Speech and Language Processing

Lecture 4
Neural network based speech recognition and synthesis

Information and Communications Engineering Course

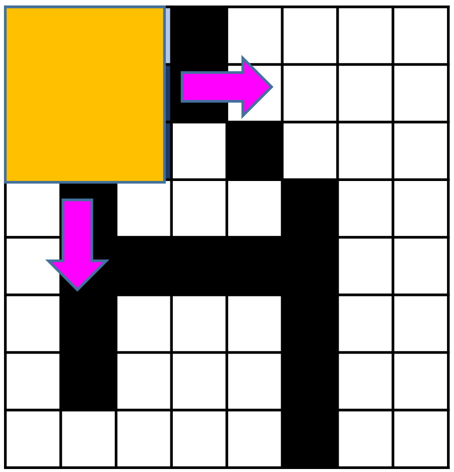
Takahiro Shinozaki

Manabu Okumura

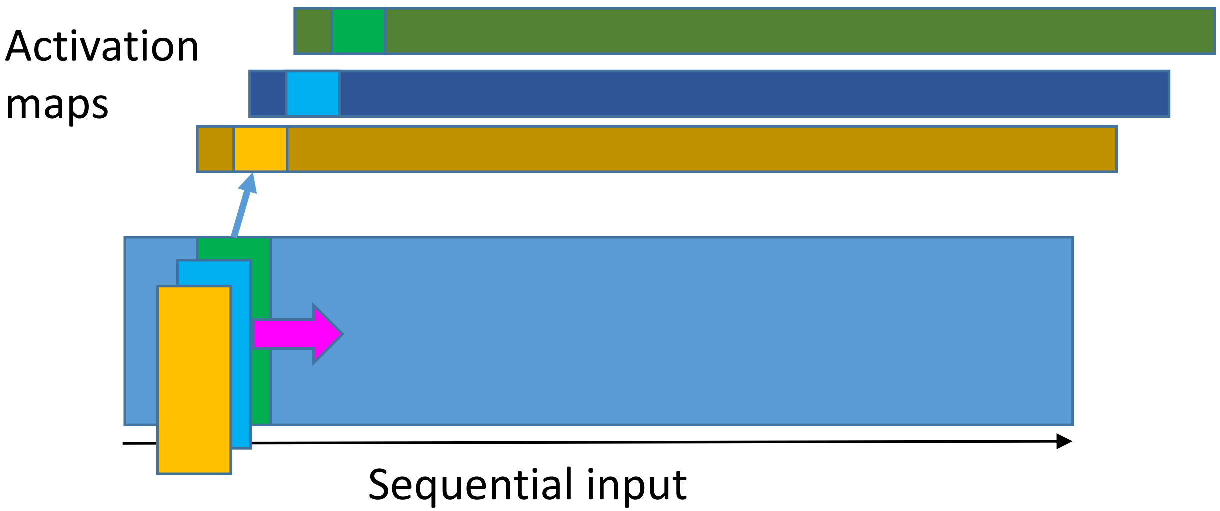
2024/10/1

Some Basic Functional Elements

1D-CNN



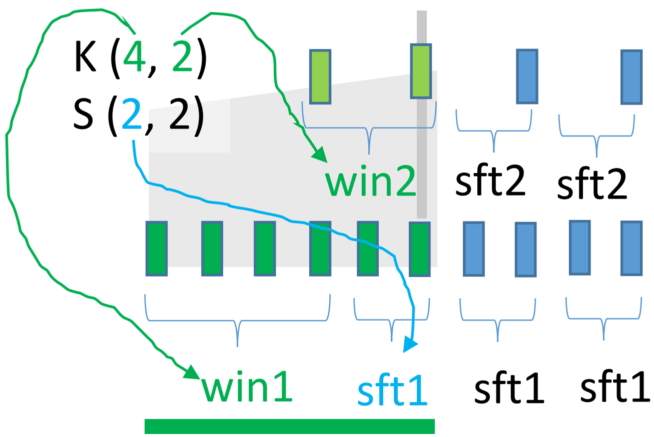
2D-CNN



1D-CNN

Receptive Field Length of Cascaded 1D Convolution

1st convolution: Kernel (window) Size = 4, Stride (shift) = 2.
2nd convolution: Kernel (window) Size = 2, Stride (shift) = 2.



win1 + (win2-1)*sft1
=4+
(2-1)*2

K (10, 8, 4, 4, 4), S (5, 4, 2, 2, 2)

((((10 ← Window width of 1st layer @ input sample rate

+(8-1)*5) ← Window width of 2st layer @ input sample rate

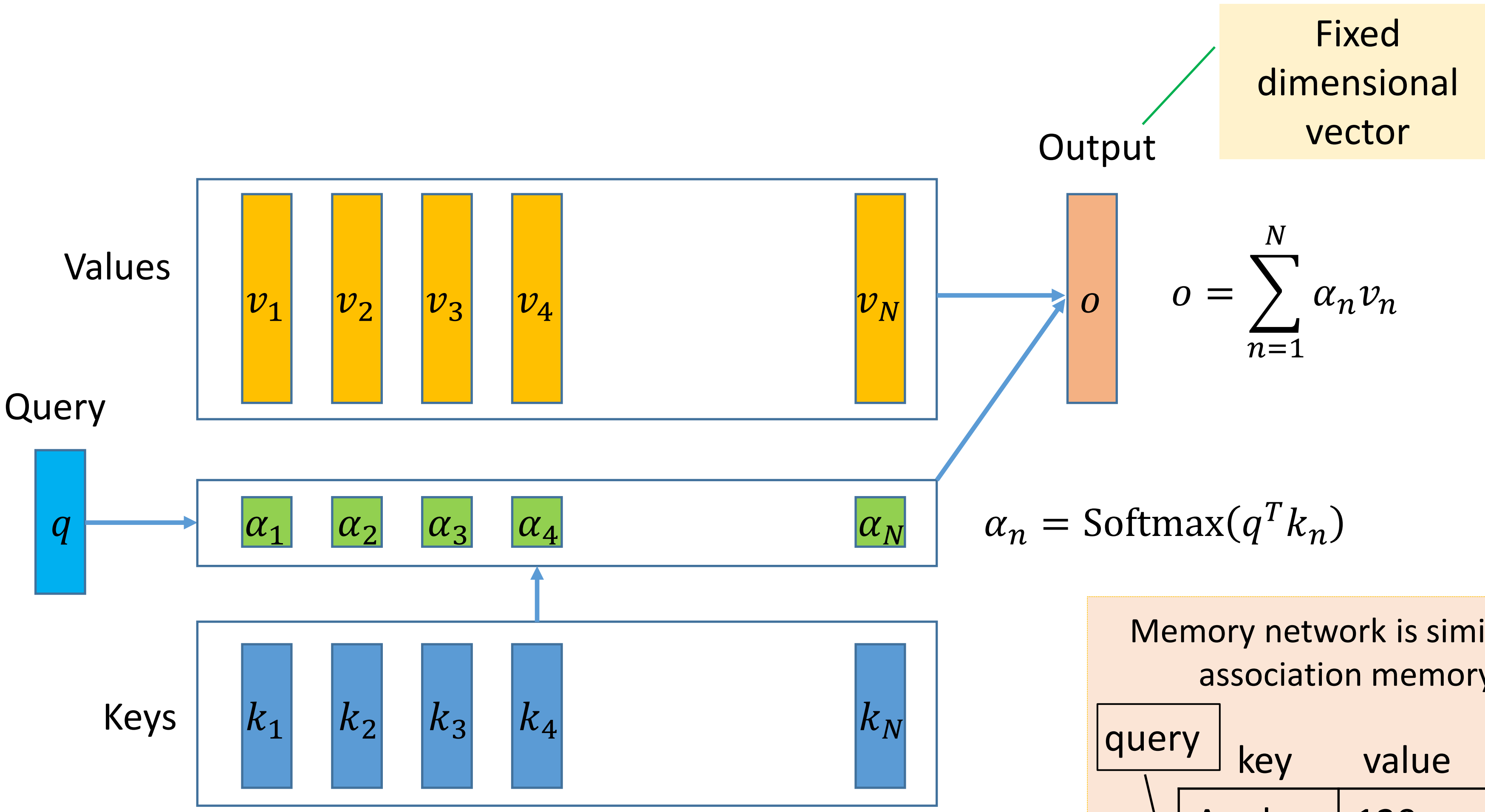
+(4-1)*(5*4)) ← Window width of 3st layer @ input sample rate

+(4-1)*(5*4*2)) ← Window width of 4st layer @ input sample rate

+(4-1)*(5*4*2*2)) ← Window width of 5st layer @ input sample rate

=465

Memory Network



$$o = \sum_{n=1}^{N} \alpha_n v_n$$

$$\alpha_n = \text{Softmax}(q^T k_n)$$

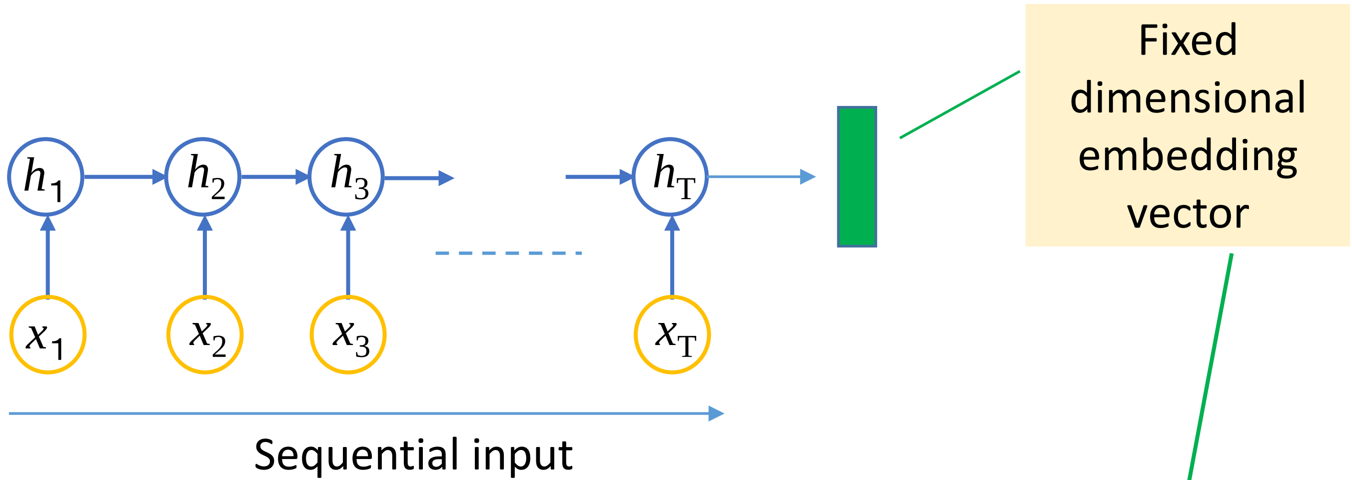
Memory network is similar to association memory

query

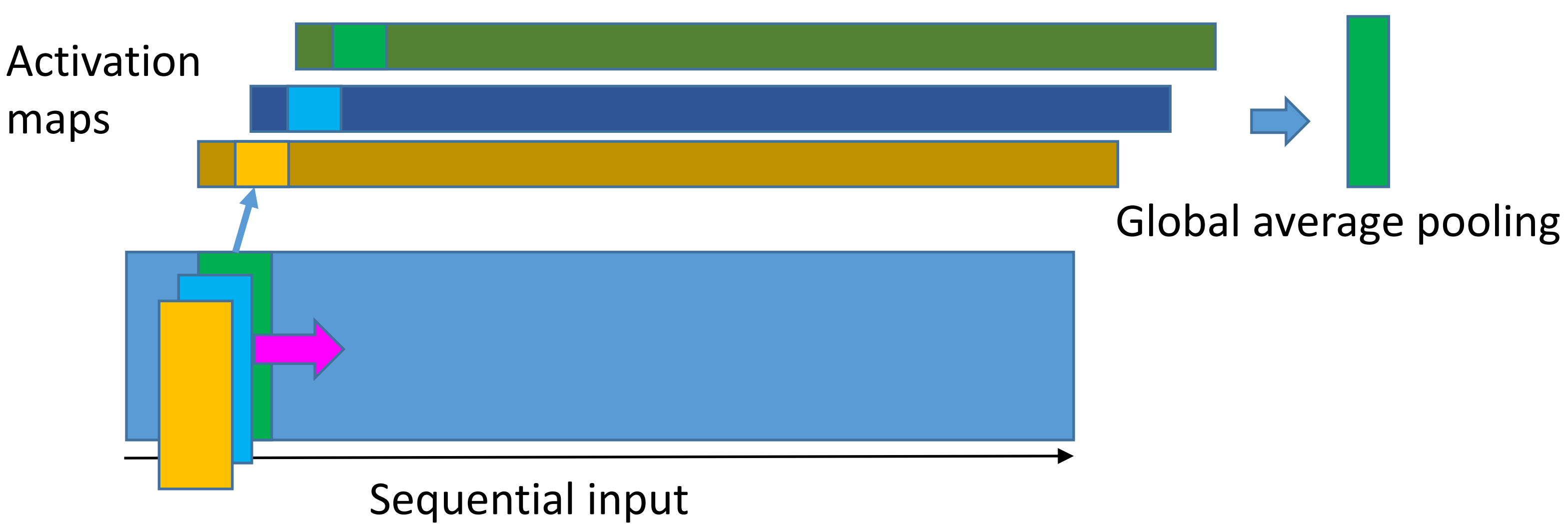
key	value
Apple	120
Banana	100
Durian	5500

Fixed-Dimensional Embeddings of Sequences

- Use RNN



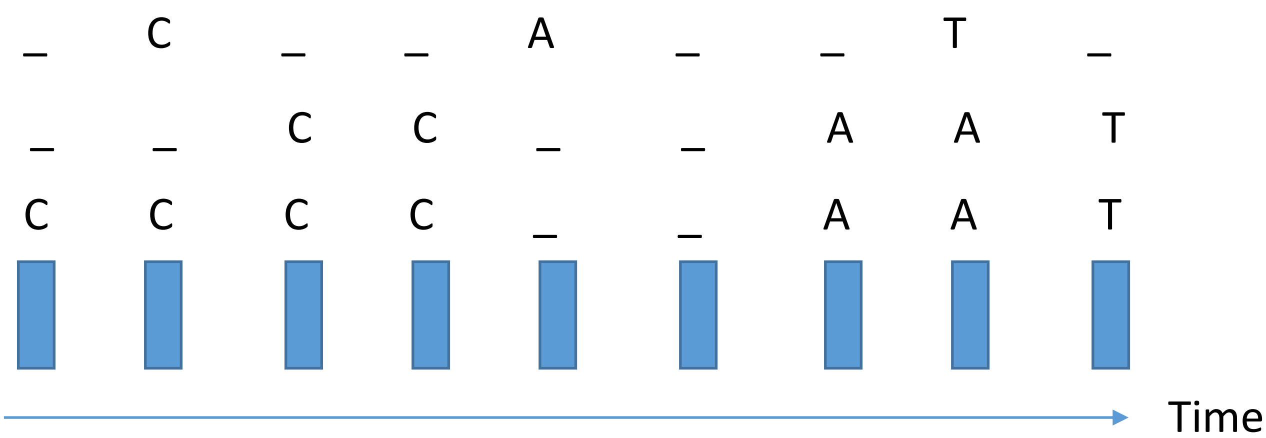
- 1D CNN with Global average pooling



Connectionist Temporal Classification (CTC) Loss

- Assume:
  - We have frame-wise character (or word etc.) prediction for a sequence of time frames
  - A blank label is included as a special character
  - Reference text is a sequence of characters whose length is smaller than the frame sequence
- Matching of the prediction and the reference
  - Form output by collapsing repeated characters and removing blank character from the predicted sequence

Reference: C A T



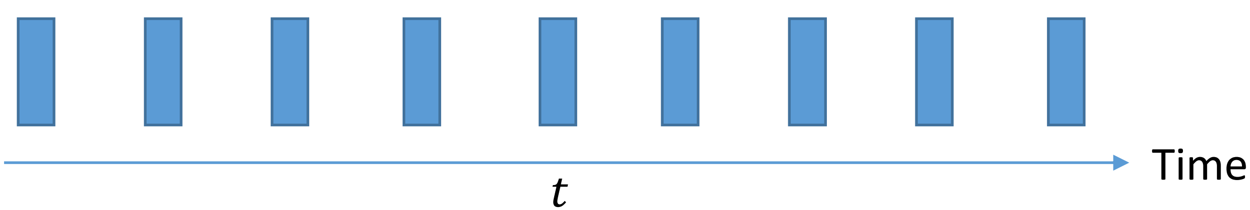
A. Graves+, “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks,” ICML, 2006https://www.cs.toronto.edu/~graves/icml_2006.pdf

Probability of Predicted Sequence

- Probability of Predicted Sequence is a product of frame-wise prediction probabilities

$$P(\_\ \_\ \mathrm{C\,C}\,\_\,\_\,\mathrm{A\,A\,T}) = P_1(\_)P_2(\_)P_3(C)P_4(C)P_5(\_)P_6(\_)P_7(A)P_8(A)P_9(T)$$

–	0.1	0.15			$P_t(\_)$				
A	0.02	0.03			$P_t(A)$				
C	0.05				$P_t(C)$				
T	0.01				$P_t(T)$				
⋮									



Probability of Character Sequence

- Probability of character sequence (like the reference) is a sum of probabilities of all the matching predicted sequences

$$P(CAT) = P(\_\ \ \_\ \mathrm{C\,C}\,\_\,\_\,\mathrm{A\,A\,T}) + P(\mathrm{C}\ \ \mathrm{C}\ \mathrm{C\,C}\,\_\,\_\,\mathrm{A\,A\,T}) \cdots$$

$$= \sum_{\pi \in \mathcal{B}^{-1}(CAT)} P(\pi) = \sum_{\pi \in \mathcal{B}^{-1}(CAT)} \prod_{t} P_t(\pi_t)$$

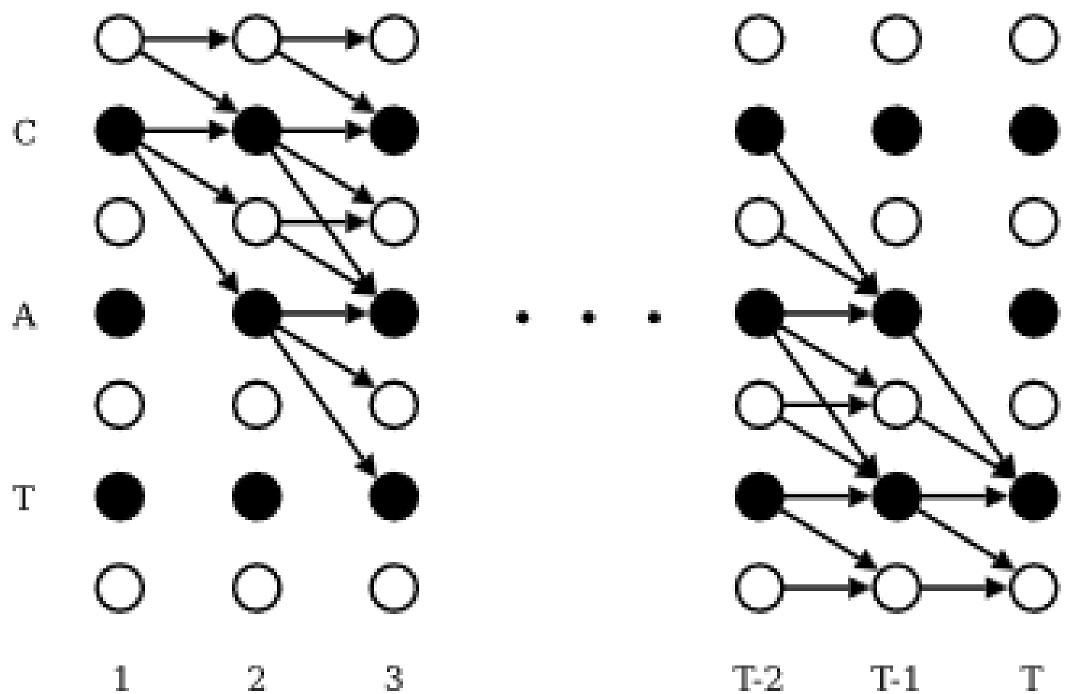
$\mathcal{B}$: The contraction function of CTC.

e.g. $\mathcal{B}(\_\ \ \_\ \mathrm{C\,C}\,\_\,\_\,\mathrm{A\,A\,T}) = \mathcal{B}(\mathrm{C}\ \ \mathrm{C}\ \mathrm{C\,C}\,\_\,\_\,\mathrm{A\,A\,T}) = CAT$

$\mathcal{B}^{-1}$: Inverse of the contraction function (one-to-many mapping)

Efficient Probability Evaluation

The probability of the character sequence is efficiently evaluated by the forward algorithm



Black circles represent labels, and white circles represent blanks

Figure is cited from A. Graves+ 2006

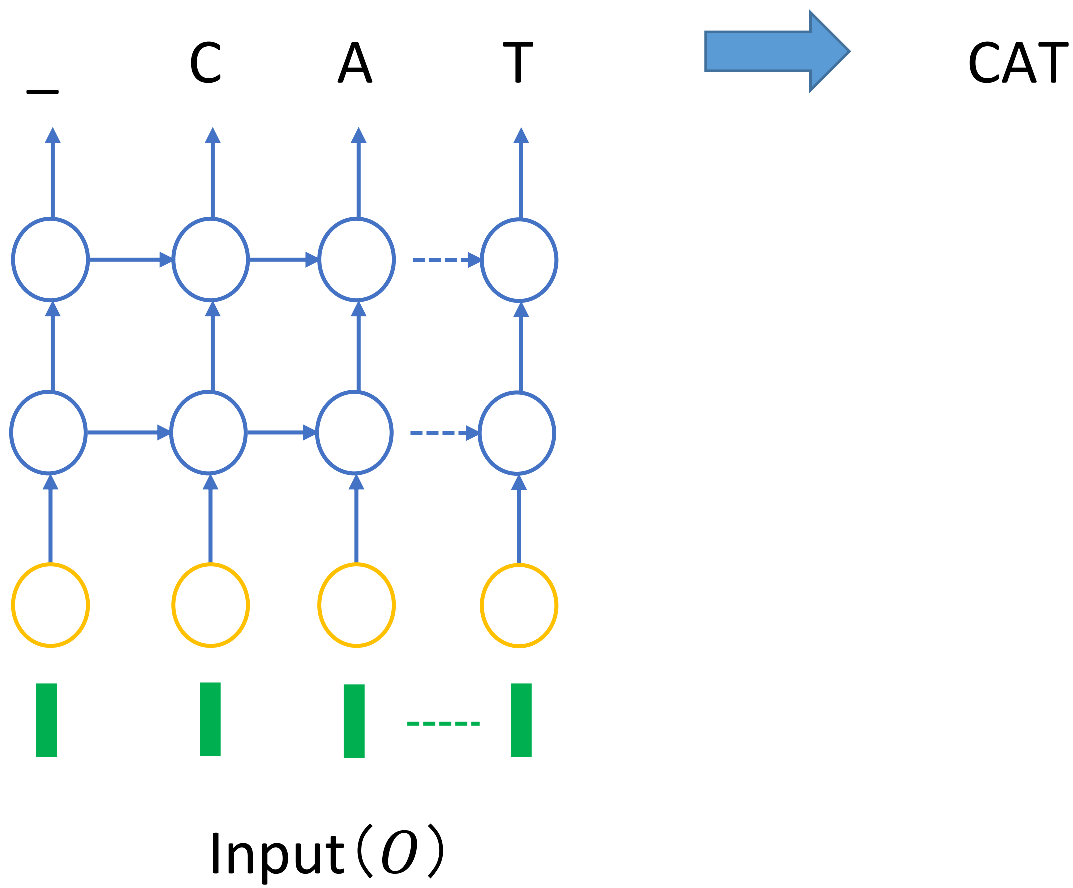
CTC Loss

CTC loss $$\mathcal{L}(S) = -logP(S) = -log \sum_{\pi \in \mathcal{B}^{-1}(S)} \prod_{t} P_t(\pi_t)$$

It's gradient $$\frac{\partial}{\partial P_u(c)} \mathcal{L}(S) = -\frac{\sum_{\pi \in \mathcal{B}^{-1}(S), \pi_u = c} \prod_{t \neq u} P_t(\pi_t)}{\sum_{\pi \in \mathcal{B}^{-1}(S)} \prod_t P_t(\pi_t)}$$

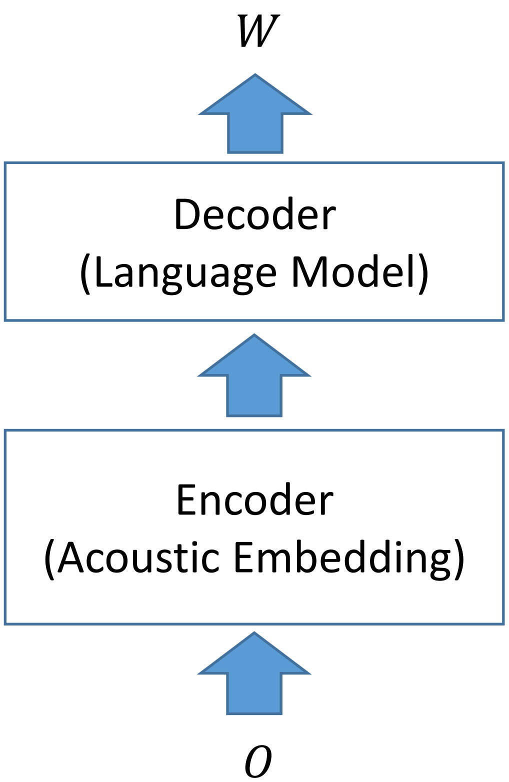
Neural Network Based Speech Recognition

RNN+CTC



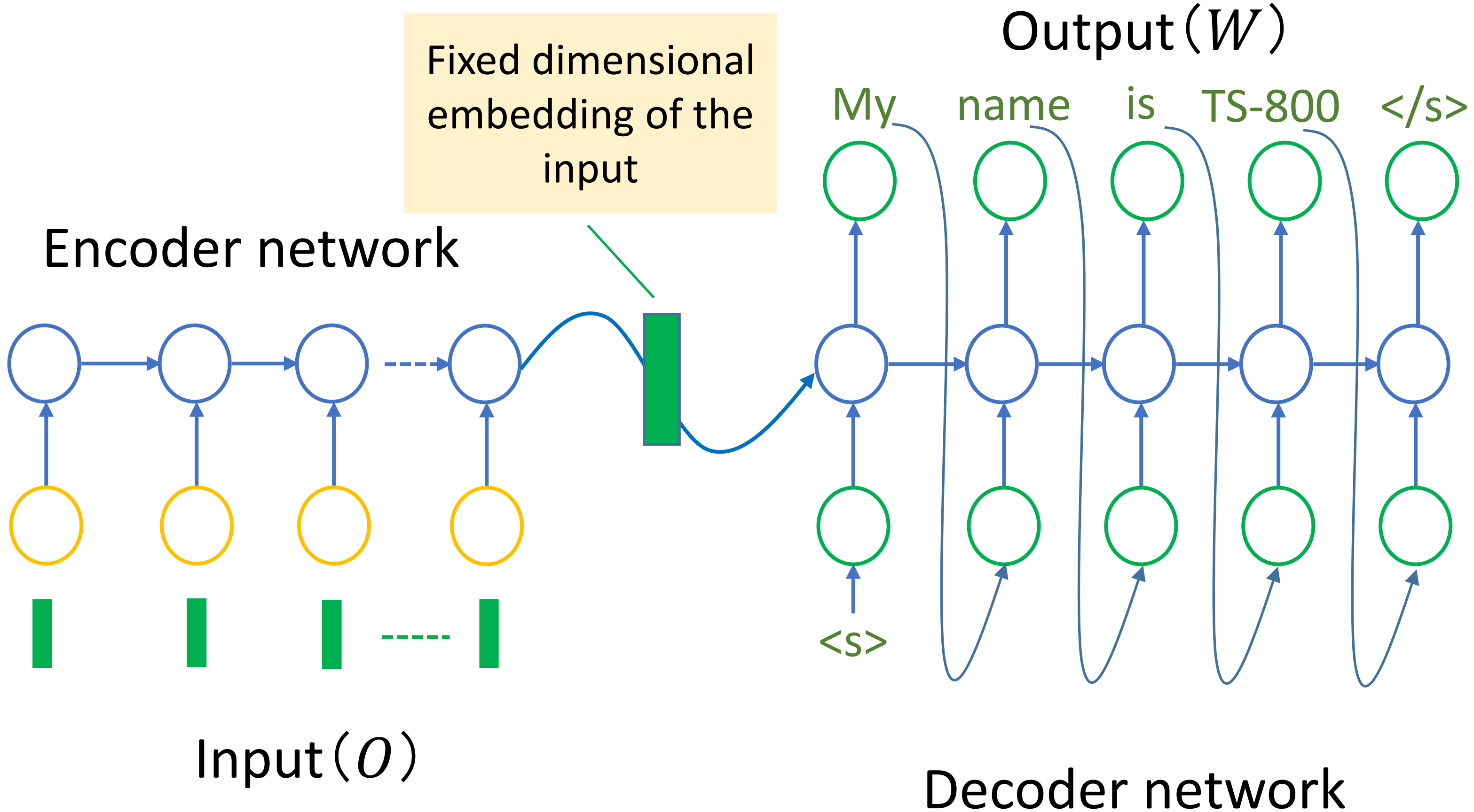
Encoder-Decoder Networks

Language model models a probability of a sentence $P(W)$.
By conditioning it with an acoustic input O, we get the discriminative modeling based speech recognizer $P(W|O)$.

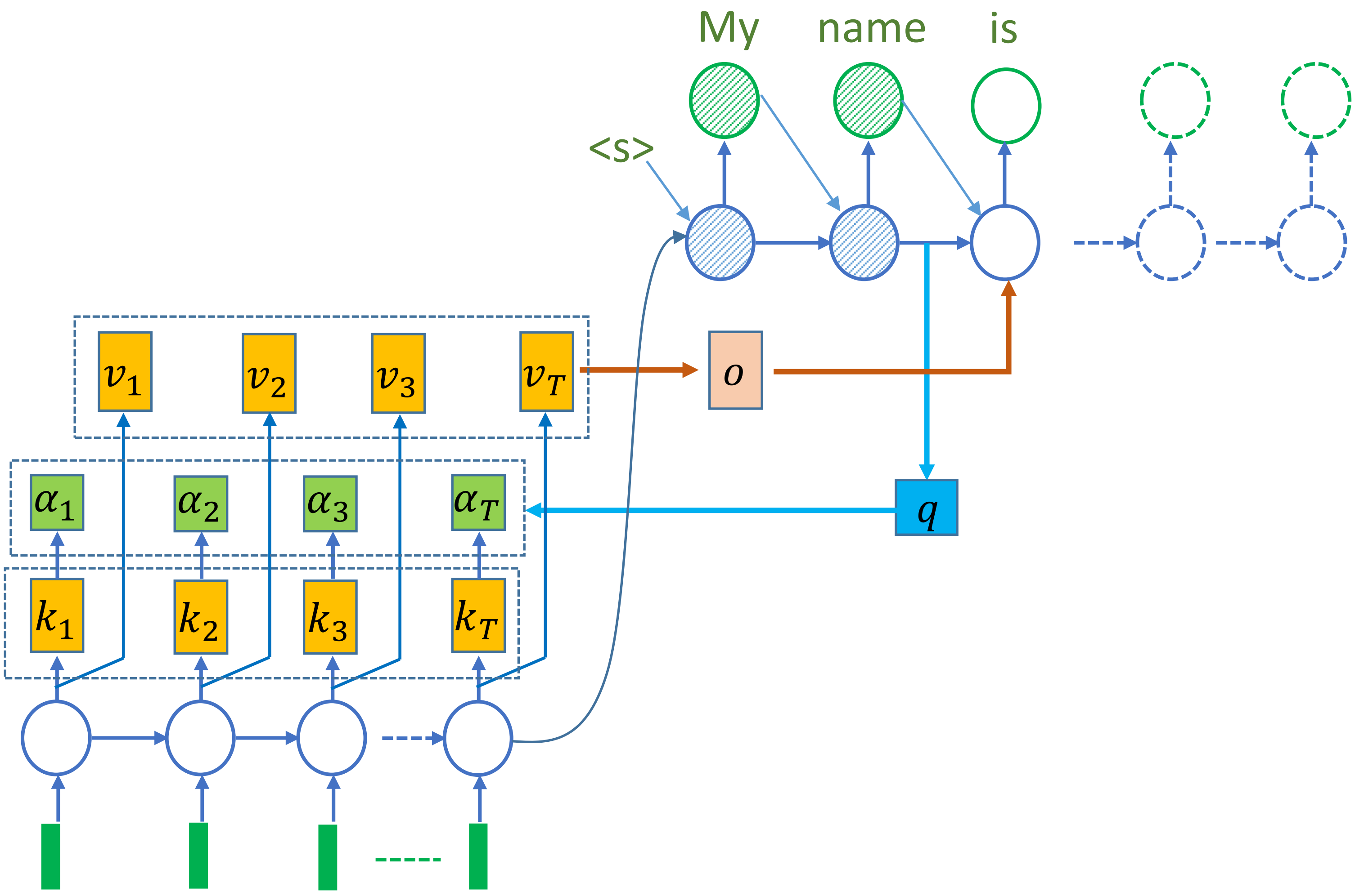


Encoder-Decoder Network

Directly models $P(W|O)$



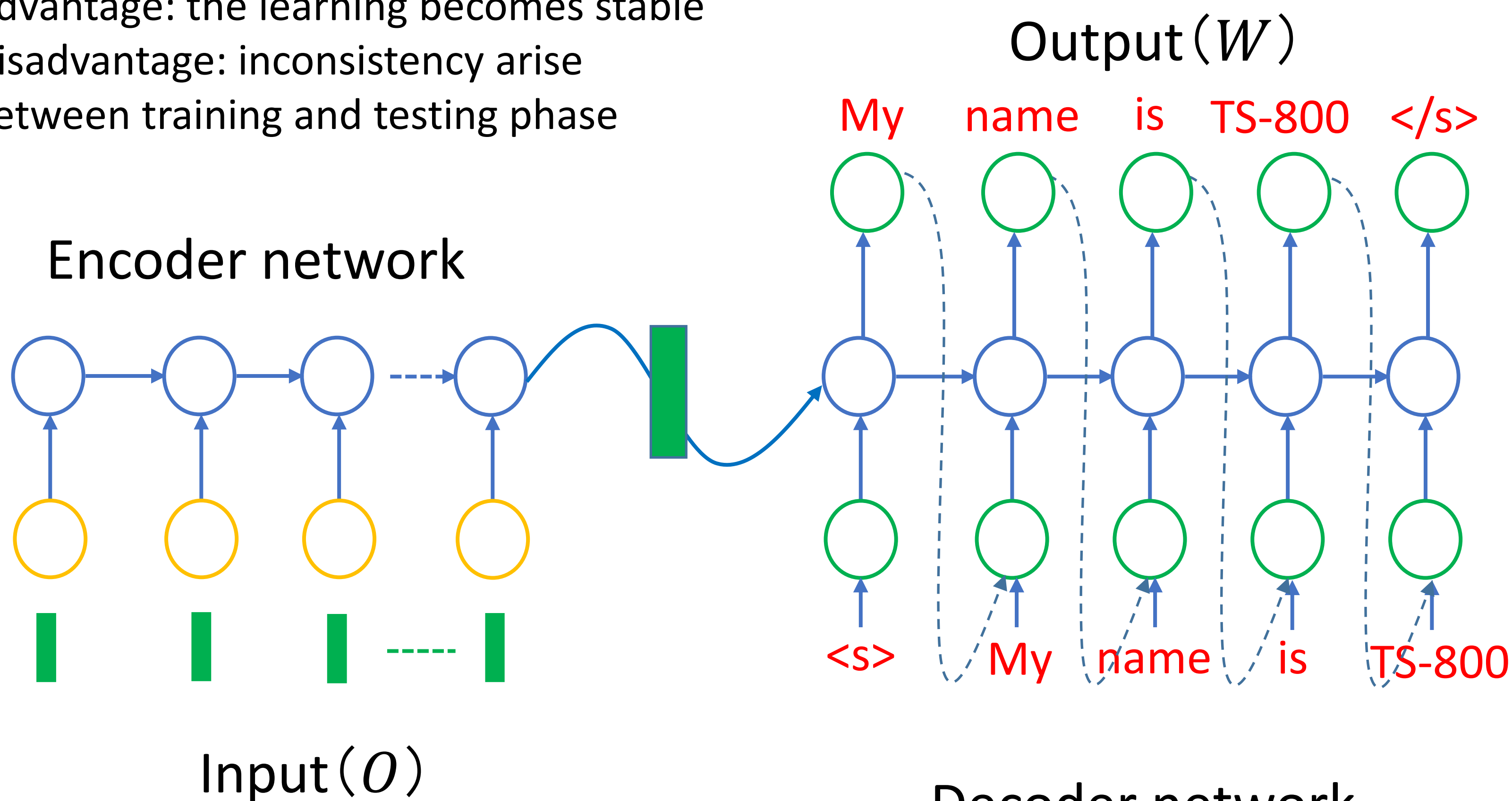
Attention Encoder-Decoder



Teacher Forcing

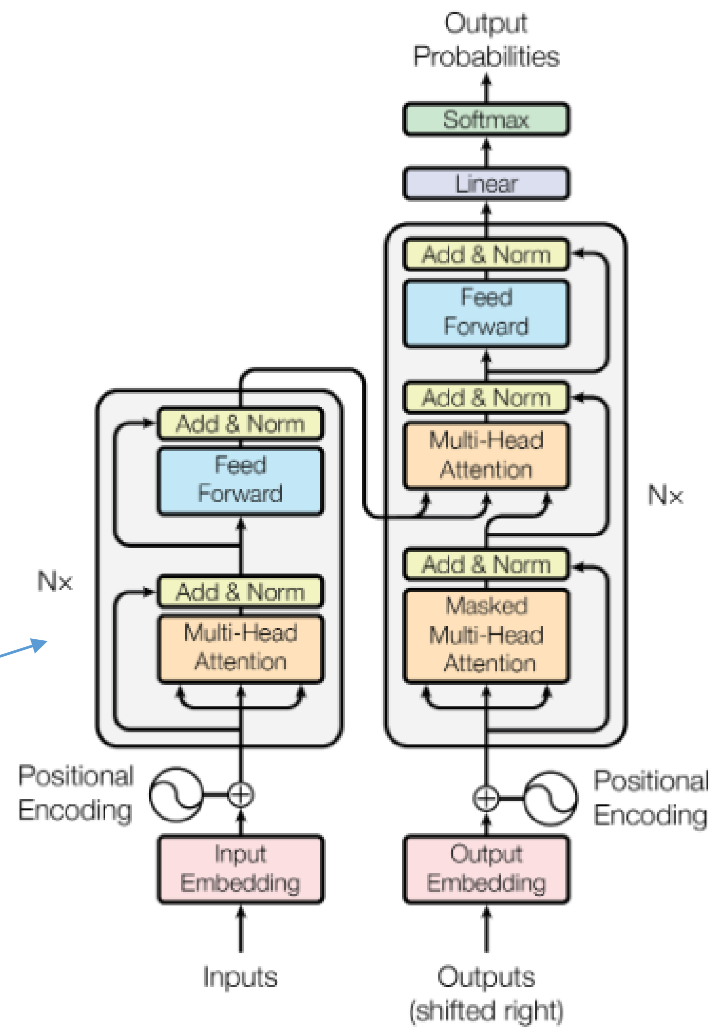
When training the encoder-decoder network, teacher forcing uses reference words in the decoder input instead of the predicted words

- Advantage: the learning becomes stable
- Disadvantage: inconsistency arise between training and testing phase



Transformer

- Faster than RNN to train
- Higher performance than RNN



Feed-forward network

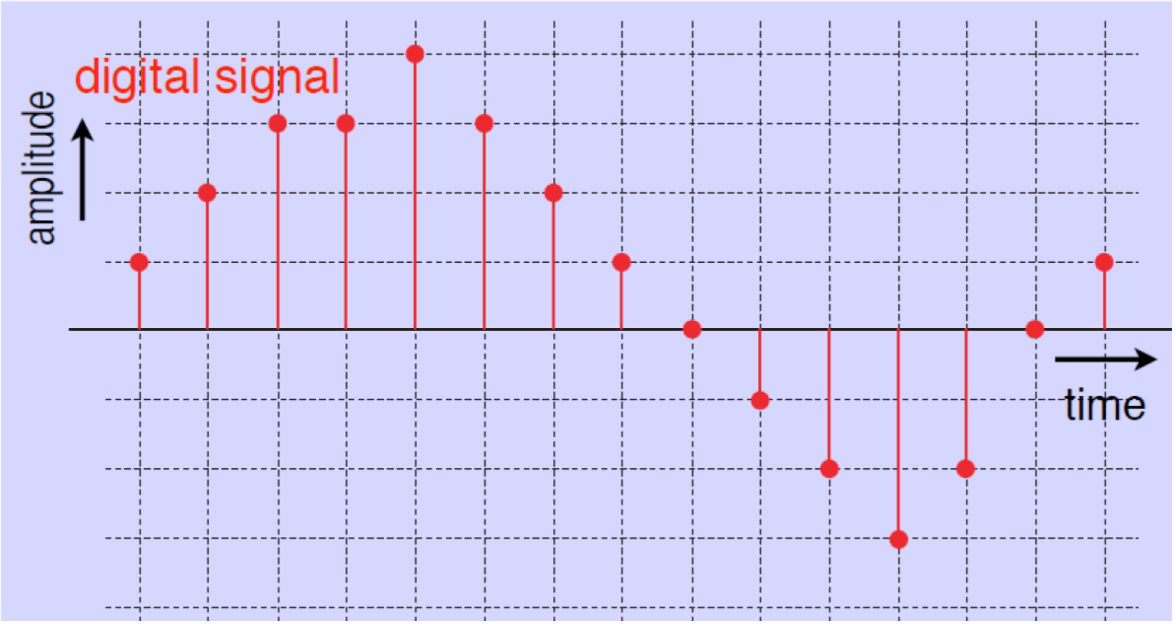
A. Vaswani+, Attention Is All You Need, 2017

Neural Network Based Speech Synthesis

WAVENET

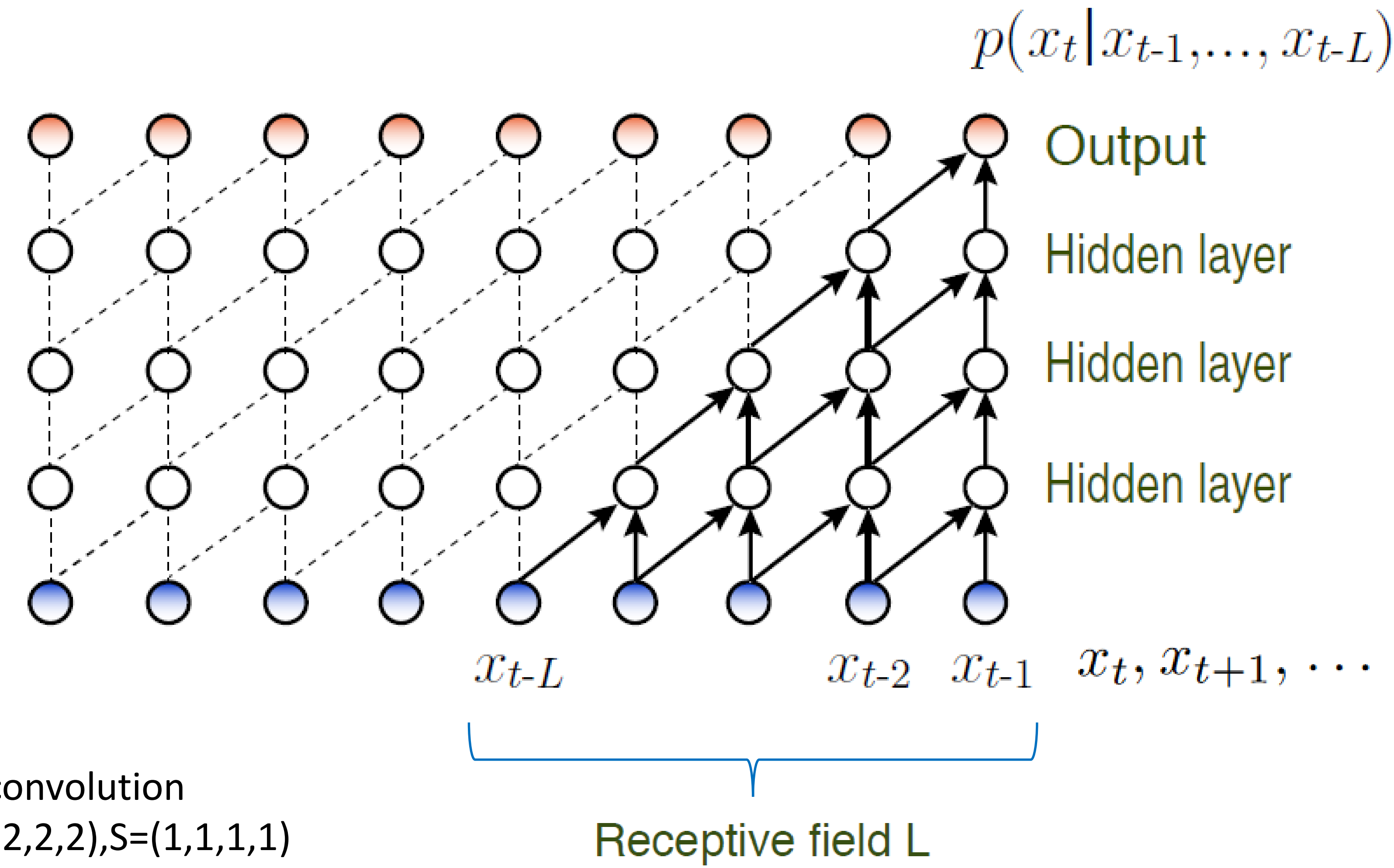
- A DNN based generative raw waveform model
  [van den Oord, et al., 2016]

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$



Causal Convolution

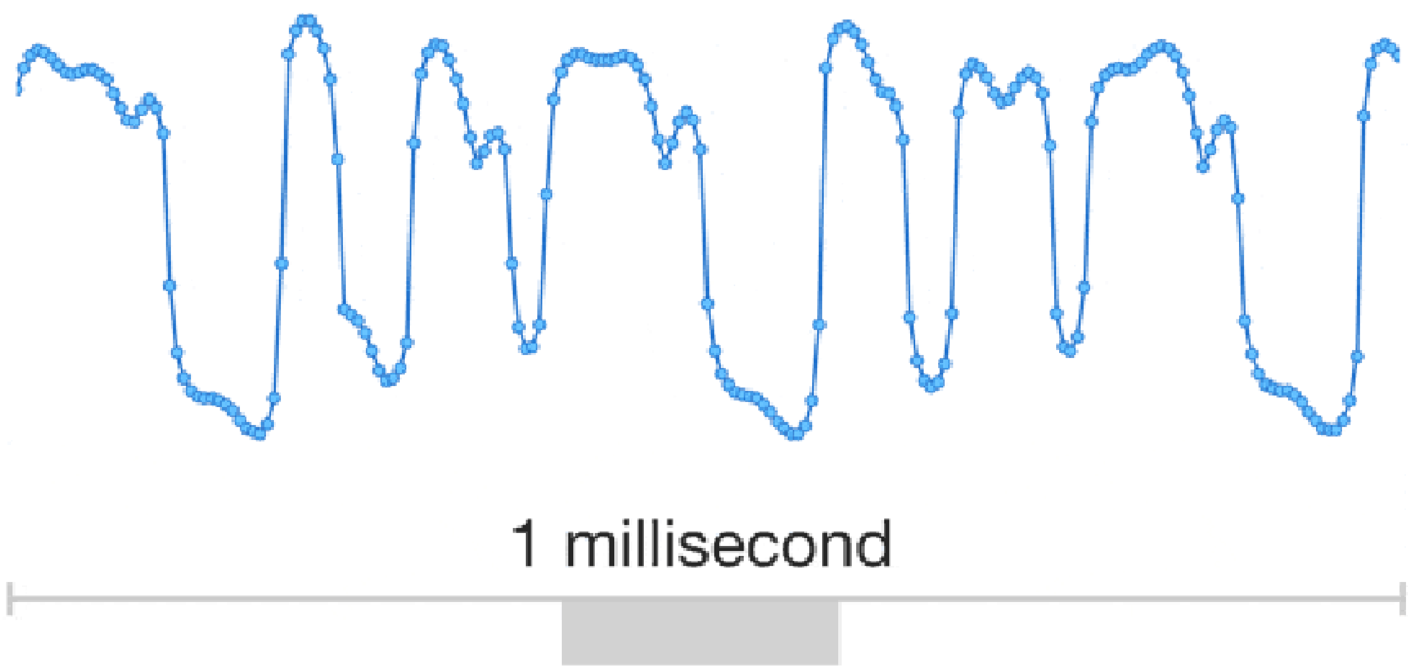
The prediction emitted at time step t is independent of future time steps $t, t+1, \cdots$



1-D convolution
K=(2,2,2,2),S=(1,1,1,1)

Categorical Prediction of Amplitude

- Discrete prediction by the softmax function is used, as it is found to work better than continuous regression

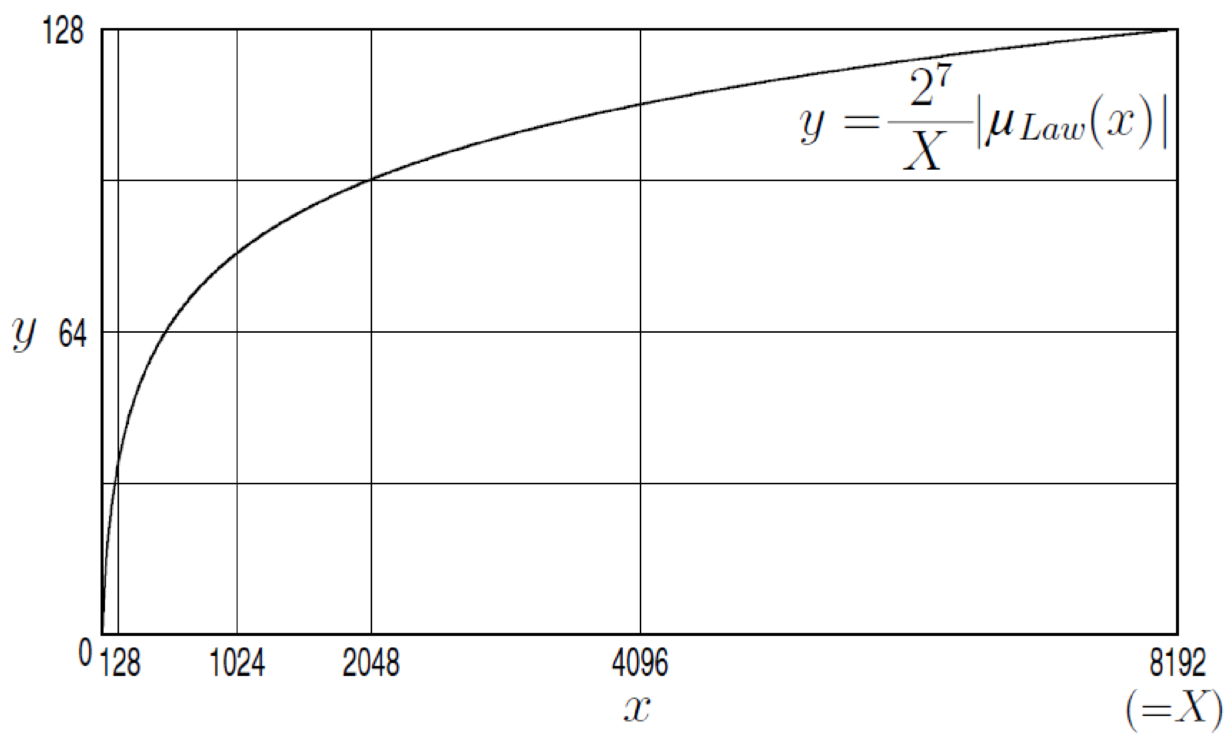


The figure is from DeepMind web page
Last visited 2021/5/24

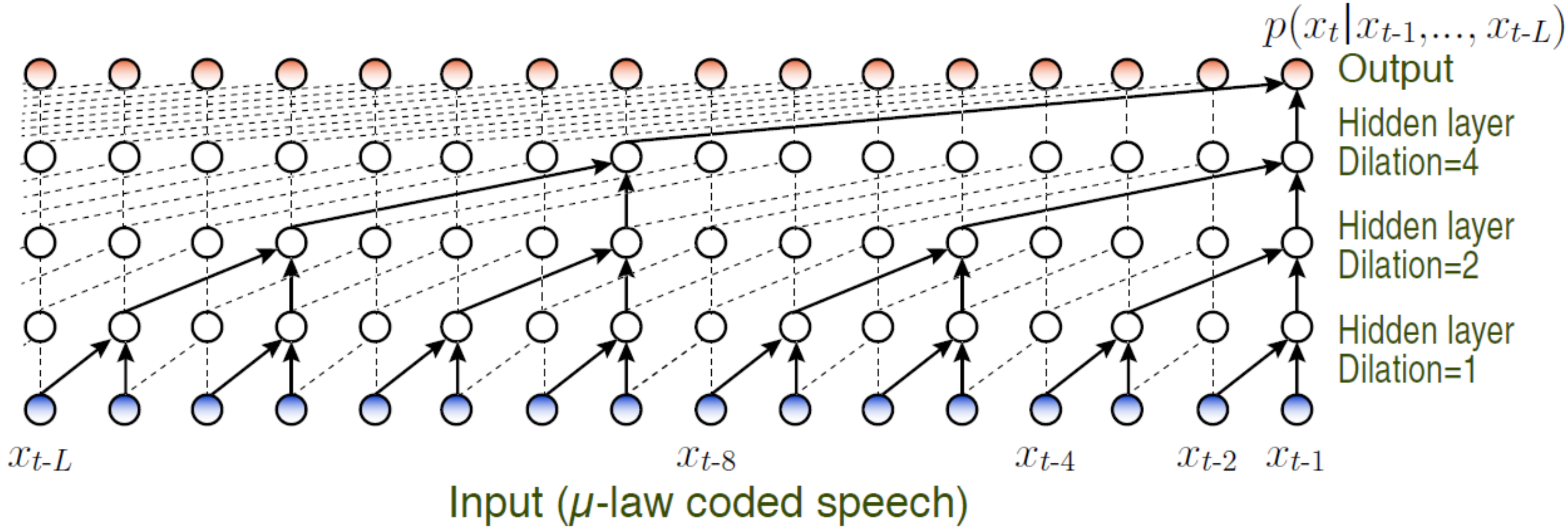
μ-Law Coding

$$\mu_{Law}(x) = \operatorname{sign}(x) X \frac{\log(1+\mu|x|/X)}{\log(1+\mu)}, \quad |x| \le X$$

- $\mu = 255$ for 8bit PCM



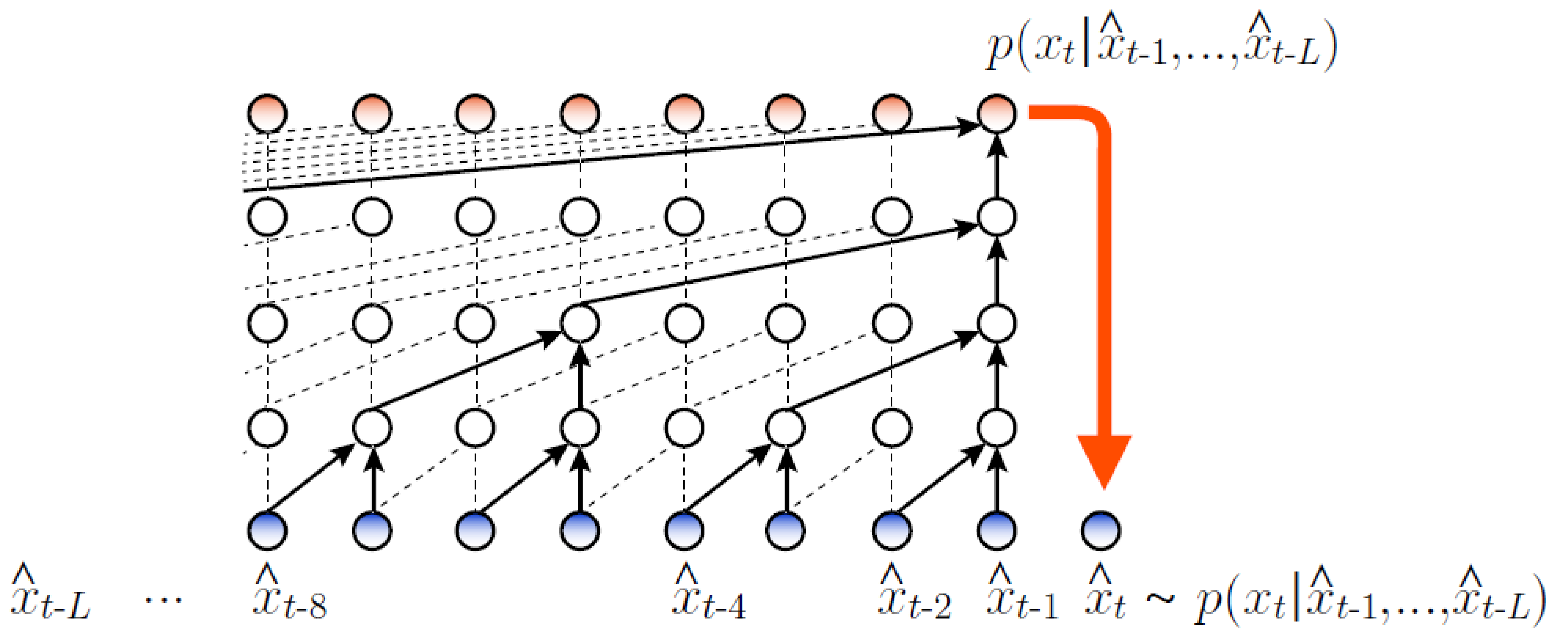
Dilated Causal Convolution



1-D convolution
K=(2,2,2,2),S=(2,2,2,2)

Signal Generation

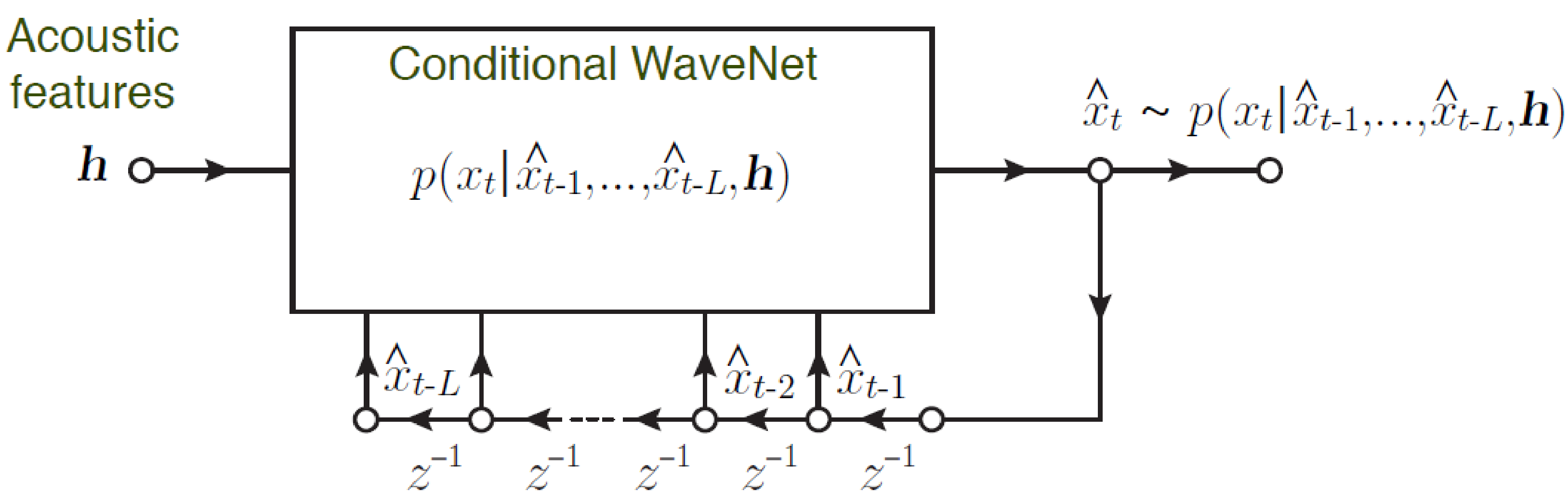
- Random sampling from estimated distribution



Conditional WavNet

$$\ln p(\mathbf{x} \mid \boldsymbol{h}) \approx \sum_{t=1}^{T} \ln p(x_t | x_{t\text{-}1},\ldots,x_{t\text{-}L},\boldsymbol{h})$$

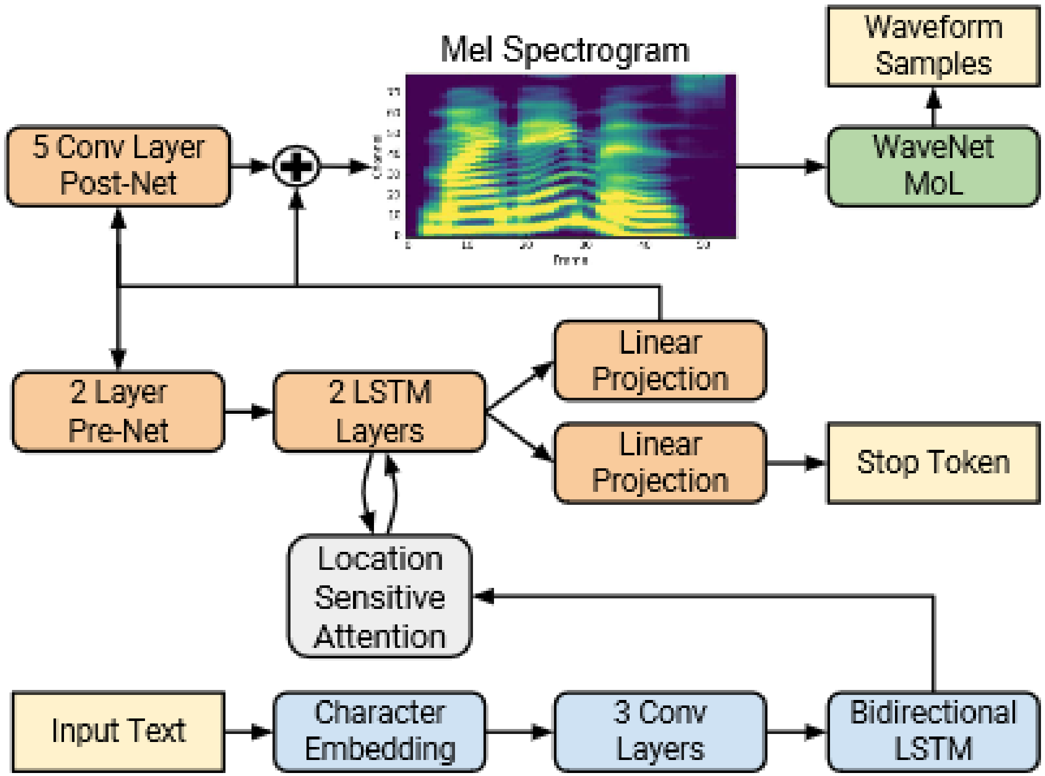
- Auxiliary input $\boldsymbol{h}$: F0, mel spectrum, spectrogram, etc.
- Receptive field L: several hundreds milliseconds



https://deepmind.com/blog/article/wavenet-generative-model-raw-audio
Last visited 2023/5/23

Tacotron 2

A neural network architecture for speech synthesis directly from text



*The figure is cited from J. Shen et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram predictions", ICASSP 2018.

https://google.github.io/tacotron/publications/tacotron/index.html
Last visited 2023/5/23

Exercise (Q4.1, Q4.2)

Q4.1

What is the receptive field length of 1-D convolution when K=(2,2,2,2,2),S=(1,1,1,1,1)?

Q4.2

What is the receptive field length of 1-D convolution when K=(2,2,2,2,2),S=(2,2,2,2,2)?