Обзор технологий и подходов в задаче end-to-end ASR

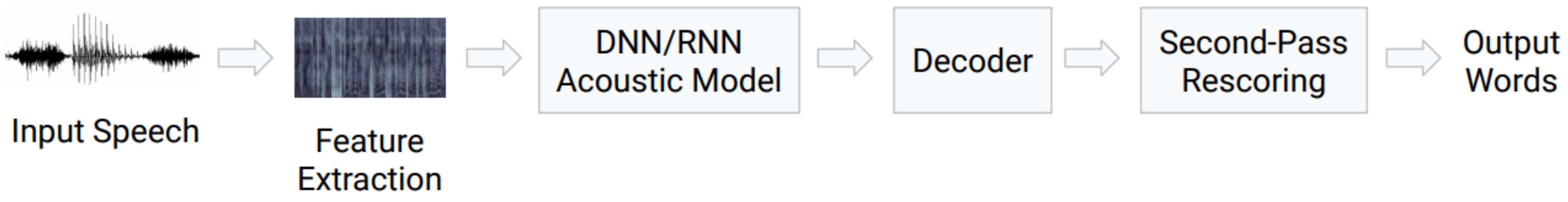
ПАО «МТС»
Ведущий разработчик
Семенов Никита

Ведите бизнес вперёд

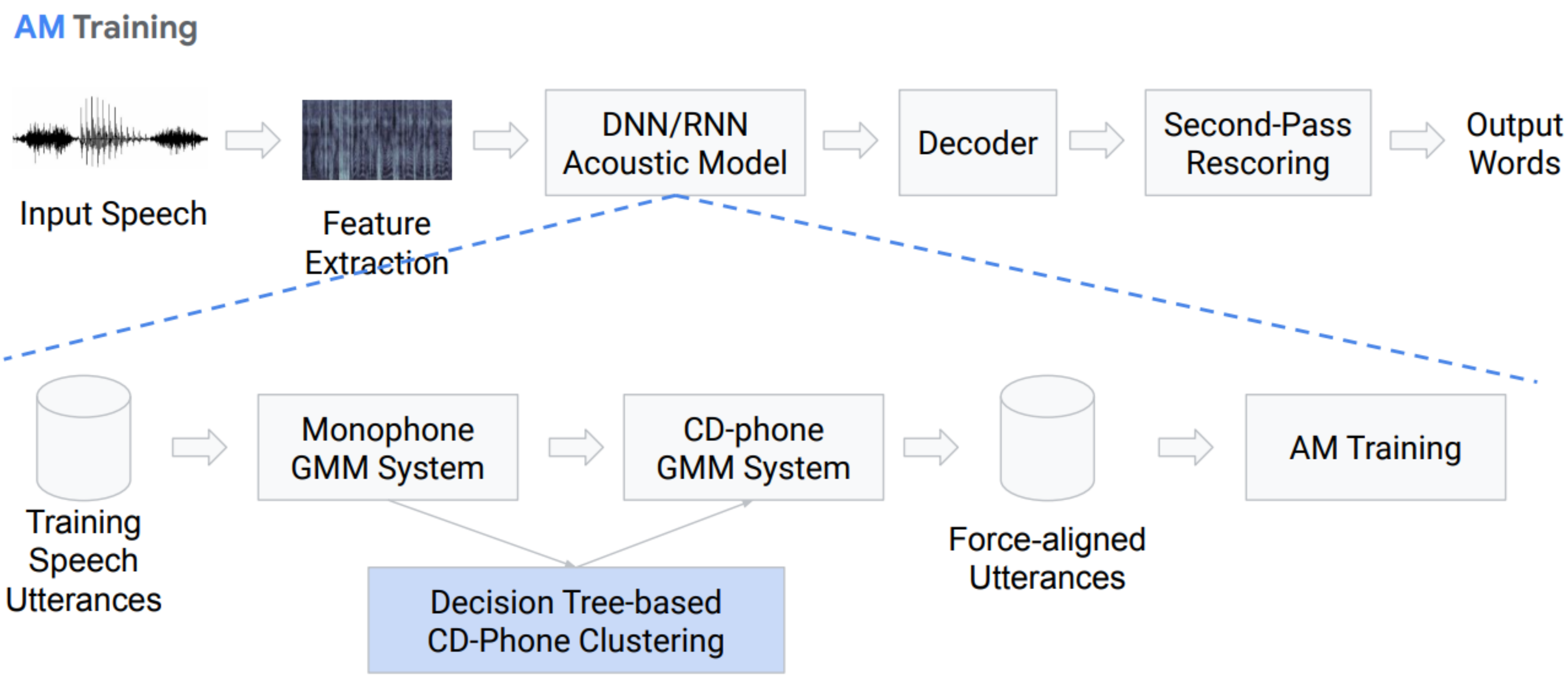
Классический подход

Conventional ASR

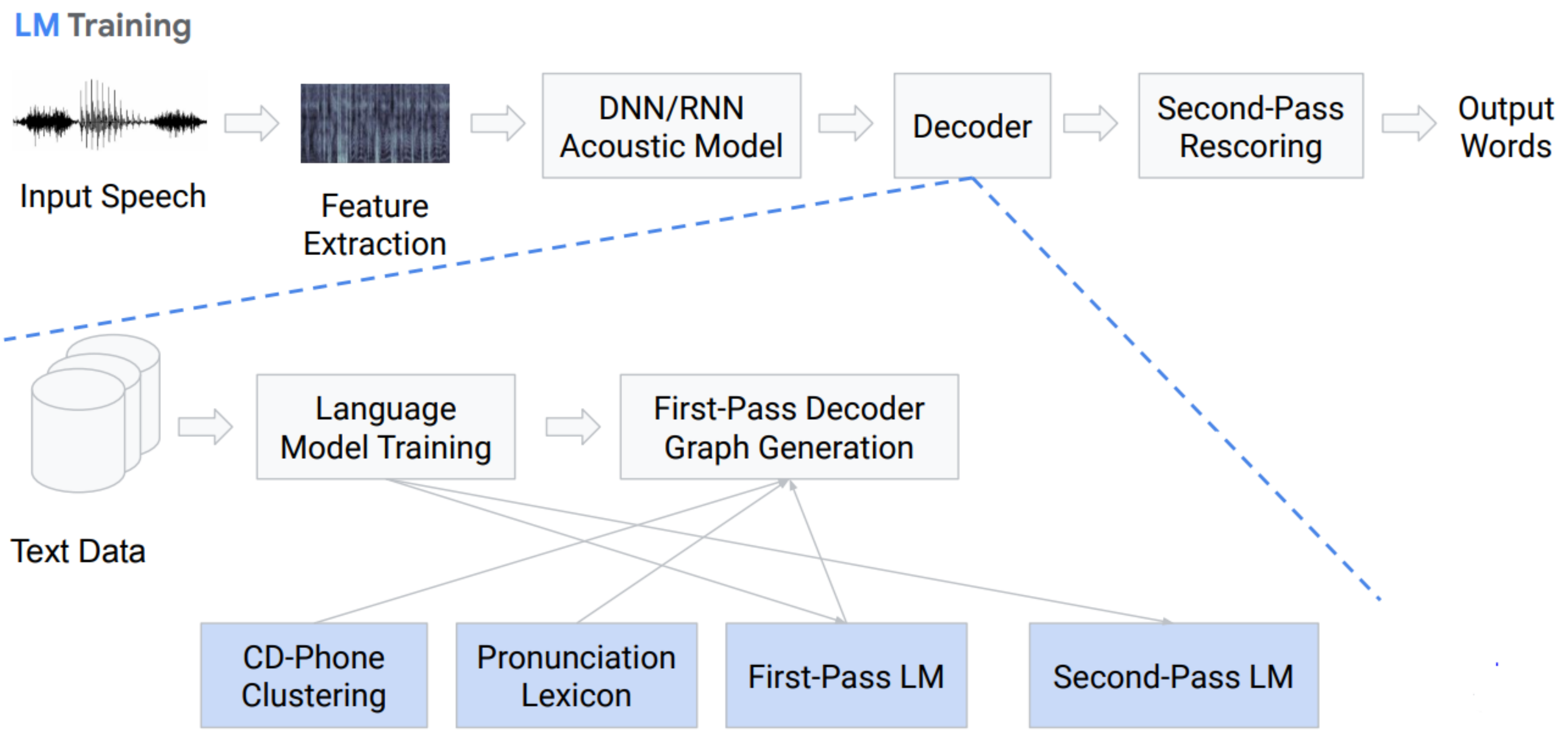
Pipeline



Классический подход



Классический подход. Первый шаг



Что же такое end2end ASR ?

Система, которая напрямую отображает последовательность входных акустических признаков в последовательность графем или слов

Система, которая обучена на оптимизации критериев, которые являются наиболее релевантными для итоговой метрики. (в задачах распознавания речи традиционно – WER)

Мотивация

Typical Speech System

Key Takeaway

A single end-to-end trained sequence-to-sequence model, which directly outputs words or graphemes, could greatly simplify the speech recognition pipeline.

CTC based models

Connectionist Temporal Classification (CTC)

- CTC was proposed by [Graves et al., 2006] as a way to train an acoustic model without requiring frame-level alignments
- Early work, used CTC with phoneme output targets - not “end-to-end”
- CD-phoneme based CTC models achieve state-of-the-art performance for conventional ASR, but word-level lagged behind [Sak et al., 2015]

Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Alex Graves[1] ALEX@IDSIA.CH
Santiago Fernández[1] SANTIAGO@IDSIA.CH
Faustino Gomez[1] TINO@IDSIA.CH
Jürgen Schmidhuber[1,2] JUERGEN@IDSIA.CH
[1] Istituto Dalle Molle di Studi sull’Intelligenza Artificiale (IDSIA), Galleria 2, 6928 Manno-Lugano, Switzerland
[2] Technische Universität München (TUM), Boltzmannstr. 3, 85748 Garching, Munich, Germany

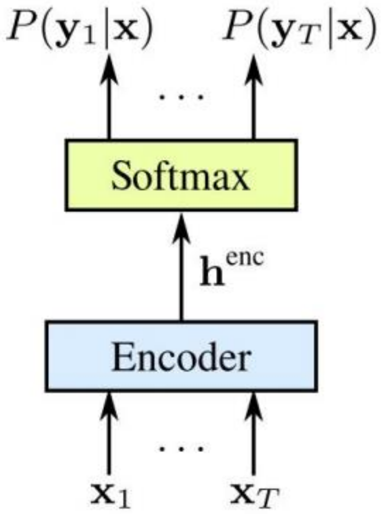
Abstract

Many real-world sequence learning tasks require the prediction of sequences of labels from noisy, unsegmented input data. In

belling. While these approaches have proved successful for many problems, they have several drawbacks: (1) they usually require a significant amount of task specific knowledge, e.g. to design the state models for HMMs, or choose the input features for CRFs; (2)

[Graves et al., 2006] ICML

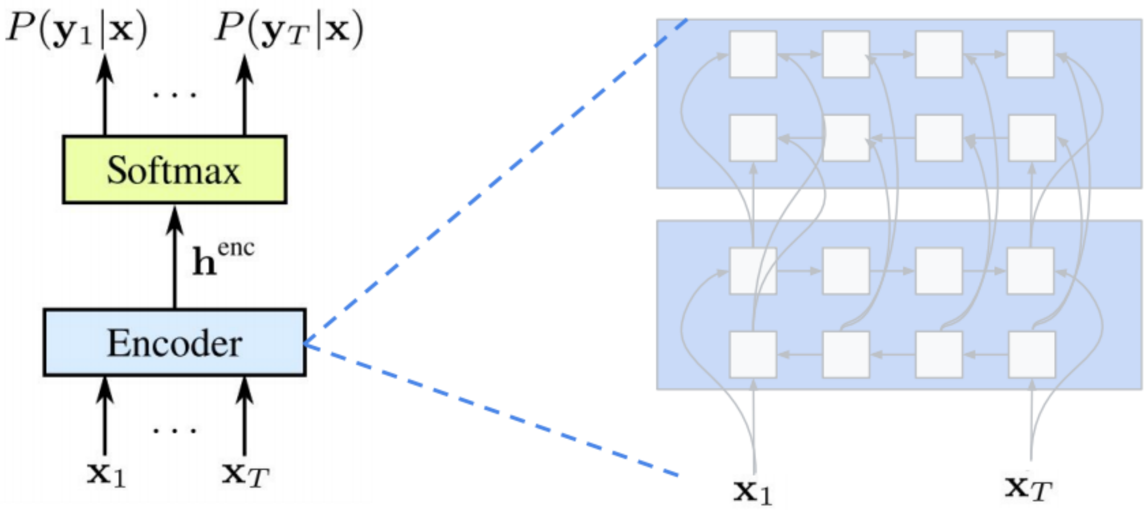
Connectionist Temporal Classification (CTC)



Key Takeaway

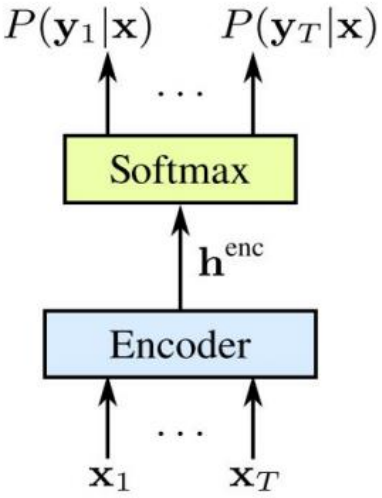
CTC allows for training an acoustic model without the need for frame-level alignments between the acoustics and the transcripts.

Connectionist Temporal Classification (CTC)



Key Takeaway	Encoder: Multiple layers of Uni- or Bi-directional RNNs (often LSTMs).

Connectionist Temporal Classification (CTC)



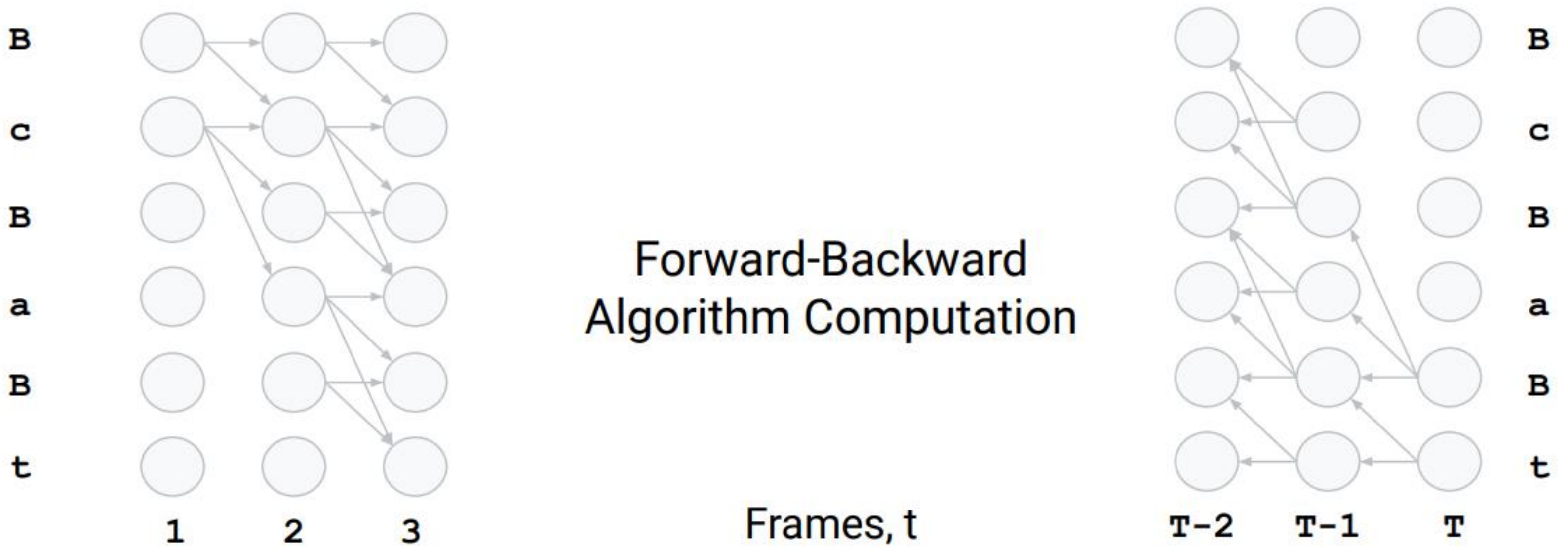
B	B	**c**	B	B	**a**	**a**	B	B	**t**
B	**c**	**c**	B	**a**	B	B	B	B	**t**
				...					
B	**c**	B	B	**a**	B	B	**t**	**t**	B

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\hat{\mathbf{y}} \in \mathcal{B}(\mathbf{y},\mathbf{x})} \prod_{t=1}^{T} P(\hat{y}_t|\mathbf{x})$$

Key Takeaway

CTC introduces a special symbol - blank (denoted by B) - and maximizes the total probability of the label sequence by marginalizing over all possible alignments

Connectionist Temporal Classification (CTC)



Key Takeaway

Computing the gradients of the loss requires the computation of the alpha-beta variables using the forward-backward algorithm [Rabiner, 1989]

CTC-based End-to-End ASR

- Graves and Jaitly proposed a system with character-based CTC which directly output word sequences given input speech
- Using an external LM was important for getting good performance. Results reported by rescoring a baseline system.
- Also proposed minimizing expected transcription error [WSJ: 8.7% → 8.2%]

Towards End-to-End Speech Recognition with Recurrent Neural Networks

Alex Graves GRAVES@CS.TORONTO.EDU
Google DeepMind, London, United Kingdom

Navdeep Jaitly NDJAITLY@CS.TORONTO.EDU
Department of Computer Science, University of Toronto, Canada

Abstract

This paper presents a speech recognition system that directly transcribes audio data with text, without requiring an intermediate phonetic representation. The system is based on a combination

fits of holistic optimisation tend to outweigh those of prior knowledge.

While automatic speech recognition has greatly benefited from the introduction of neural networks (Bourlard & Morgan, 1993; Hinton et al., 2012), the networks are at present

[Graves and Jaitly, 2014] ICML

CTC-based ASR

Refinements since [Graves & Jaitly, 2014]

- LM incorporated into first-pass decoding; easy integration with WFSTs
  - [Hannun et al., 2014] [Maas et al., 2015]: Direct first-pass decoding with an LM as opposed to rescoring as in [Graves & Jaitly, 2014]
  - [Miao et al., 2015]: EESEN framework for decoding with WFSTs, open source toolkit

- Large-scale GPU training; data augmentation; multiple languages
  - [Hannun et al., 2014; DeepSpeech] [Amodei et al., 2015; DeepSpeech2]: Large scale GPU training; Data Augmentation; Mandarin and English

- Using longer span units: words instead of characters
  - [Soltau et al., 2017]: Word-level CTC targets, trained on 125,000 hours of speech. Performance close to or better than a conventional system, even without using an LM!
  - [Audhkhasi et al., 2017]: Direct Acoustics-to-Word Models on Switchboard
- And many others ...

Недостатки CTC

- Для эффективности CTC делает важное предположение о независимости - выходные данные сети в разных кадрах условно независимы
- Для достижения хорошей производительности требуется использование внешней языковой - прямое жадное декодирование работает неудовлетворительно

Attention based models

Attention-based End-to-End ASR

- Attention-based Encoder-Decoder Models emerged first in the context of neural machine translation.
- Were first applied to ASR by [Chan et al., 2015] [Chorowski et al., 2015]

Listen, Attend and Spell

William Chan
Carnegie Mellon University
williamchan@cmu.edu

Navdeep Jaitly, Quoc V. Le, Oriol Vinyals
Google Brain
{ndjaitly,qvl,vinyals}@google.com

[Chan et al., 2015]

Attention-Based Models for Speech Recognition

Jan Chorowski
University of Wrocław, Poland
jan.chorowski@ii.uni.wroc.pl

Dzmitry Bahdanau
Jacobs University Bremen, Germany

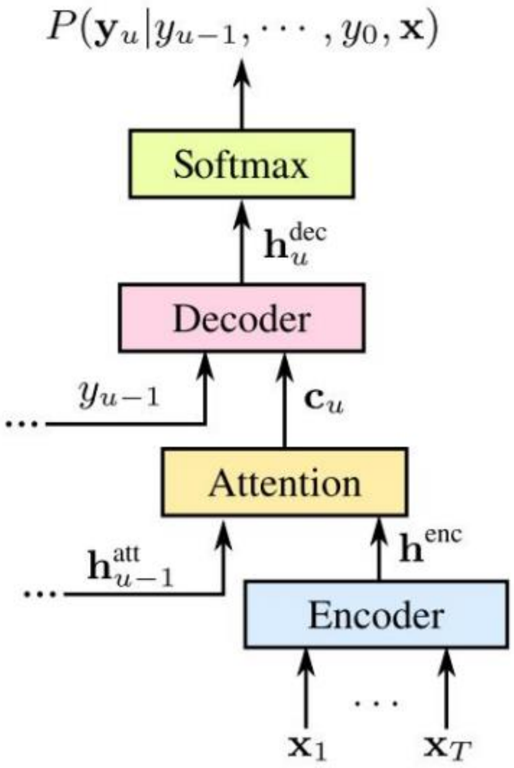
Dmitriy Serdyuk
Université de Montréal

Kyunghyun Cho
Université de Montréal

Yoshua Bengio
Université de Montréal
CIFAR Senior Fellow

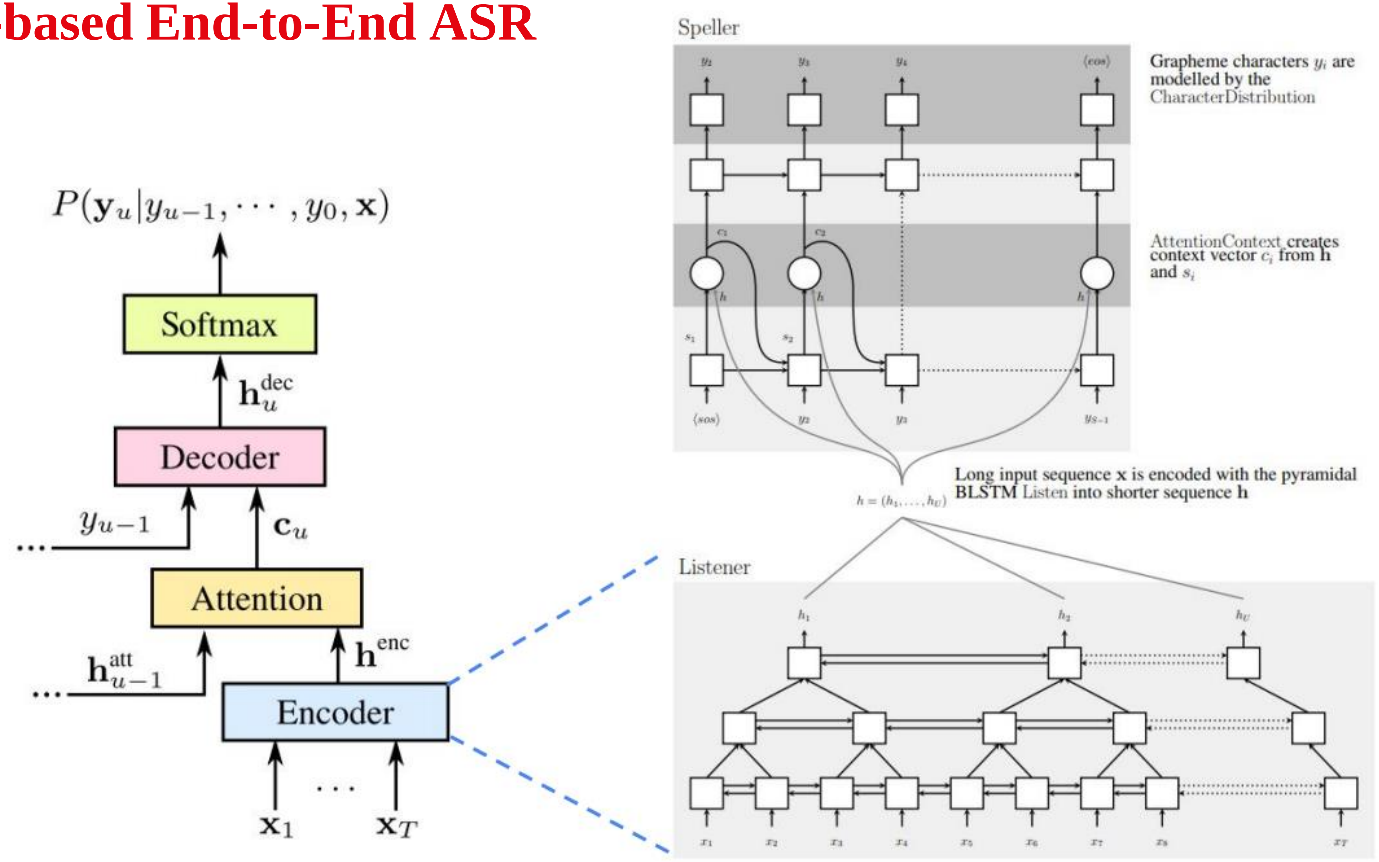
[Chorowski et al., 2015]

Attention-based End-to-End ASR



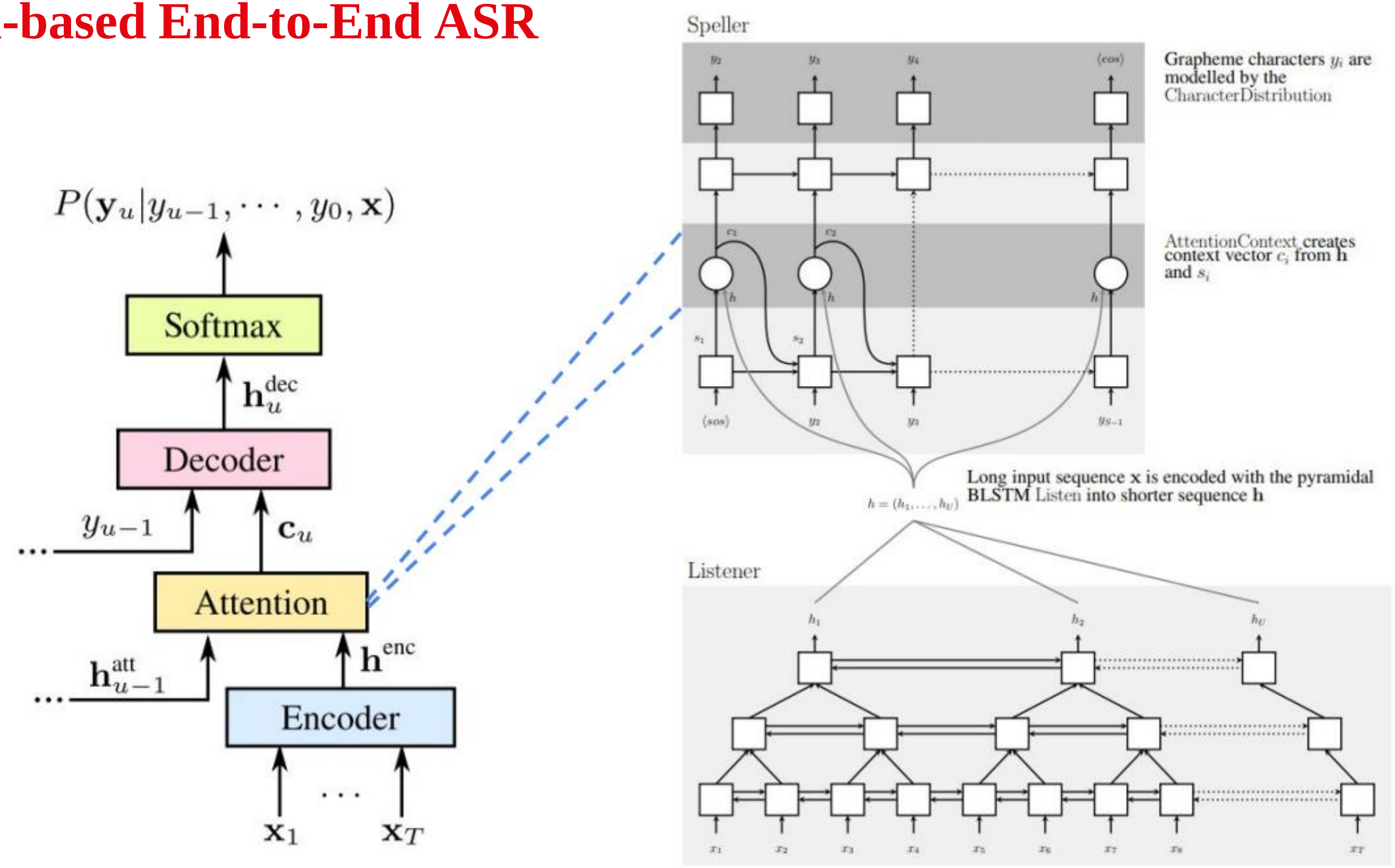
- Encoder (analogous to AM):
  - Transforms input speech into higher-level representation
- Attention (alignment model):
  - Identifies encoded frames that are relevant to producing current output
- Decoder (analogous to PM, LM):
  - Operates autoregressively by predicting each output token as a function of the previous predictions

Attention-based End-to-End ASR



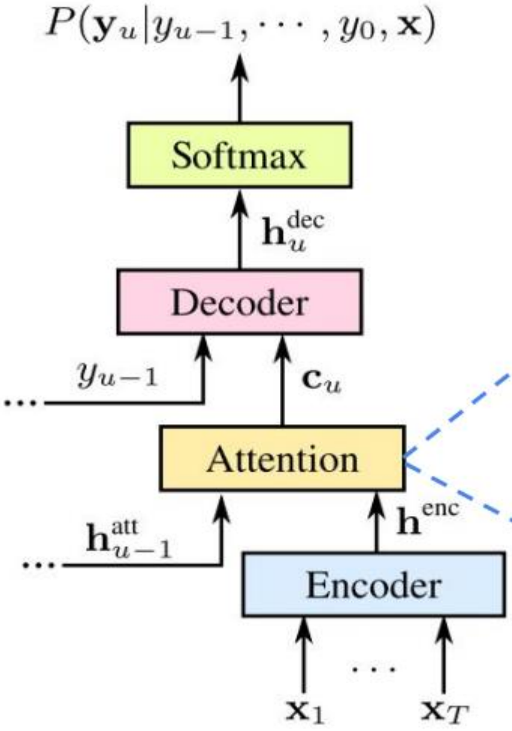
Reproduced from [Chan et al., 2015]

Attention-based End-to-End ASR



Reproduced from [Chan et al., 2015]

Attention-based End-to-End ASR



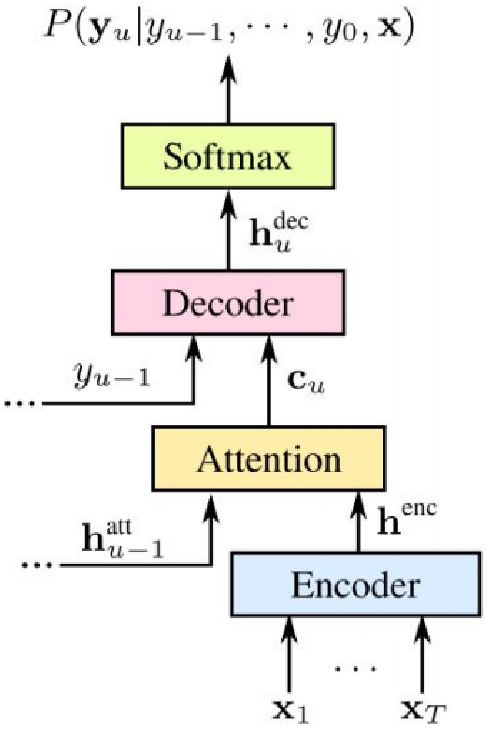
Attention module computes a similarity score between the decoder and each frame of the encoder

$$e_{u,t} = \text{score}(\mathbf{h}_{u-1}^{\text{att}}, \mathbf{h}_t^{\text{enc}})$$

$$\alpha_{u,t} = \frac{\exp(e_{u,t})}{\sum_{t'=1}^{T} \exp(e_{u,t'})}$$

$$\mathbf{c}_u = \sum_{t=1}^{T} \alpha_{u,t} h_t^{\text{enc}}$$

Attention-based End-to-End ASR



Dot-Product Attention [Chan et al., 2015]

$$e_{u,t} = \left\langle \phi(W\mathbf{h}_{u-1}^{\text{att}}),\ \psi(V\mathbf{h}_t^{\text{enc}}) \right\rangle$$

Additive Attention [Chorowski et al., 2015]

$$e_{u,t} = w^T \tanh(W\mathbf{h}_{u-1}^{\text{att}} + V\mathbf{h}_t^{\text{enc}} + b)$$

Возможные улучшения

Модели на кусочках слов

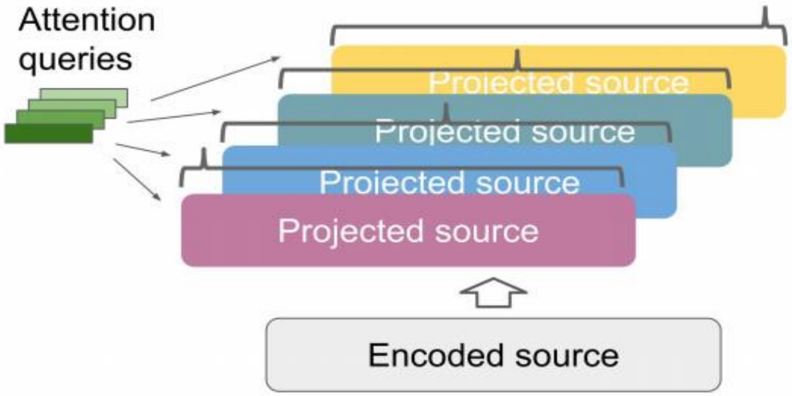
- Instead of the commonly used grapheme, we can use longer units such as wordpieces
- Motivations:
  - Typically, word-level LMs have a much lower perplexity compared to grapheme-level LMs [Kannan et al., 2018]
  - Modeling wordpiece allows for a much stronger decoder LM
  - Modeling longer units improves the effective memory of the decoder LSTMs
  - Allows the model to potentially memorize pronunciations for frequently occurring words
  - longer units require fewer decoding steps; this speeds up inference in these models significantly
- good performance for LAS and RNN-T [Rao et al., 2017].

Модели на кусочках слов

- sub-word units, ranging from graphemes all the way up to entire words.
- there are no out-of-vocabulary words with word piece models
- The word piece models are trained to maximize the language model likelihood over the training set
- the word pieces are “position-dependent”, in that a special word separator marker is used to denote word boundaries.
- Words are segmented deterministically and independent of context, using a greedy algorithm.

Multi-headed Attention

- Multi-head attention (MHA) was first explored in [Vaswani et al., 2017] for machine translation
- MHA extends the conventional attention mechanism to have multiple heads, where each head can generate a different attention distribution.

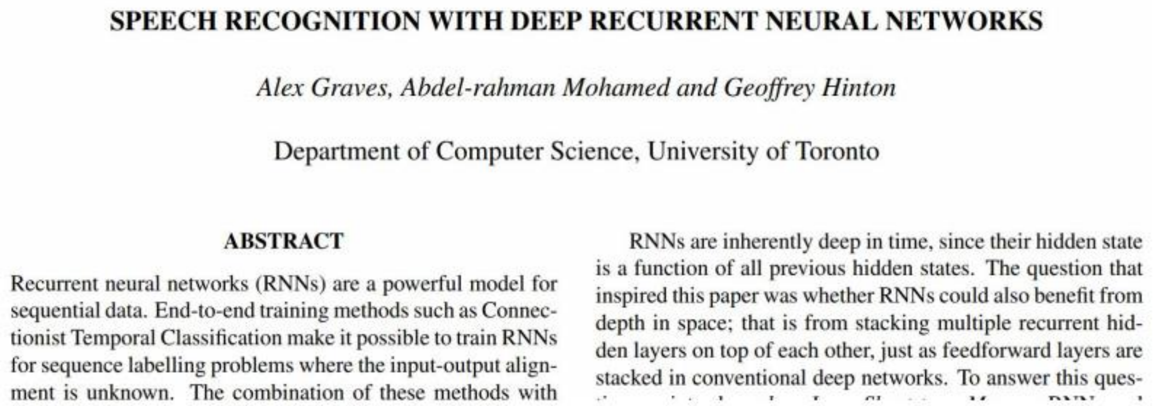


Online Models

- LAS is not streaming
- We will show a thorough comparison of different online models
  - RNN-T [Graves, 2012], [Rao et al., 2017], [He et al., 2018]
  - Neural Transducer [Jaitly et al., 2015], [Sainath et al., 2018]
  - MoChA [Chiu and Raffel, 2018]

Recurrent Neural Network Transducer (RNN-T)

- Proposed by Graves et al., RNN-T augments a CTC-based model with a recurrent LM component
- Both components are trained jointly on the available acoustic data
- As with CTC, the method does not require aligned training data.



SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS

Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton

Department of Computer Science, University of Toronto

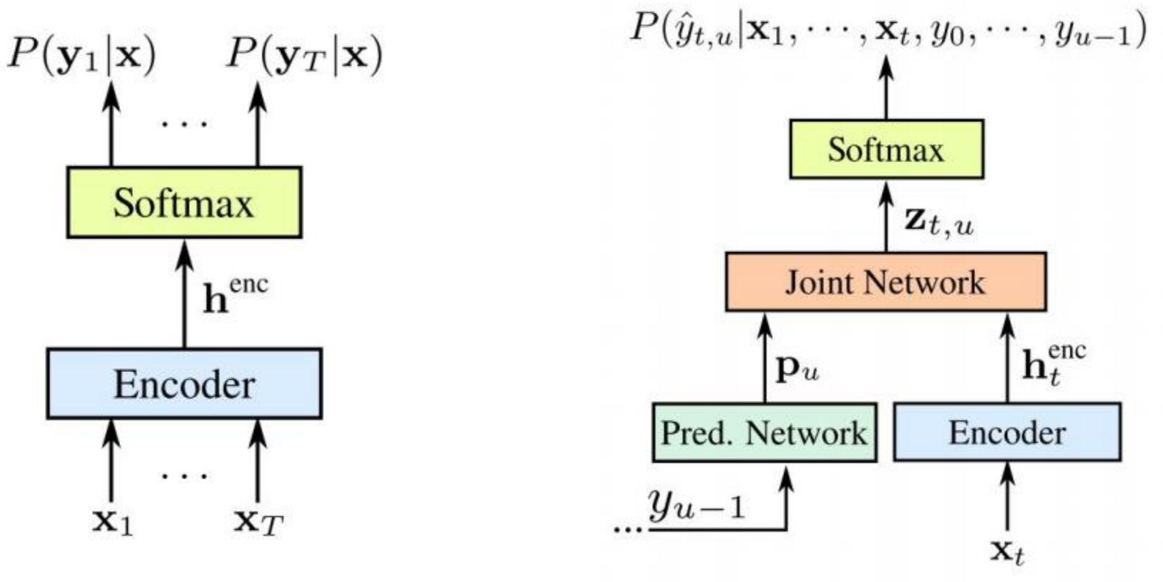
ABSTRACT

Recurrent neural networks (RNNs) are a powerful model for sequential data. End-to-end training methods such as Connectionist Temporal Classification make it possible to train RNNs for sequence labelling problems where the input-output alignment is unknown. The combination of these methods with

RNNs are inherently deep in time, since their hidden state is a function of all previous hidden states. The question that inspired this paper was whether RNNs could also benefit from depth in space; that is from stacking multiple recurrent hidden layers on top of each other, just as feedforward layers are stacked in conventional deep networks. To answer this ques-

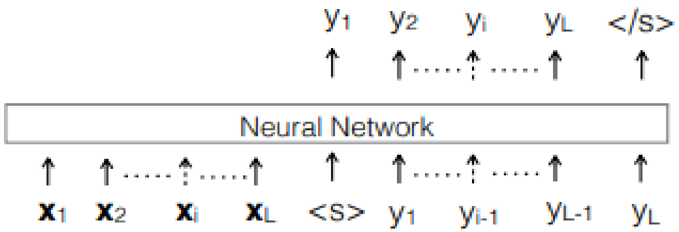
[Graves et al., 2013] ICASSP;
[Graves, 2012] ICML Representation Learning Workshop

Recurrent Neural Network Transducer (RNN-T)

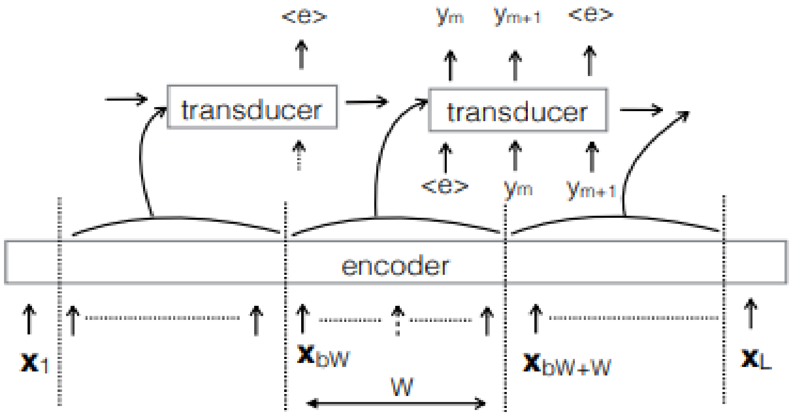


RNN-T [Graves, 2012] augments CTC encoder with a recurrent neural network LM

Neural Transducer

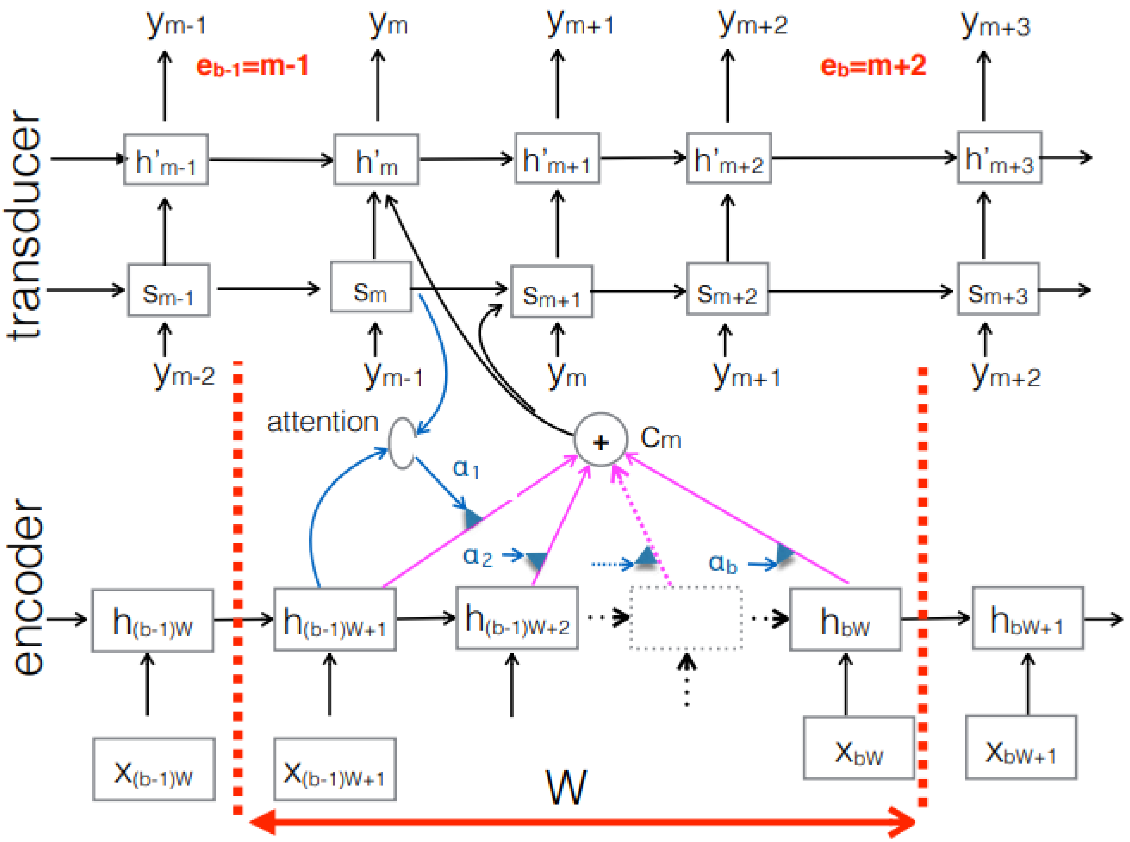


(a) seq2seq

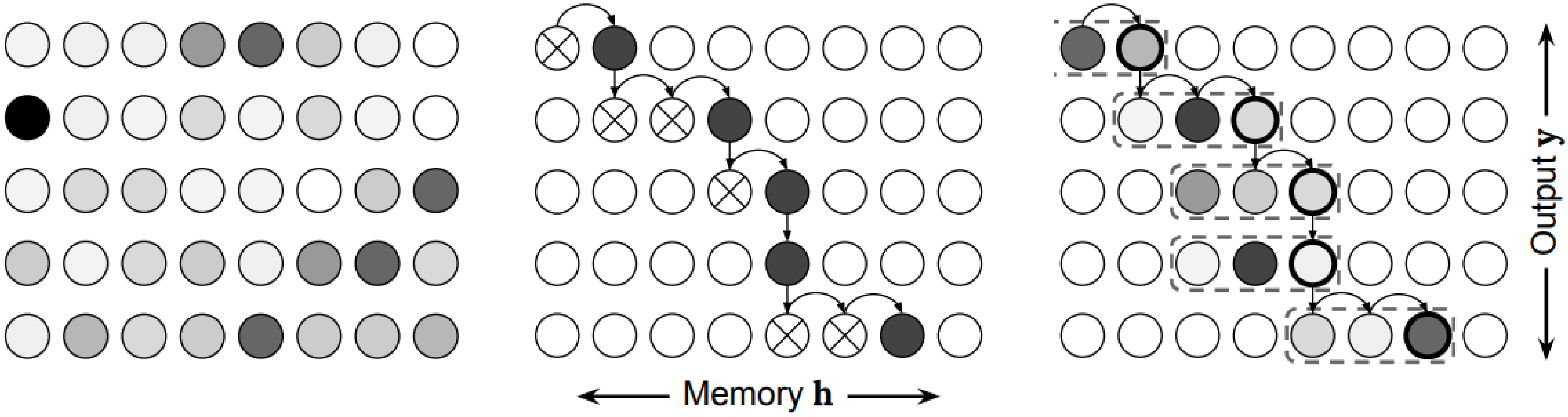


(b) Neural Transducer

Neural Transducer



Monotonic Chunkwise Attention (MoChA)



(a) Soft attention. (b) Hard monotonic attention. (c) Monotonic chunkwise attention.

Q&A

Спасибо!

Публичное акционерное общество
«Мобильные ТелеСистемы»

Ул. Марксистская, 4,
Москва, Россия, 109147
Тел.: +7 (495) 911-71-51
Факс: +7 (495) 911-65-69

www.corp.mts.ru

МТС

Ведите бизнес вперёд