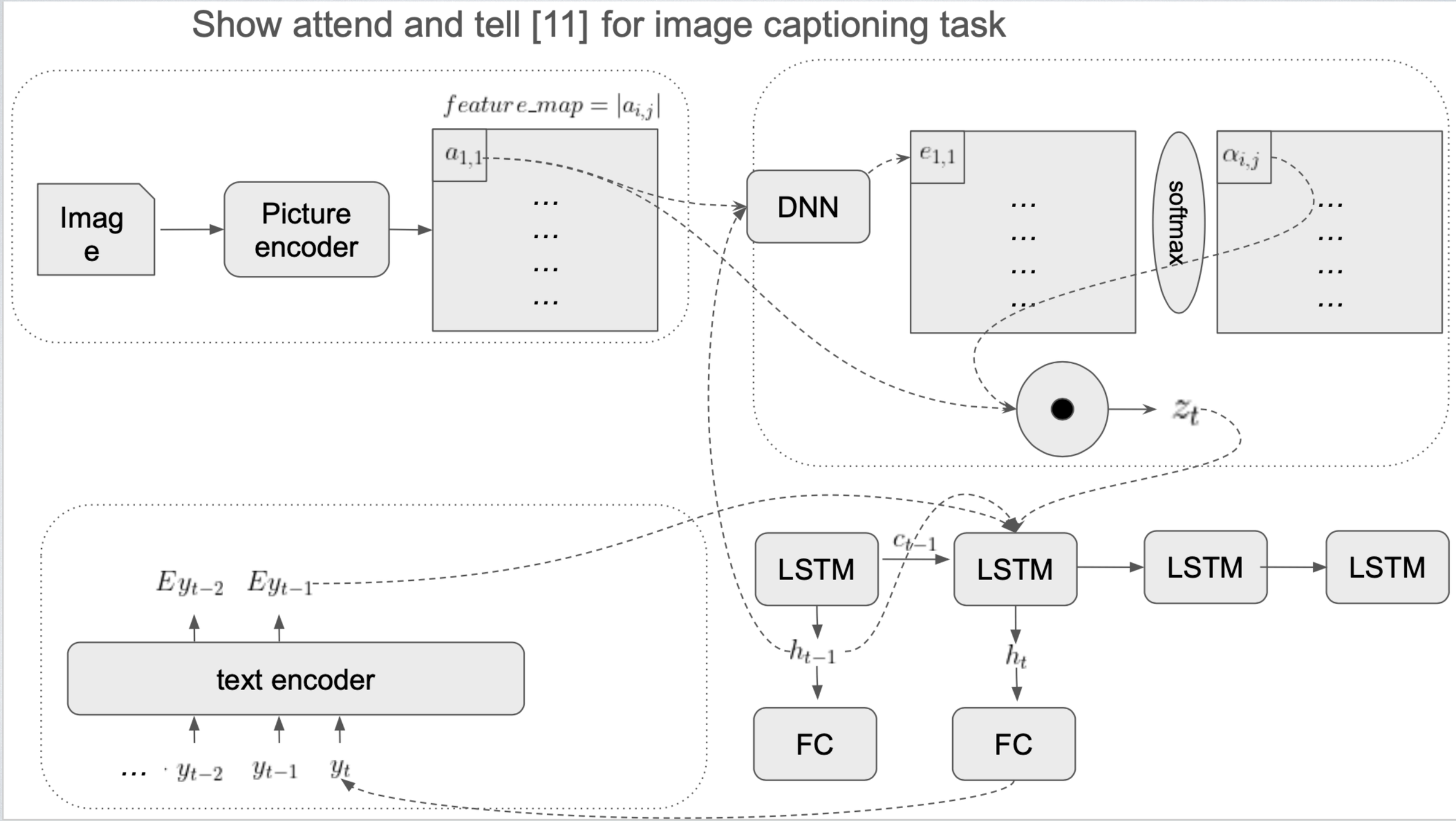
ДО НЕЙРОСЕТЕЙ

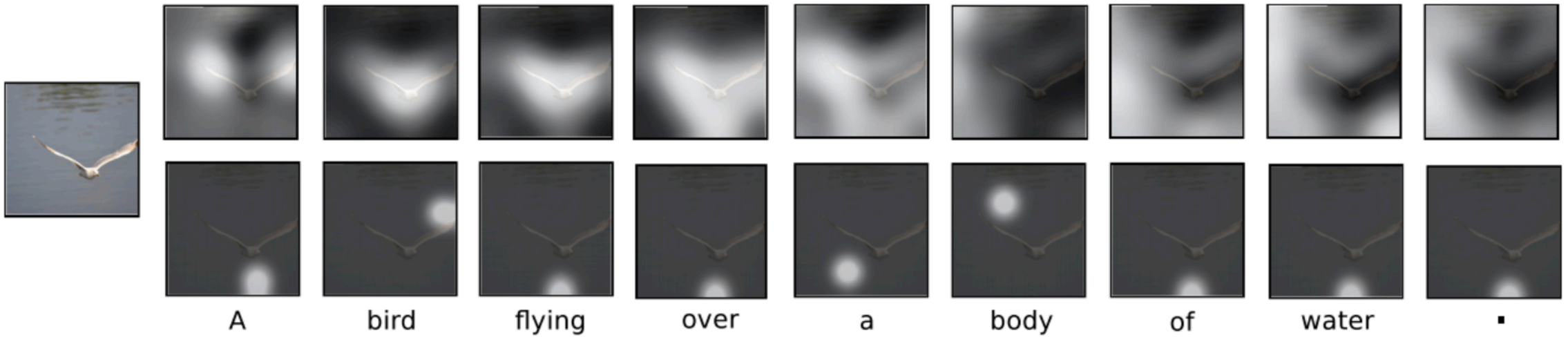


ВСПОМНИМ ATTENTION : IMAGES

A young boy is playing basketball. 	Two dogs play in the grass. 	A dog swims in the water. 
A group of people walking down a street. 	A group of women dressed in formal attire. 	Two children play in the water. 
A skier is skiing down a snowy hill. 	A little girl in a pink shirt is swinging. 	A dog jumps over a hurdle. 

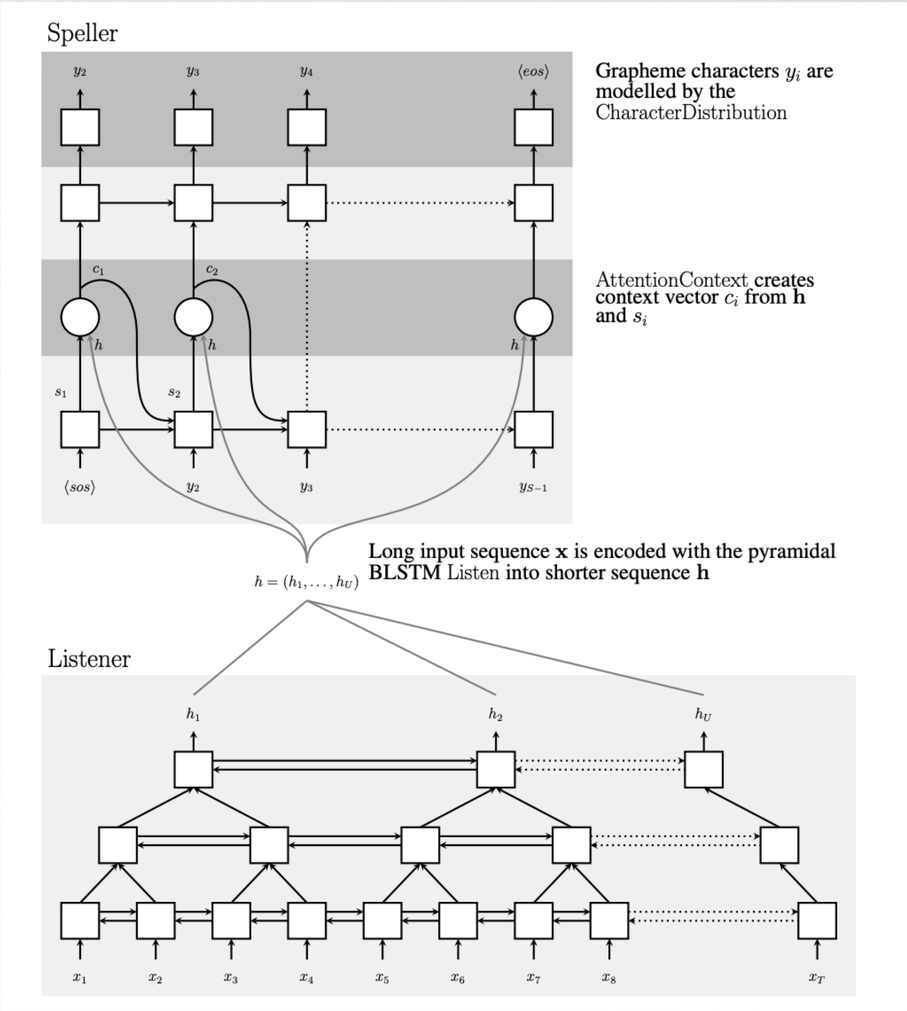
Show attend and tell [11] for image captioning task




A
bird
flying
over
a
body
of
water
.

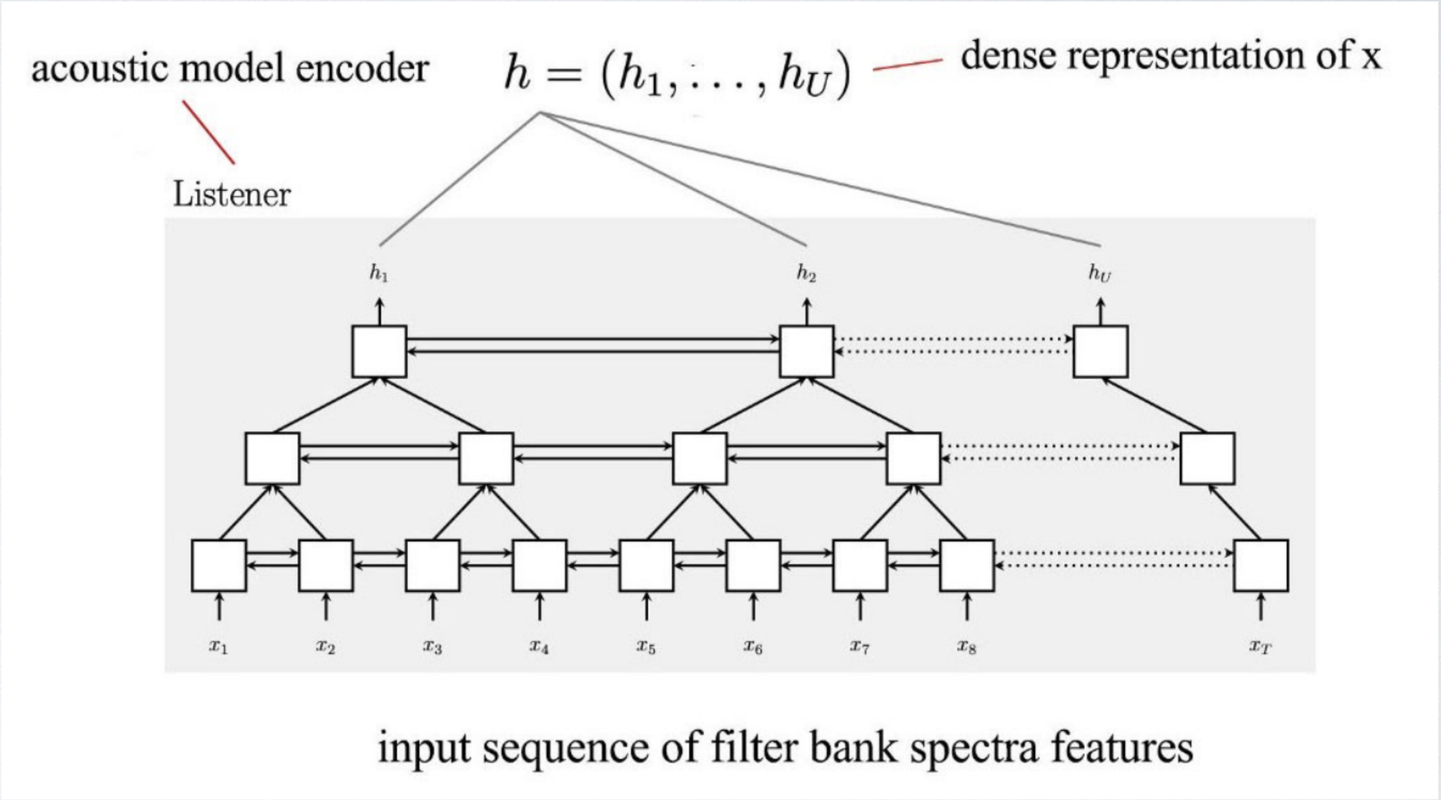
LISTEN. ATTEND.SPELL

- Цель - избавиться от независимости выходов
- Listener - енкодер
- Speller - декодер (получаем транскрипцию
- Attend - сообщает декодеру, какая часть дорожки релевантна в данный шаг

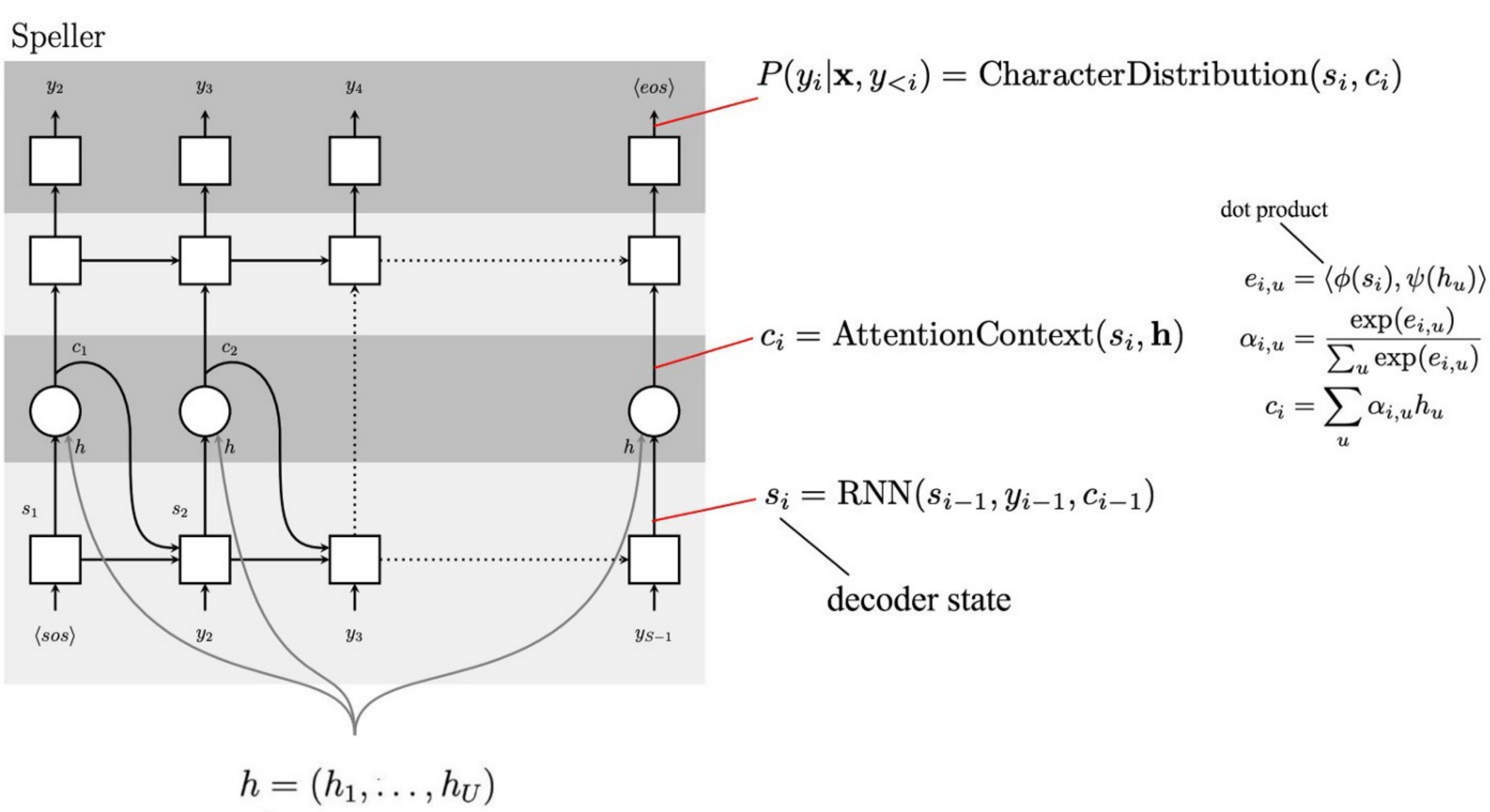


LISTENER

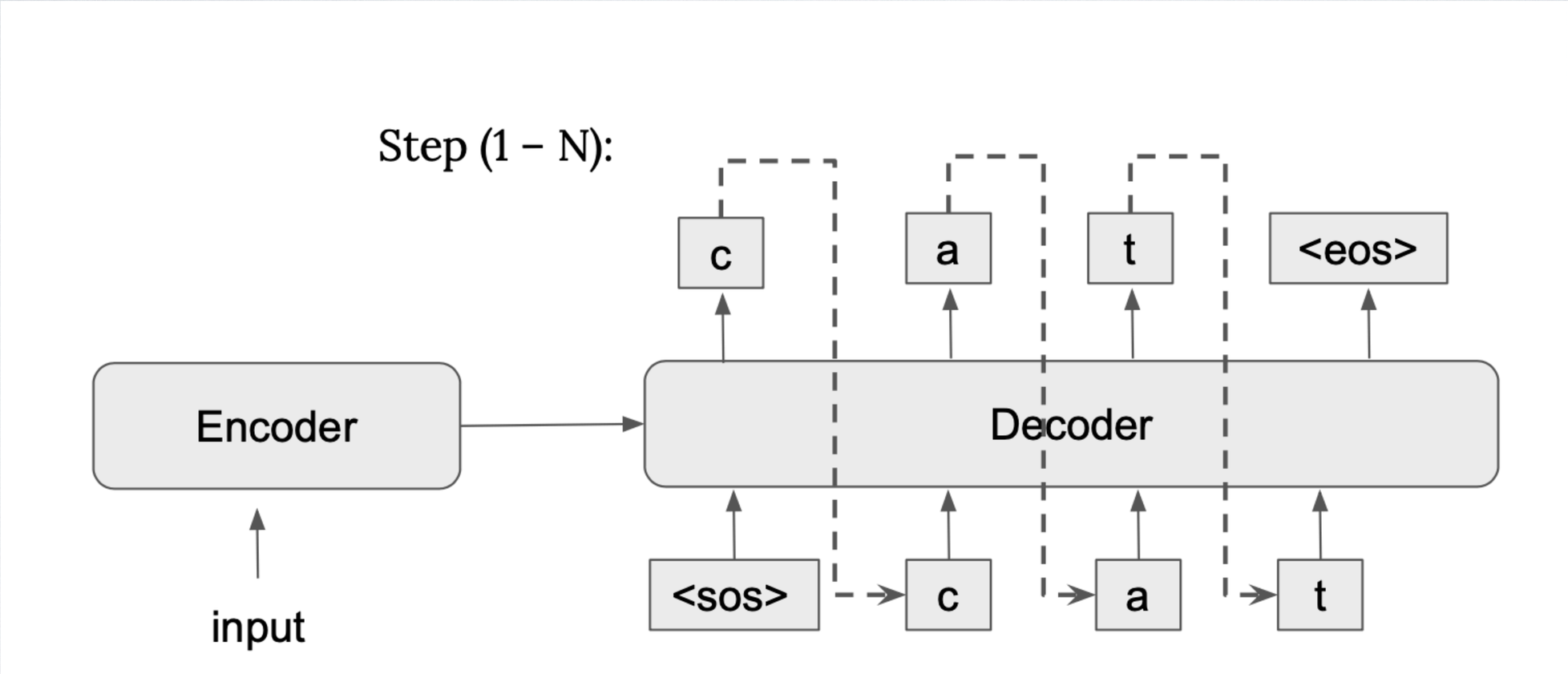
- Пирамидальная bi-LSTM



SPELLER


Speller
y_2
y_3
y_4
$\langle eos \rangle$
c_1
c_2
h
h
h
s_1
s_2
$\langle sos \rangle$
y_2
y_3
y_{S-1}
$P(y_i|\mathbf{x}, y_{<i}) = \text{CharacterDistribution}(s_i, c_i)$
$c_i = \text{AttentionContext}(s_i, \mathbf{h})$
$s_i = \text{RNN}(s_{i-1}, y_{i-1}, c_{i-1})$
decoder state
dot product
$e_{i,u} = \langle \phi(s_i), \psi(h_u) \rangle$
$\alpha_{i,u} = \frac{\exp(e_{i,u})}{\sum_u \exp(e_{i,u})}$
$c_i = \sum_u \alpha_{i,u} h_u$
$h = (h_1, \ldots, h_U)$

INFERENCE



- total time = I x run_encoder_time + T x run_decoder_time

CTC VS LAS

	CTC *beam search*	**LAS** *beam search*
Computational cost	• 1 * *run_encoder()* • T * beam_size * *expand_beam()*	• 1 * *run_encoder()* • T * beam_size * *run_decoder()*
Path merging		
Overall experience	• painful to code • joyful to run	• painful to code • painful to run

РЕЗУЛЬТАТЫ

Method	**Year**	**WER (test-clean)**	**WER (test-other)**
Human	~ 200 000 b.c.	5.83	12.69
Deep Speech 2 [5]	2015	5.15	12.73
LAS (original paper [8])	2015	----	----
LAS [9*]	2019	**3.2**	**9.8**

ВОПРОСЫ

- Что подается на вход акустической модели?
- Чем различается beam-search в CTC и LAS?
- В чем основная цель Listen, Attention, Spell подхода?

РЕСУРСЫ

- https://www.youtube.com/watch?v=3MjIkWxXigM lectures
- https://arxiv.org/pdf/1508.01211.pdf Listen.Attend.Spell
- https://habr.com/ru/company/vk/blog/579412/ Как работает распознавание речи ВК
- https://www.youtube.com/watch?v=fodf4Pttve4 Введение в обработку звука