Deep Learning with Neural Networks

Attention networks and transformers

Pablo Martínez Olmos, pamartin@ing.uc3m.es

uc3m

Fully Observed Models. Deep belief networks

- Explicit model conditional probabilities

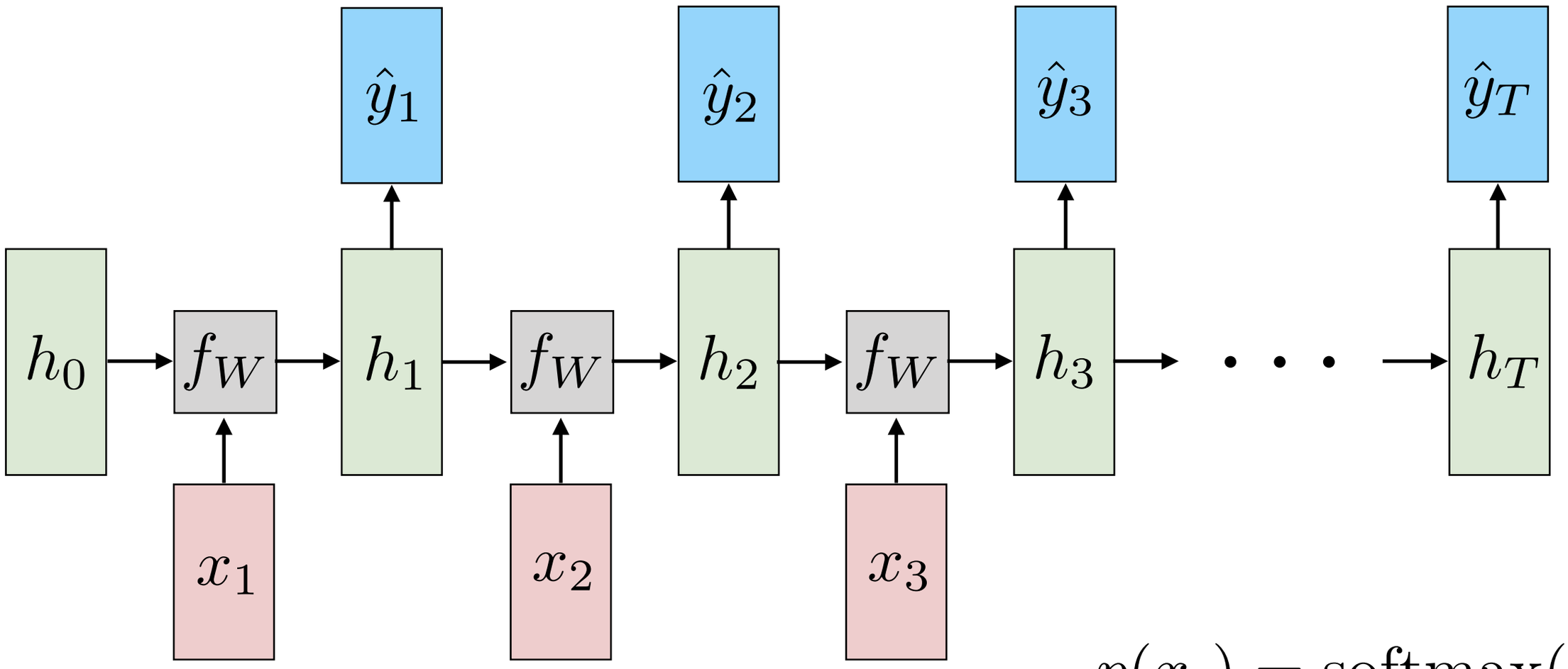
$$p_{\text{model}}(\boldsymbol{x}) = p_{\text{model}}(x_1) \prod_{i=2}^{n} p_{\text{model}}(x_i | x_1, \ldots, x_{i-1})$$

Each conditional can be a neural network

- Maximum likelihood training

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \log p_{\text{model}}(\mathbf{x}_n)$$

Recurrent Neural Language Model (Training)

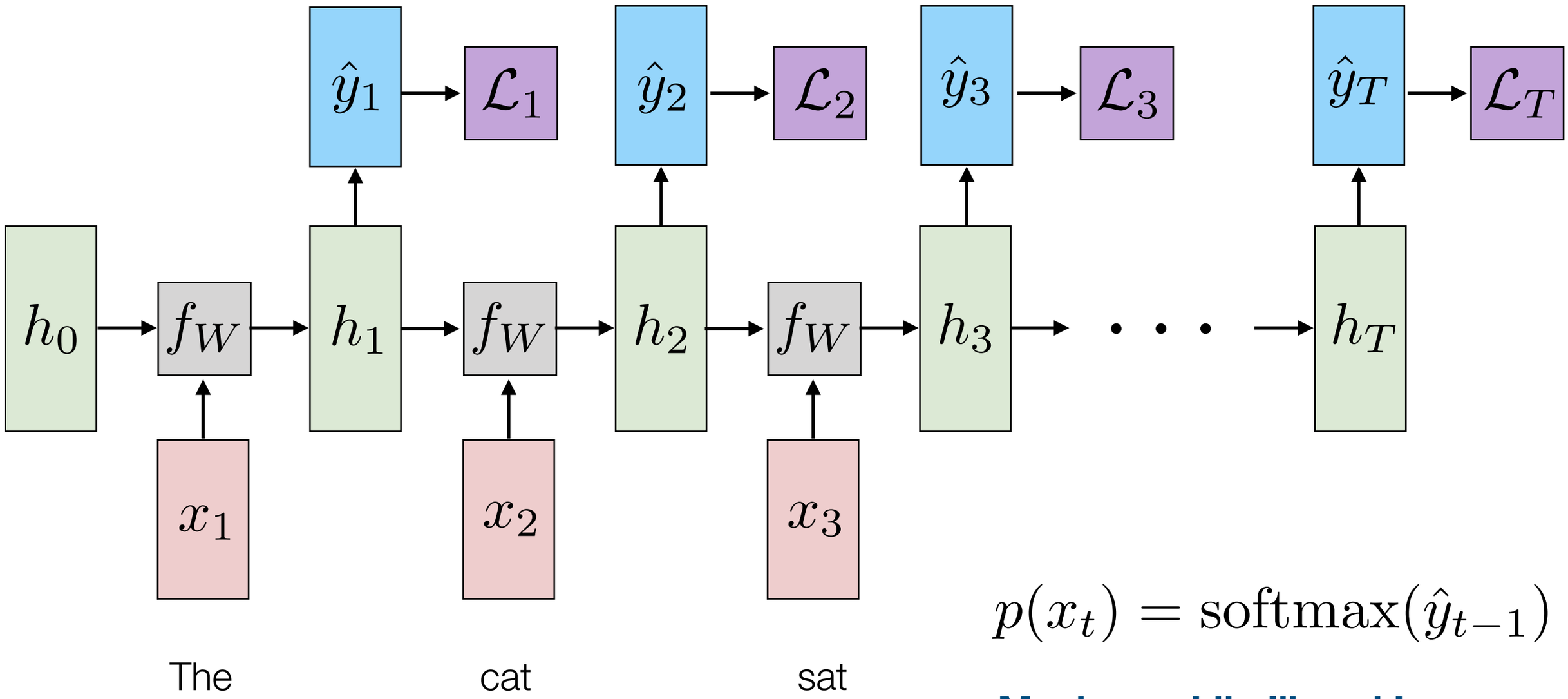


$$p(x_t) = \text{softmax}(\hat{y}_{t-1})$$

Maximum Likelihood Loss

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Training)

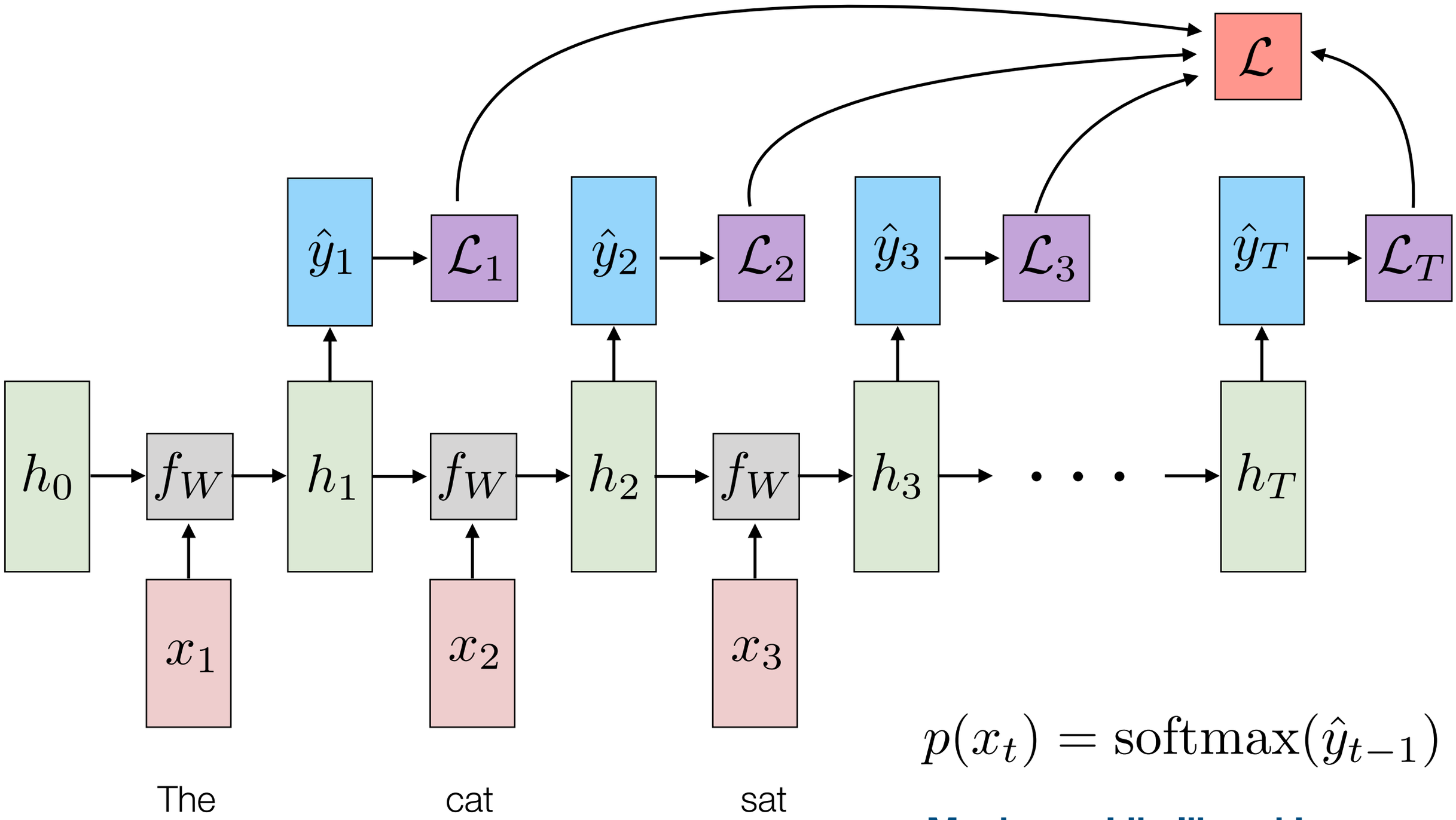


$$p(x_t) = \text{softmax}(\hat{y}_{t-1})$$

Maximum Likelihood Loss

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Training)



$$p(x_t) = \text{softmax}(\hat{y}_{t-1})$$

Maximum Likelihood Loss

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

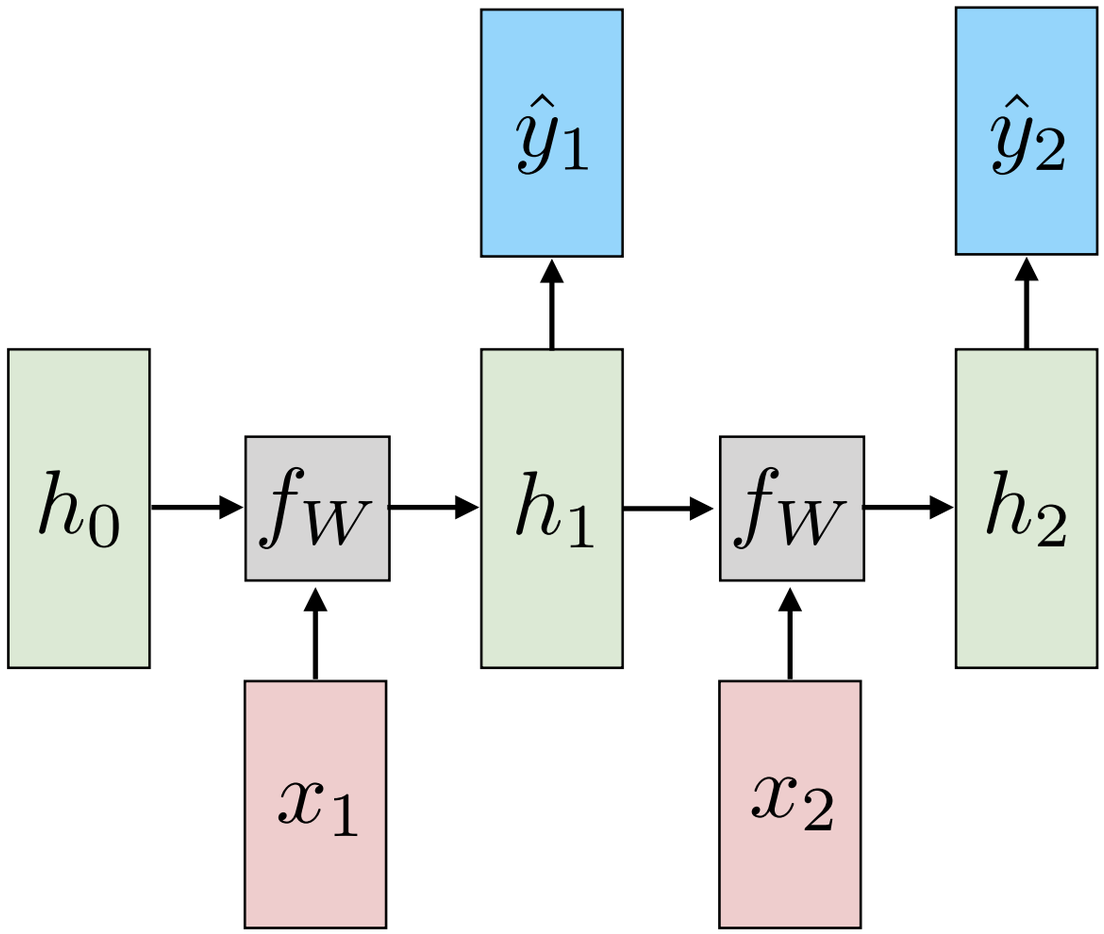
Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

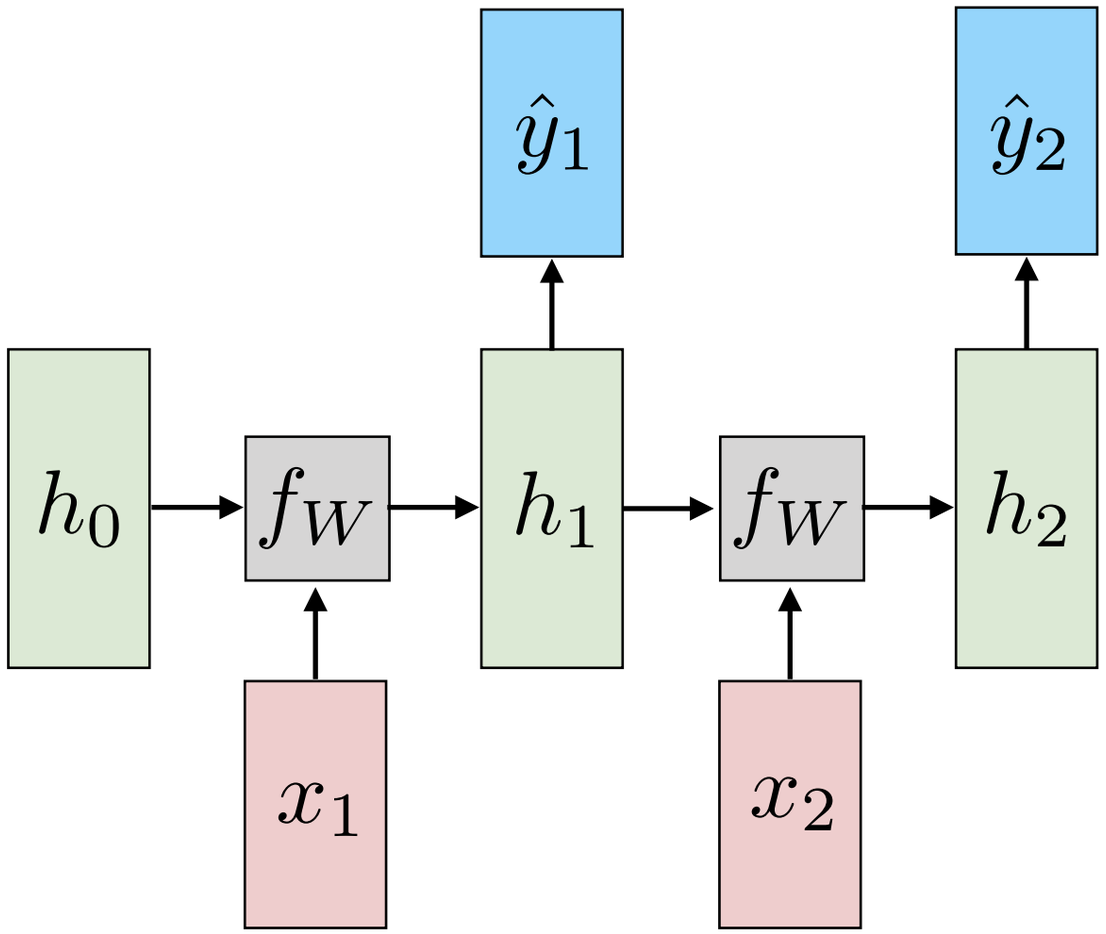
Sequentially Sample from Output Distribution



Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

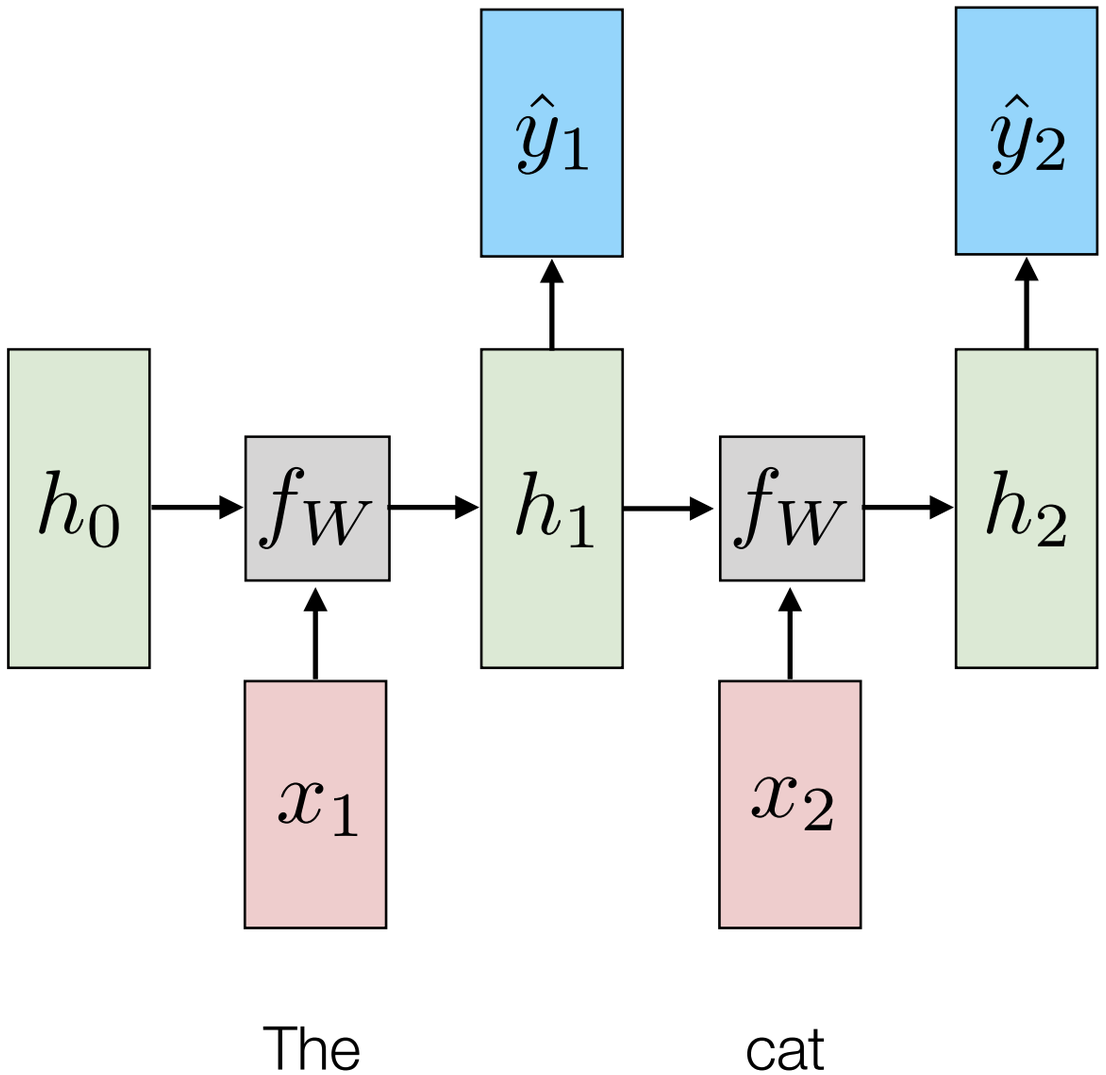
Sequentially Sample from Output Distribution



Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

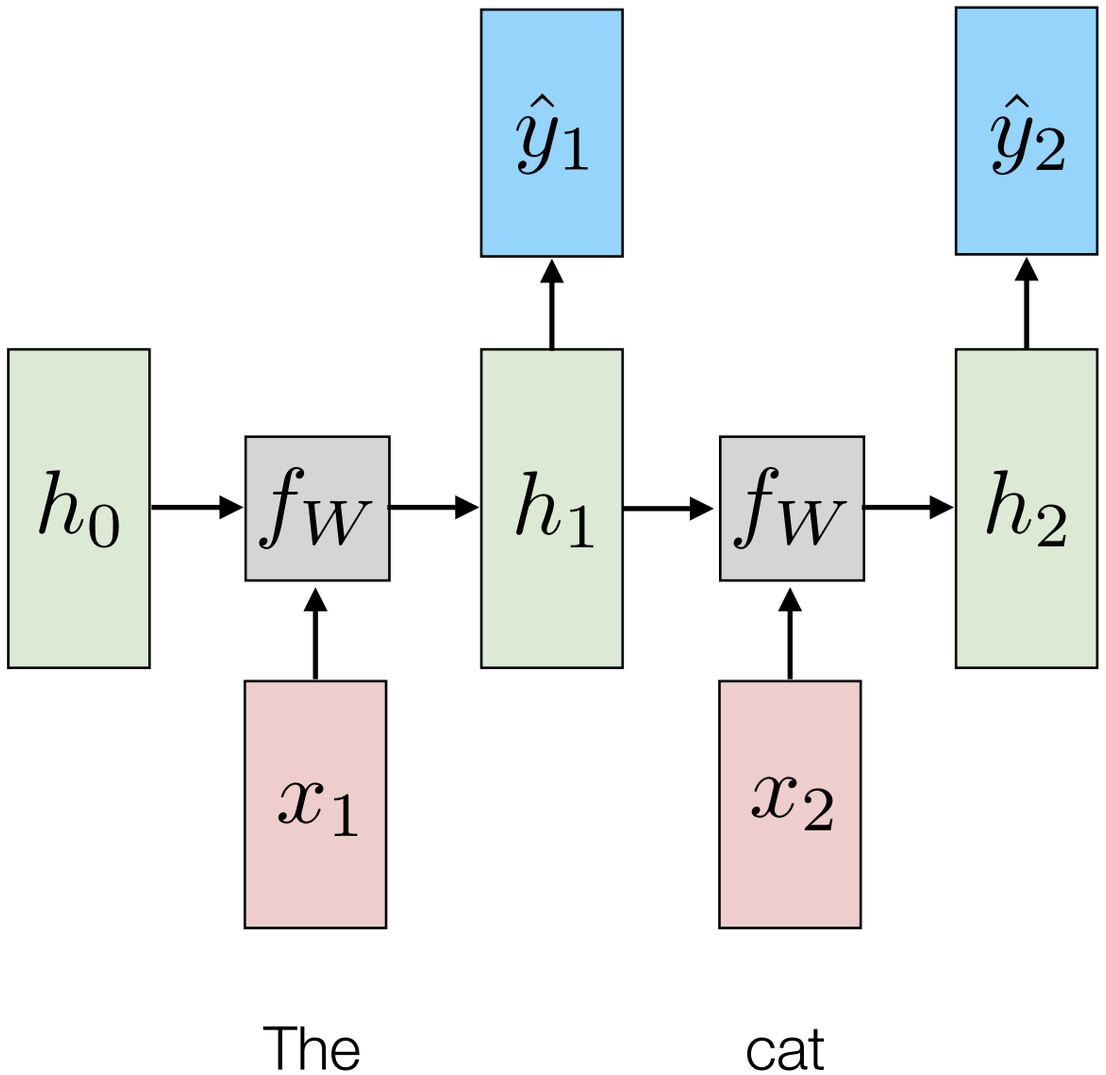


Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

$$\tilde{x}_3 \sim p(x_3) = \text{softmax}(\hat{y}_2)$$

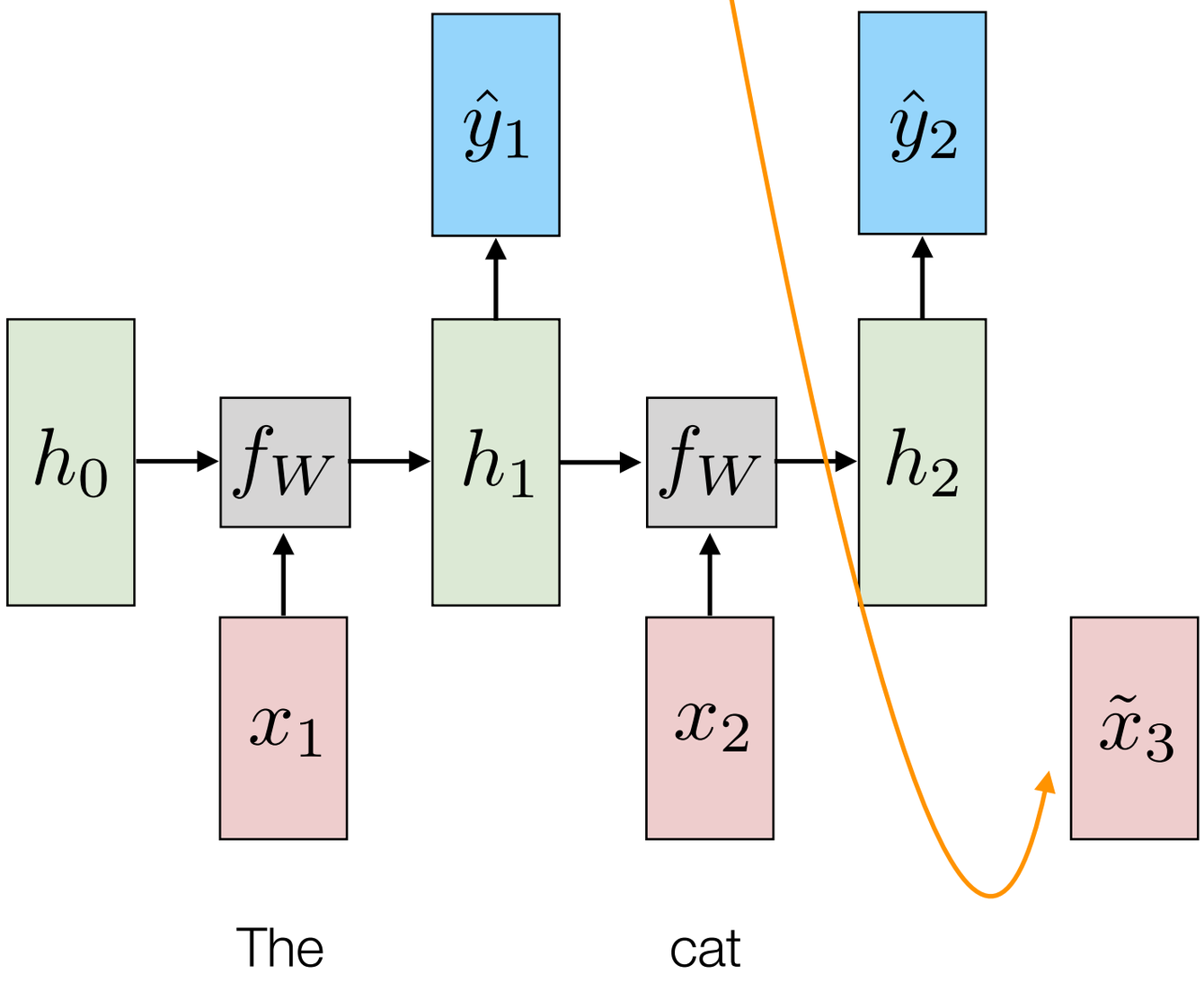


Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

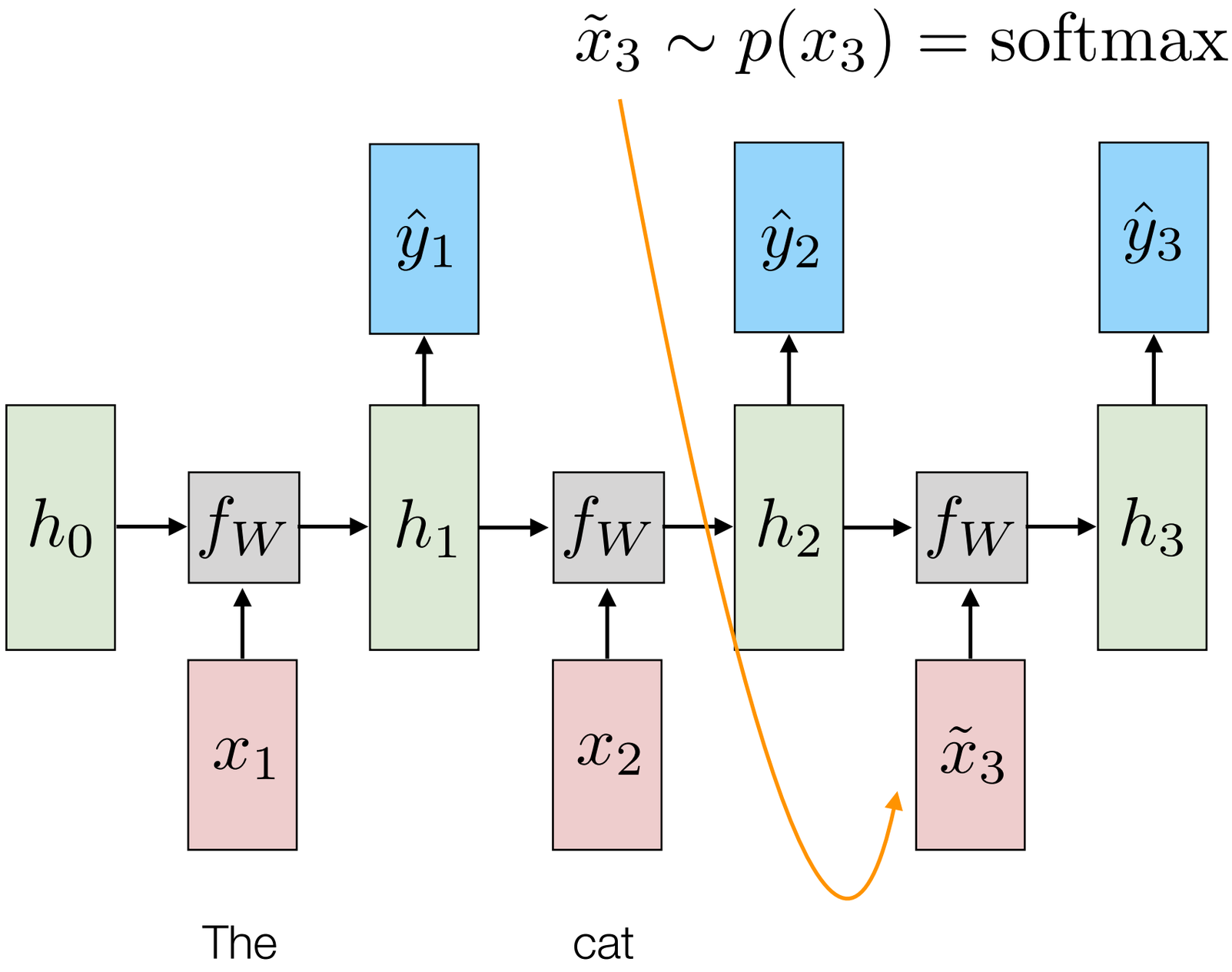
$$\tilde{x}_3 \sim p(x_3) = \text{softmax}(\hat{y}_2)$$



Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

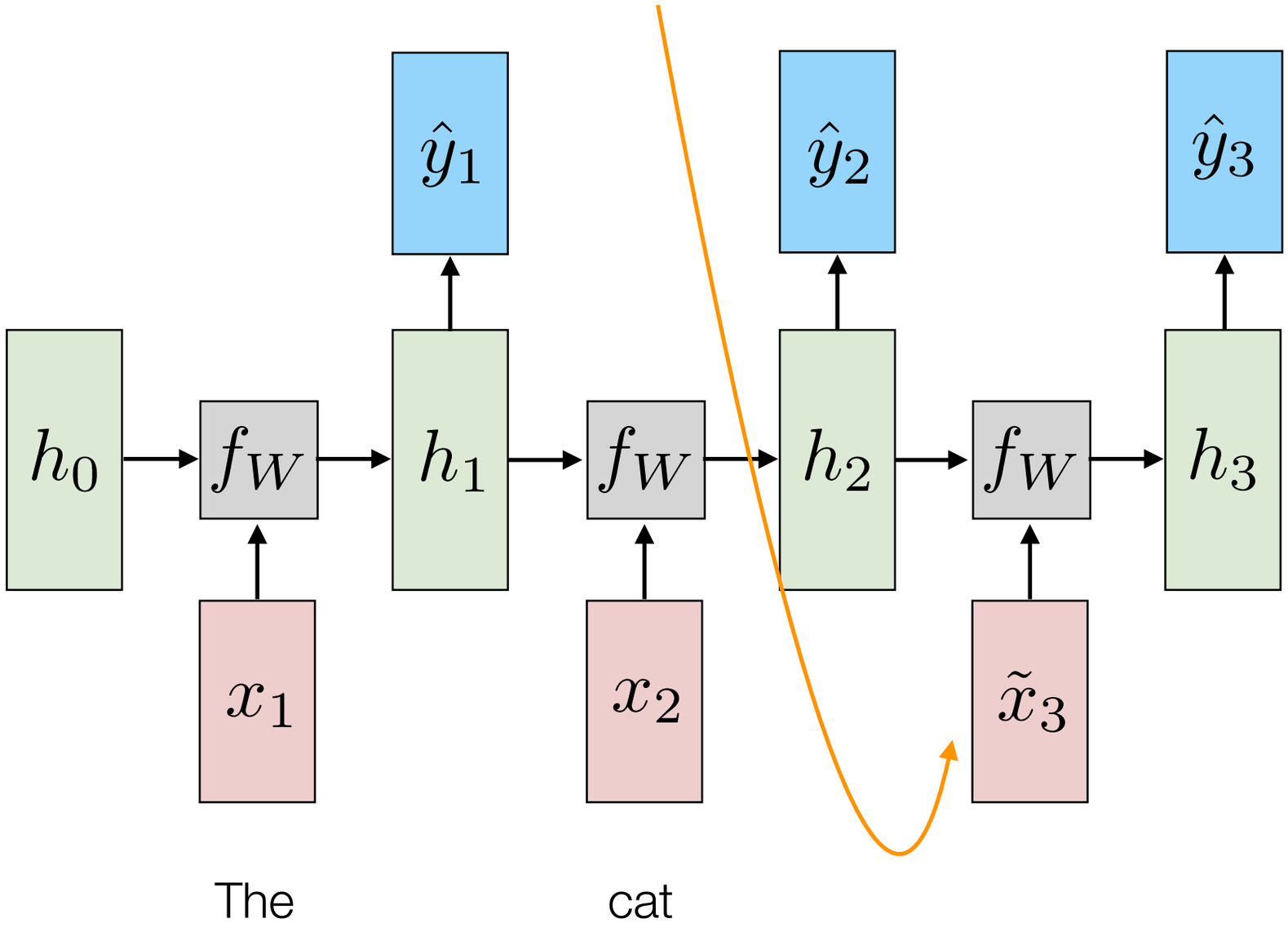
Sequentially Sample from Output Distribution



Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

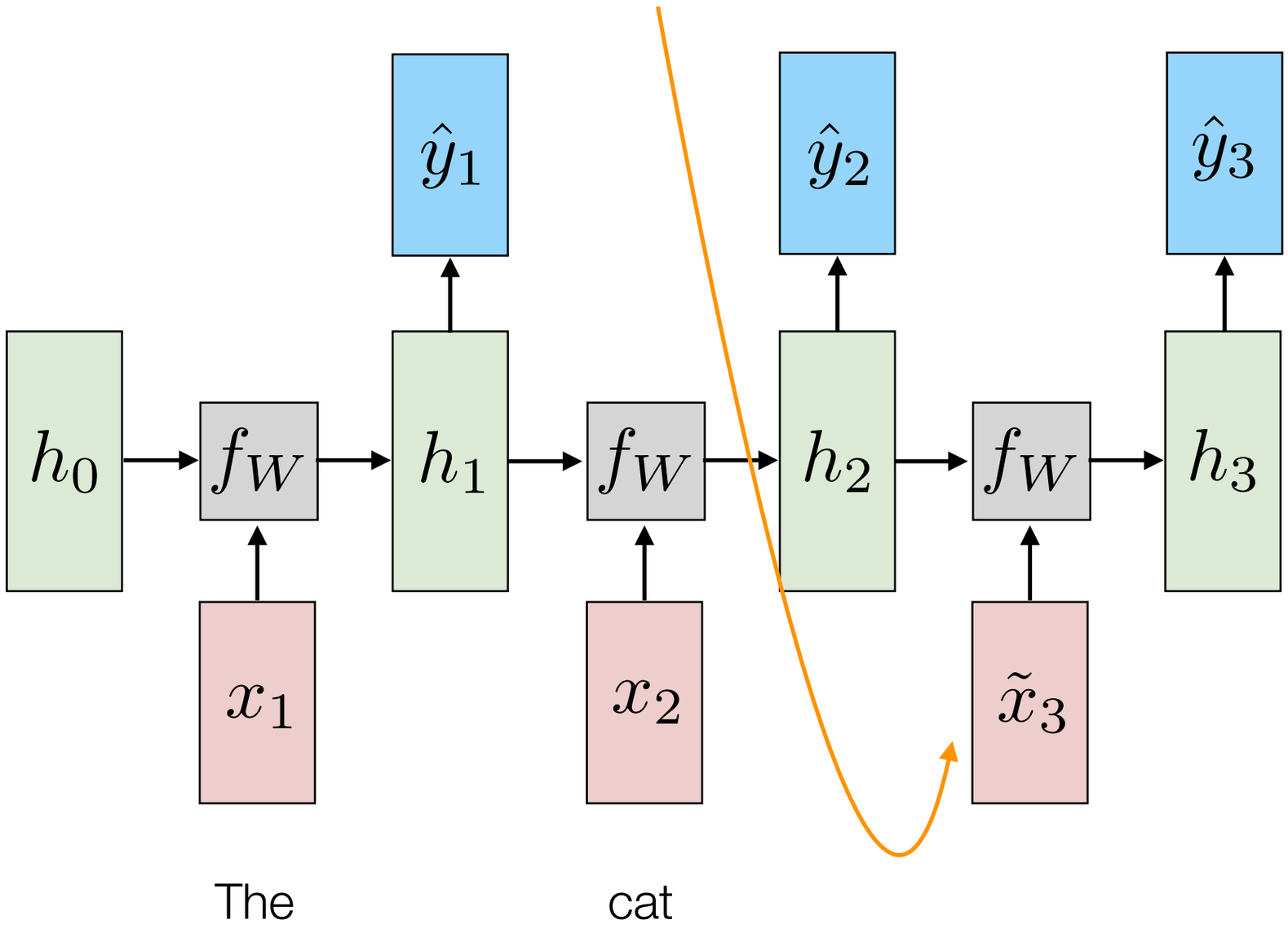


Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

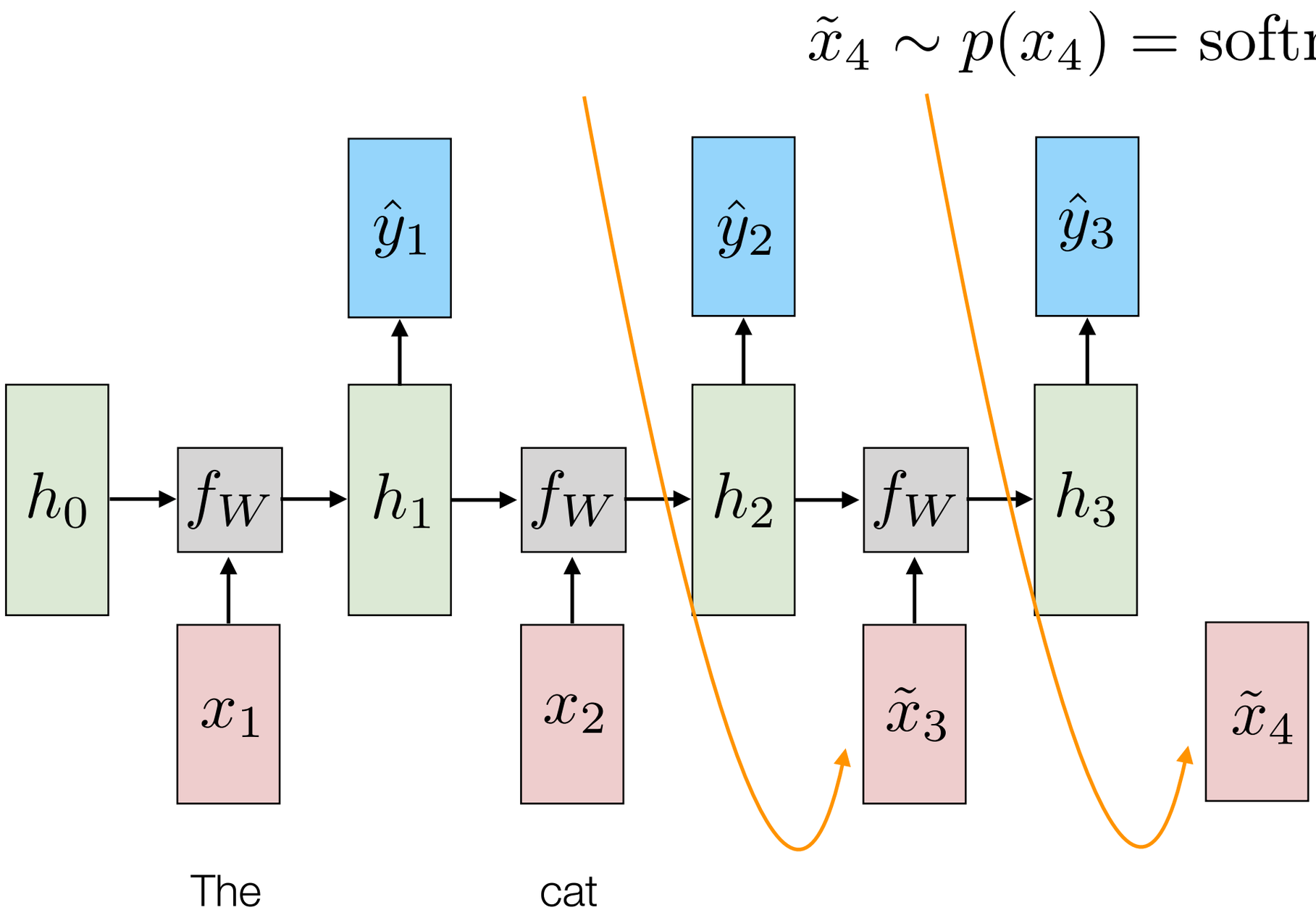
$$\tilde{x}_4 \sim p(x_4) = \text{softmax}(\hat{y}_3)$$



Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

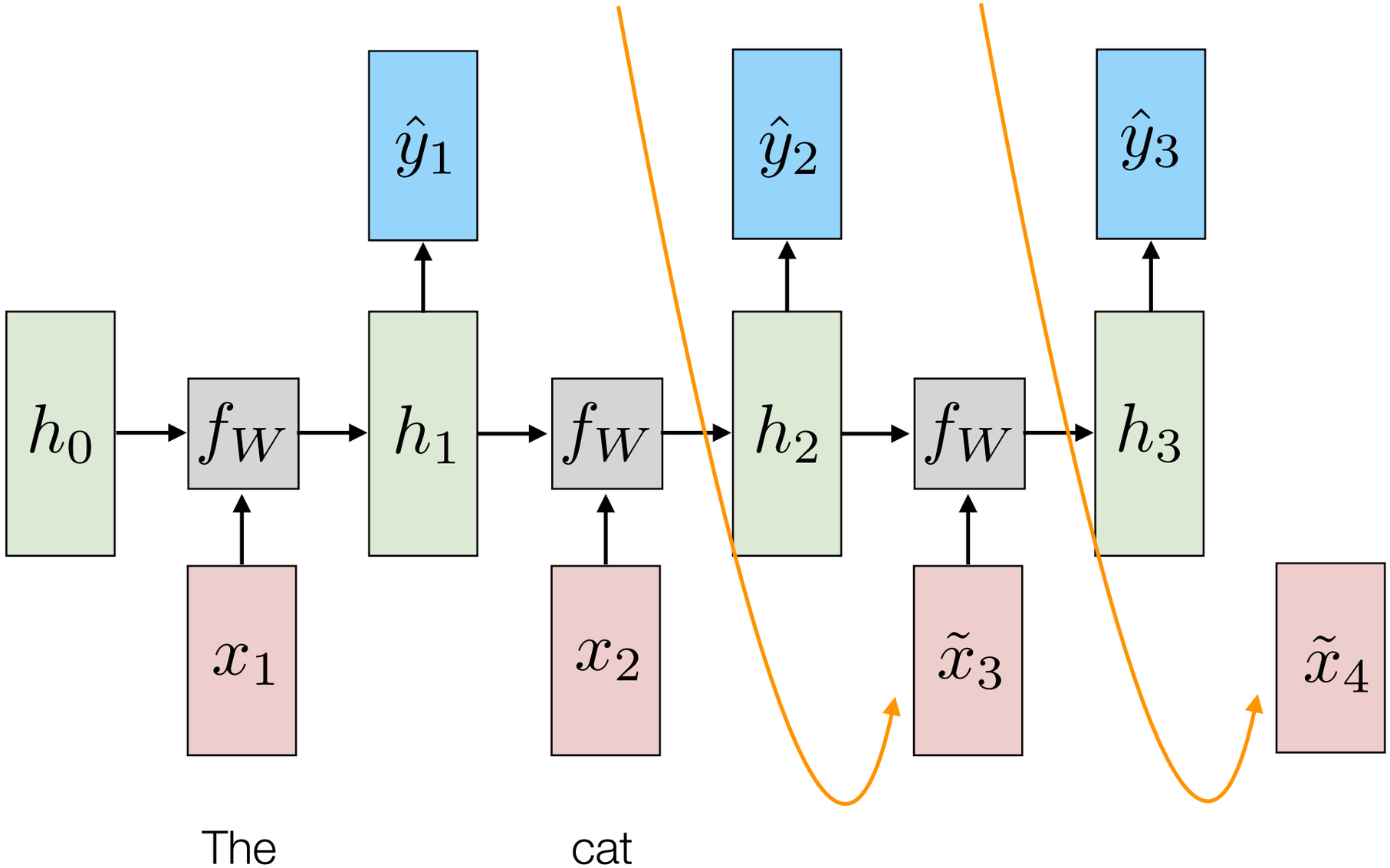
Sequentially Sample from Output Distribution



Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

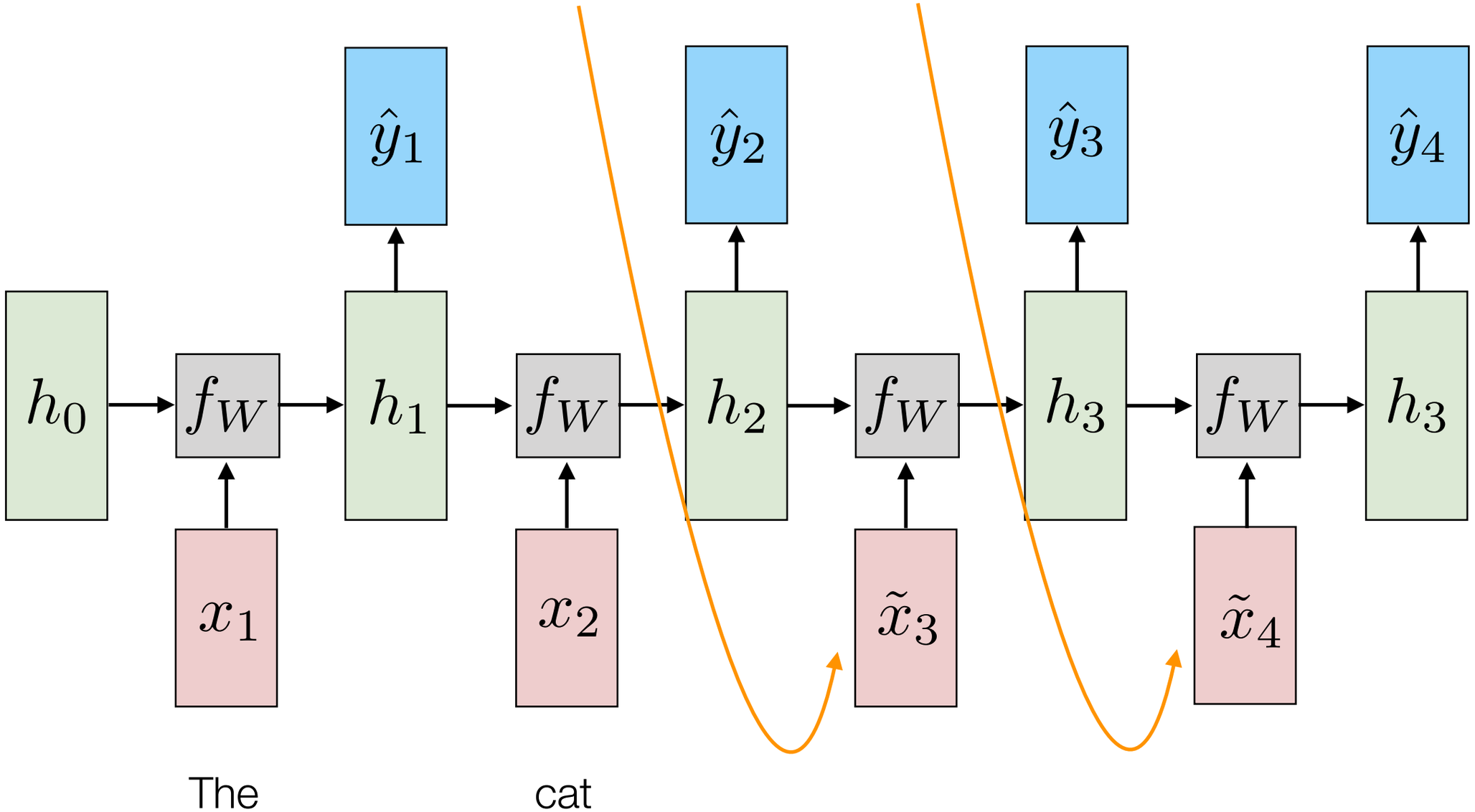
Sequentially Sample from Output Distribution



Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

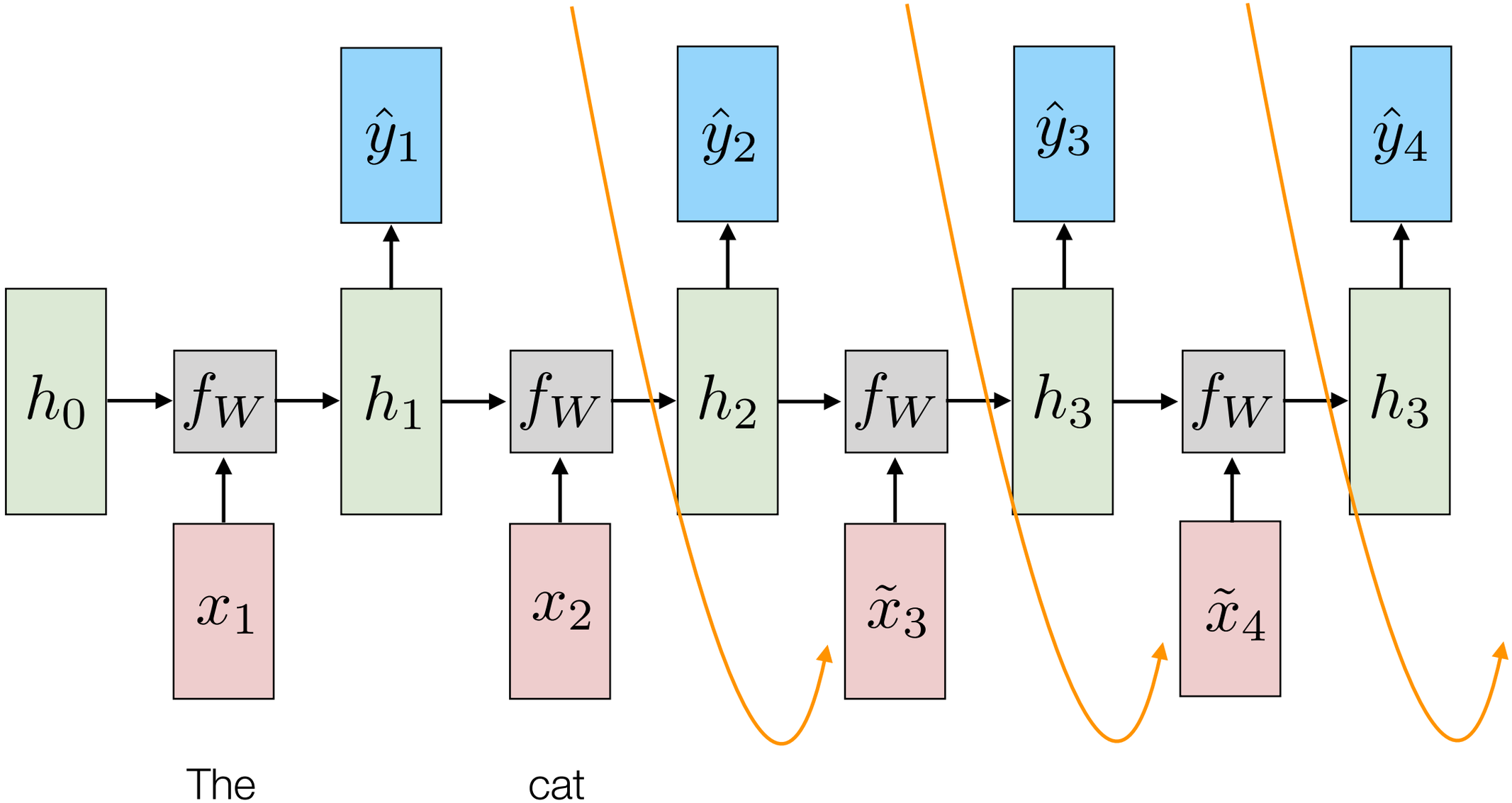


Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

$$\tilde{x}_5 \sim p(x_5) = \text{softmax}(\hat{y}_4)$$

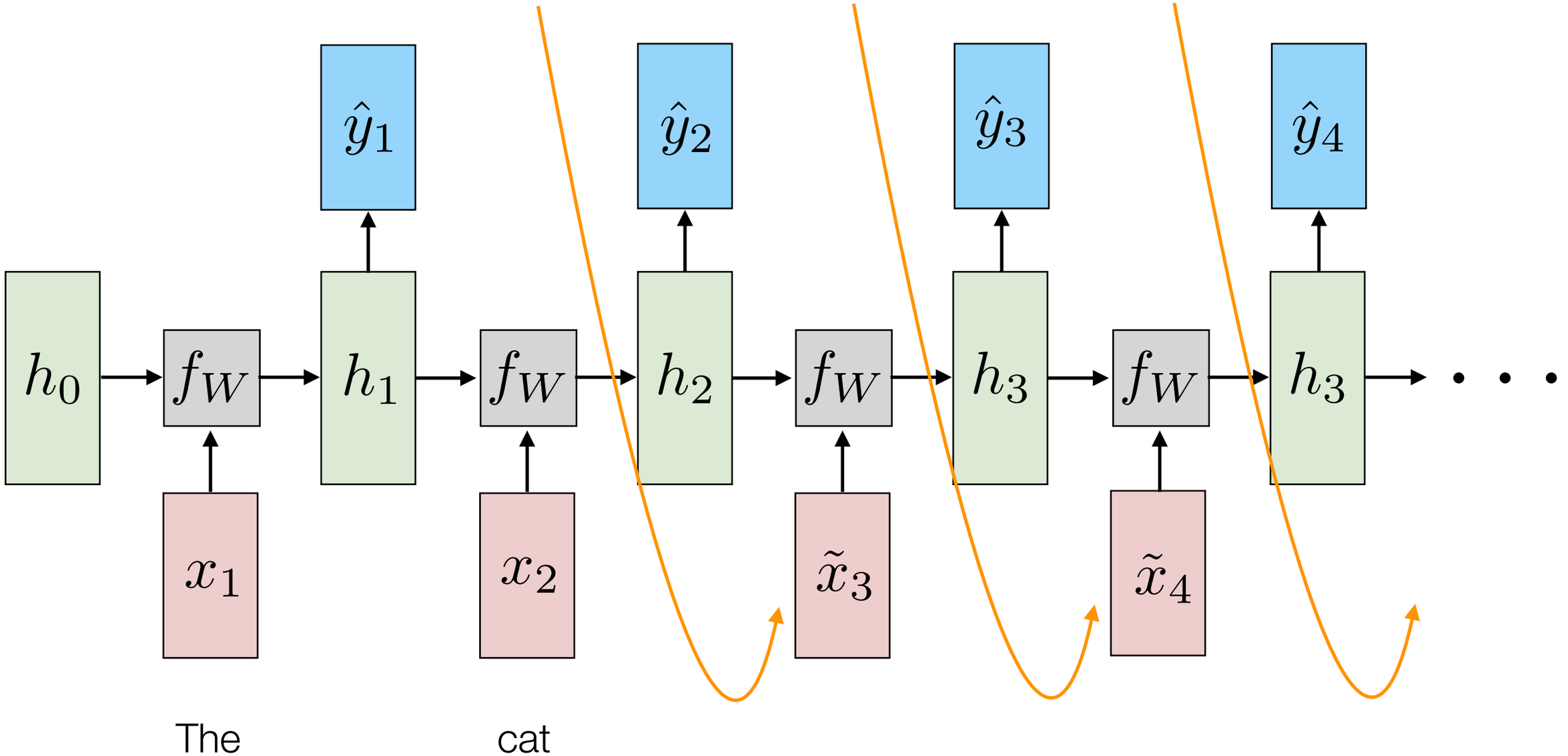


Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

Recurrent Neural Language Model (Test)

Sequentially Sample from Output Distribution

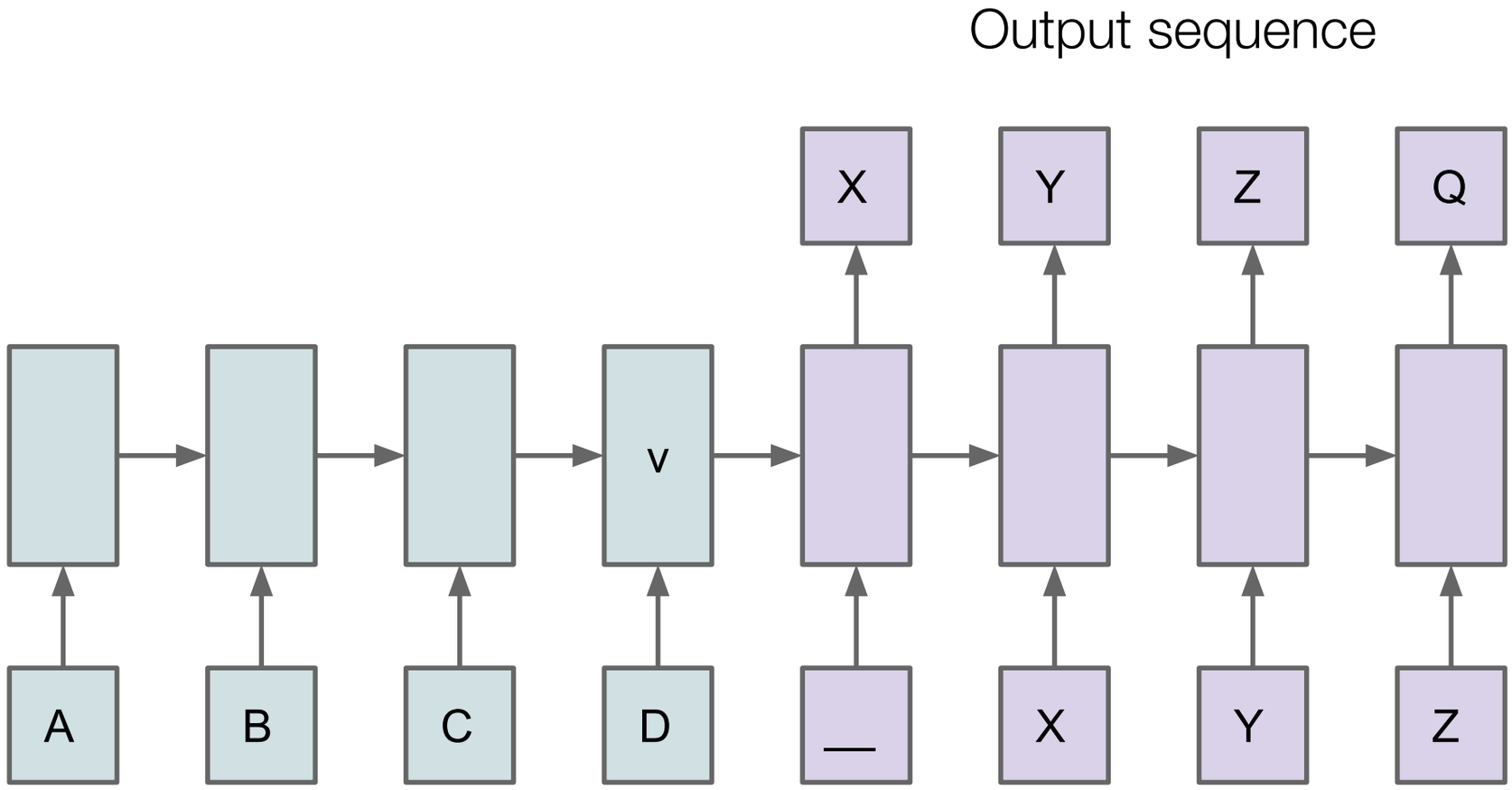
$$\tilde{x}_5 \sim p(x_5) = \text{softmax}(\hat{y}_4)$$



Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. InProceedings of INTERSPEECH

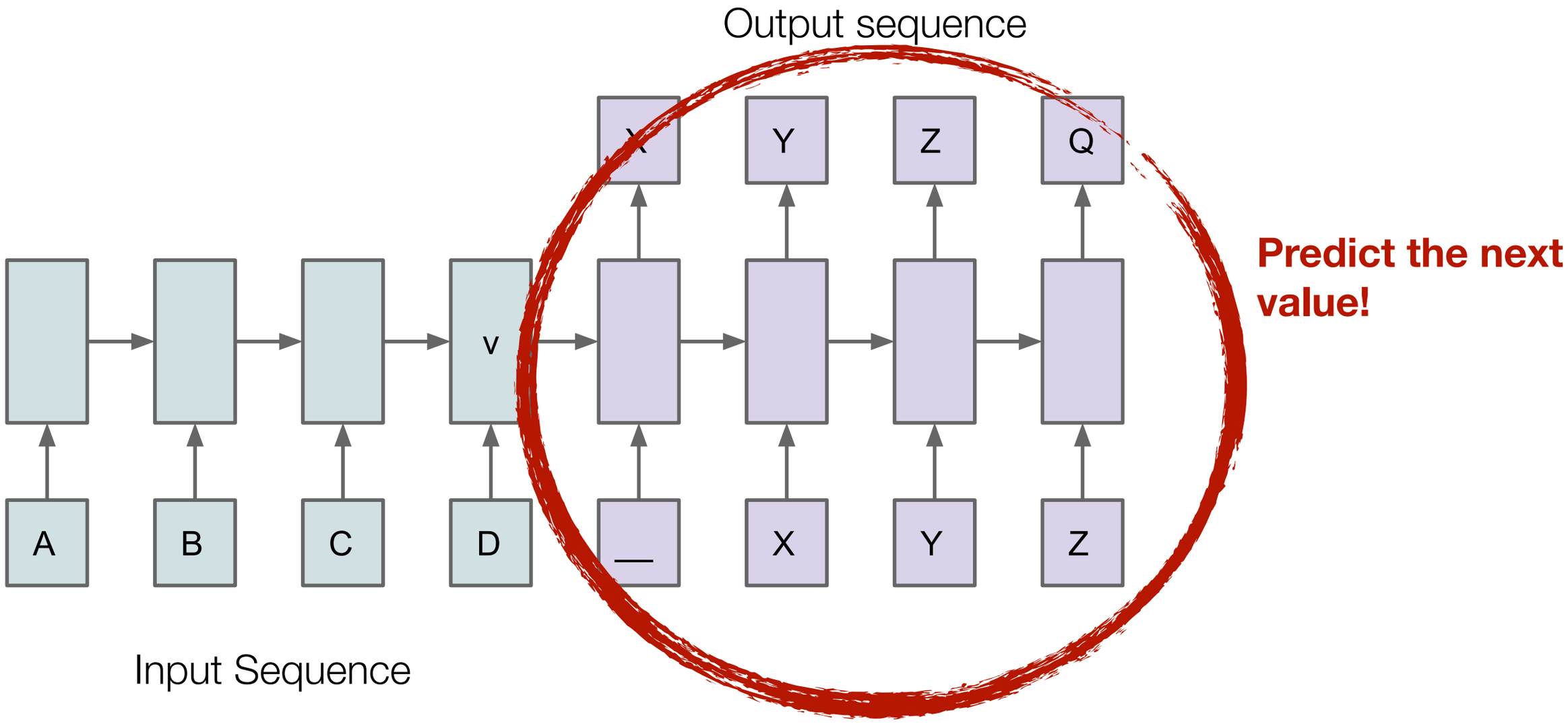
Seq2Seq & Attention Networks

Seq2Seq



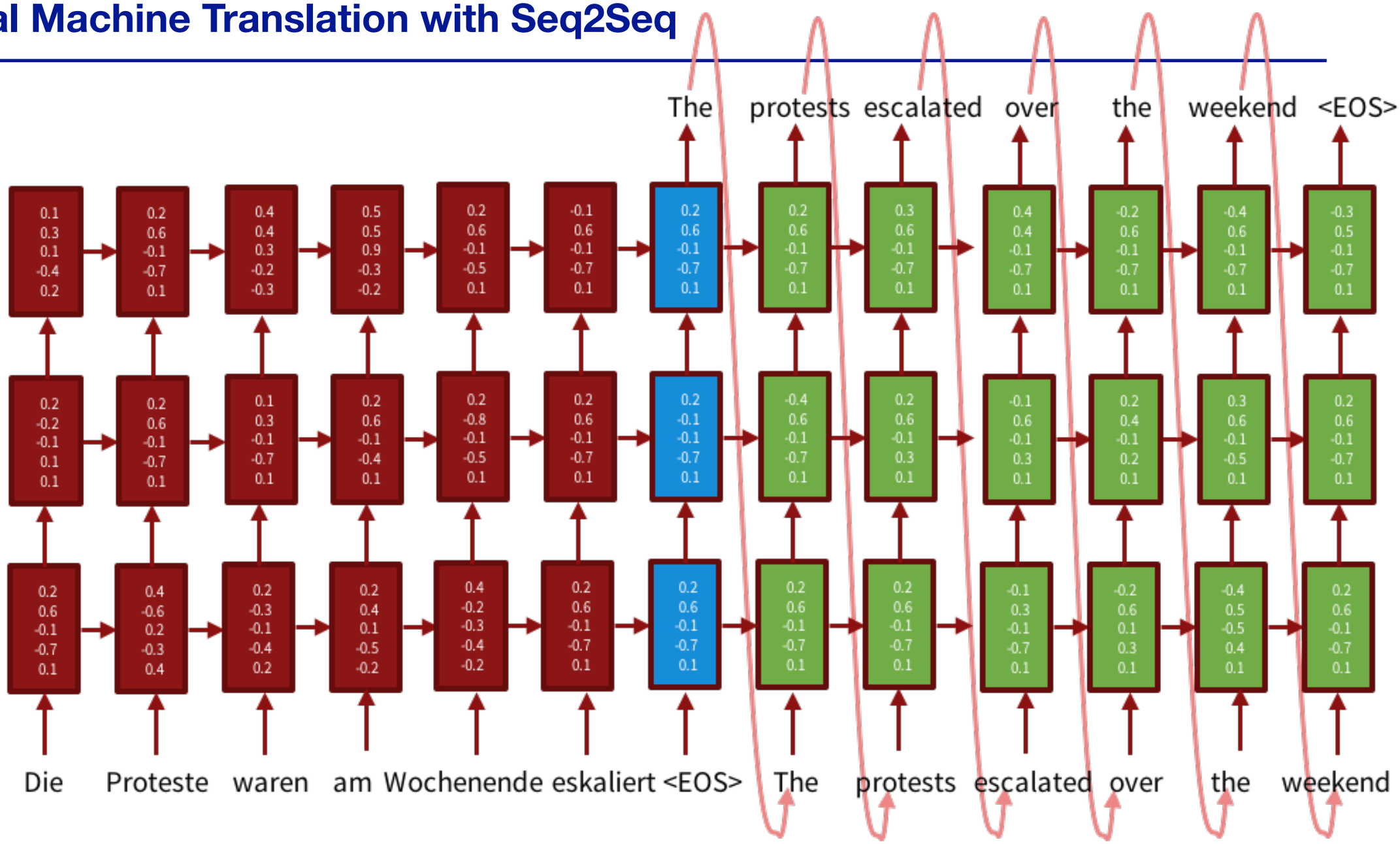
$$P\left(y_1, \ldots, y_{T'} | x_1, \ldots, x_T\right) = \prod_{t=1}^{T'} p\left(y_t | v, y_1, \ldots, y_{t-1}\right)$$

Seq2Seq



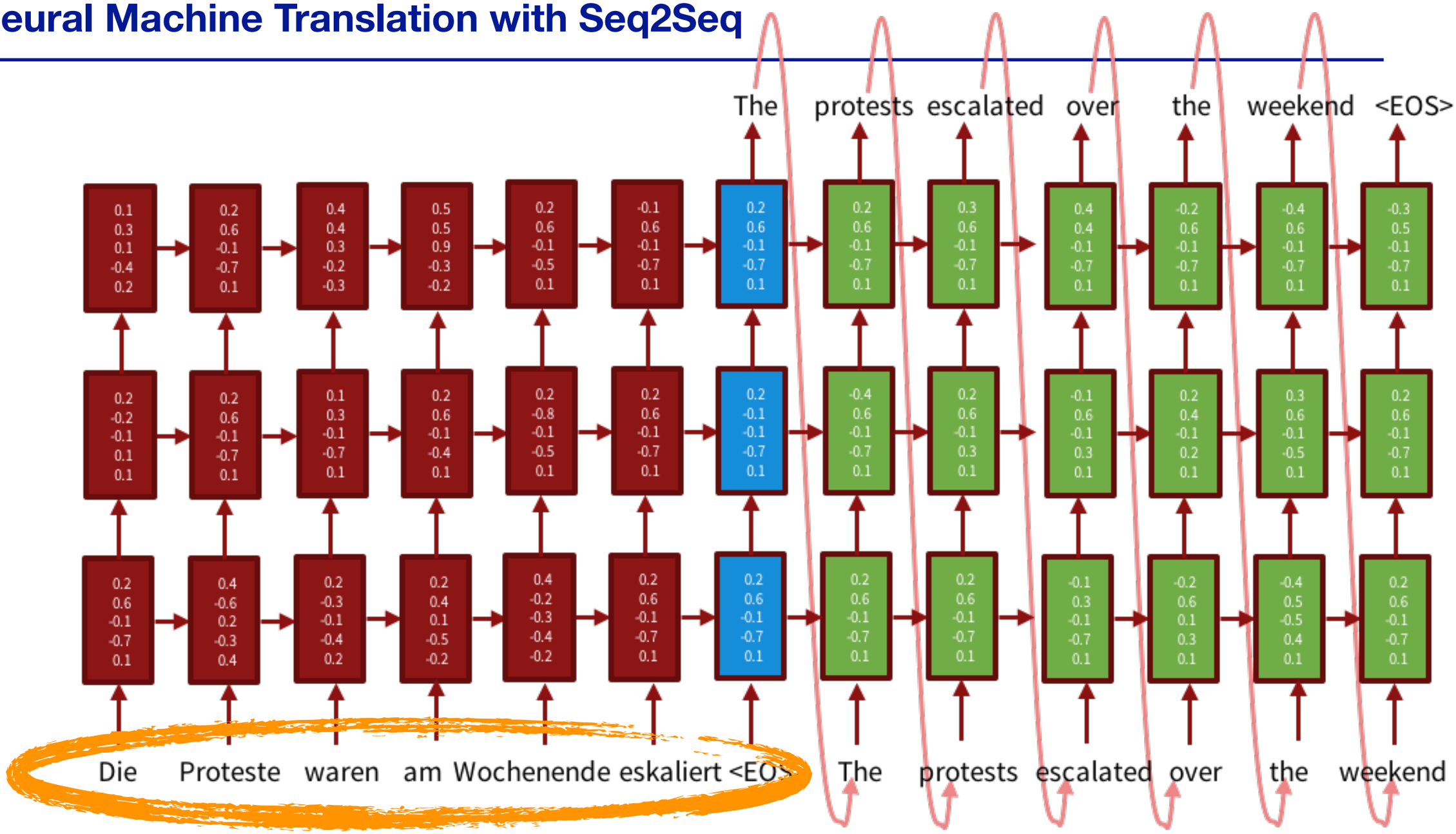
$$P\left(y_1, \ldots, y_{T'} | x_1, \ldots, x_T\right) = \prod_{t=1}^{T'} p\left(y_t | v, y_1, \ldots, y_{t-1}\right)$$

Neural Machine Translation with Seq2Seq



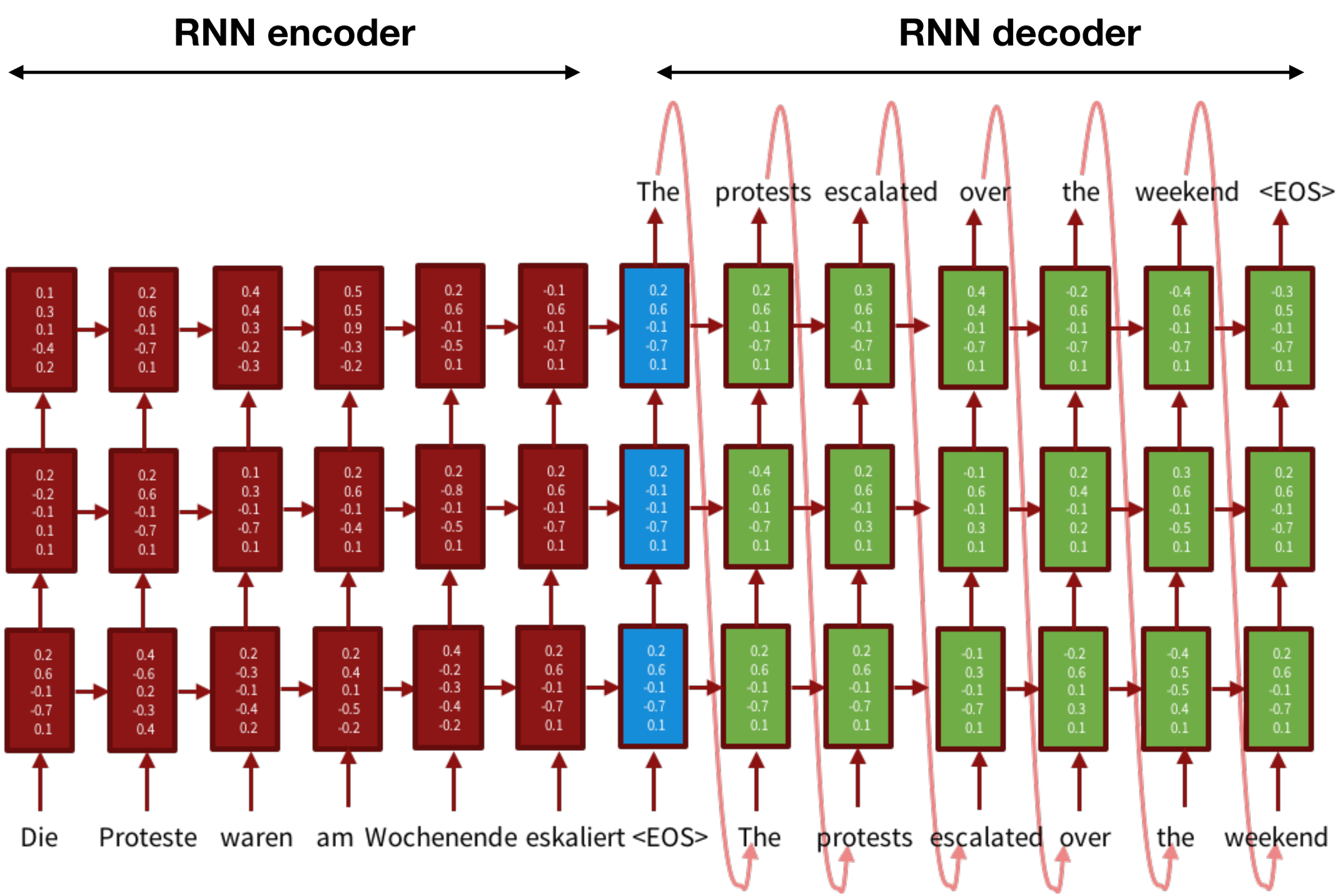
1. Auli, M., et al. "Joint Language and Translation Modeling with Recurrent Neural Networks." *EMNLP (2013)*
2. Kalchbrenner, N., et al. “Recurrent Continuous Translation Models.” *EMNLP (2013)*
3. Cho, K., et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical MT.” *EMNLP (2014)*
4. Sutskever, I., et al. “Sequence to Sequence Learning with Neural Networks.” *NIPS (2014)*

Neural Machine Translation with Seq2Seq

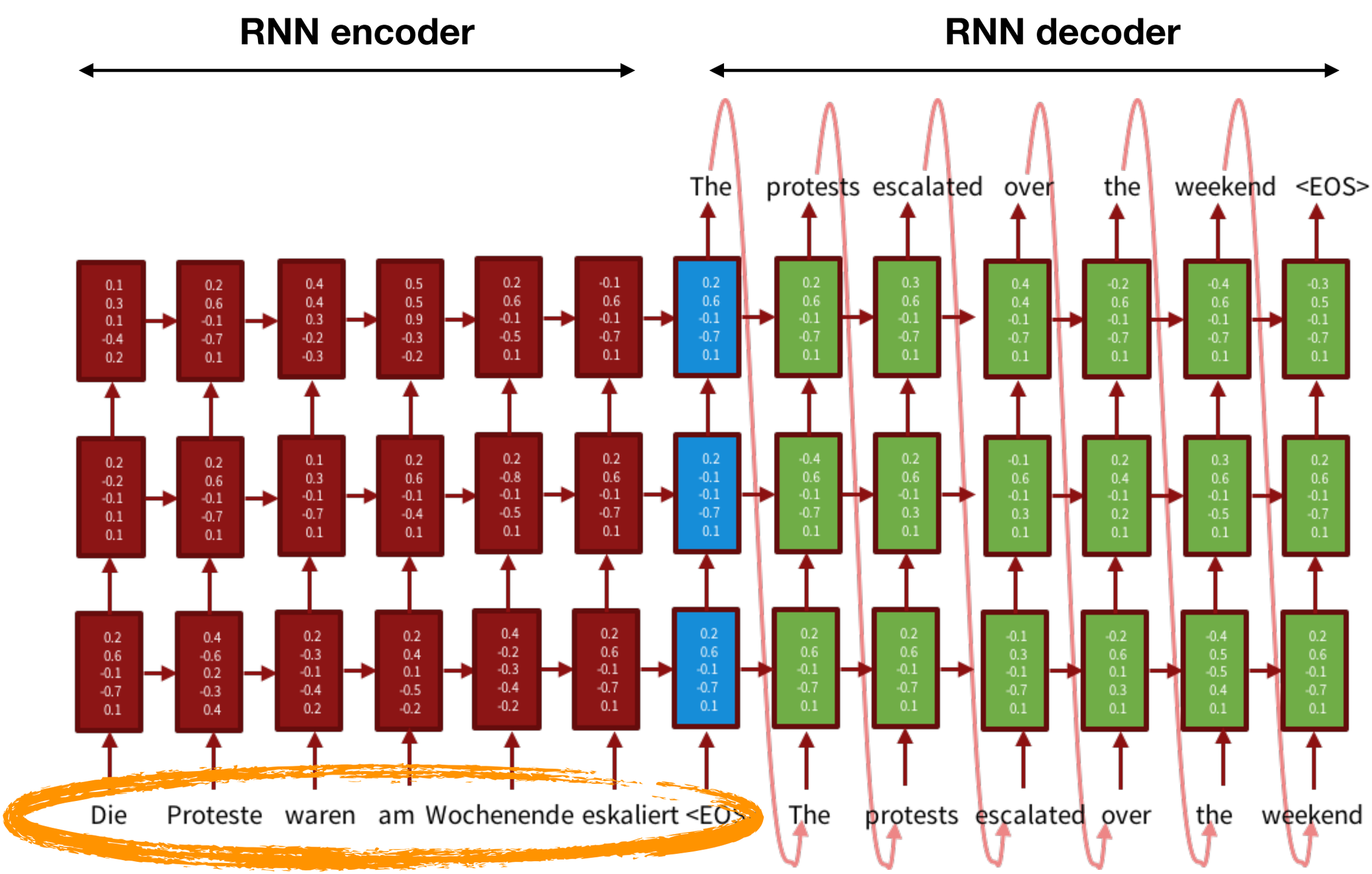


Word Embeddings!

1. Auli, M., et al. "Joint Language and Translation Modeling with Recurrent Neural Networks." *EMNLP (2013)*
2. Kalchbrenner, N., et al. “Recurrent Continuous Translation Models.” *EMNLP (2013)*
3. Cho, K., et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical MT.” *EMNLP (2014)*
4. Sutskever, I., et al. “Sequence to Sequence Learning with Neural Networks.” *NIPS (2014)*



During training, the loss function is evaluated at the decoder by the prediction loss (cross entropy) of the next word given the current word and the decoder RNN state, always feeding the decoder RNN with the true word (teacher forcing).

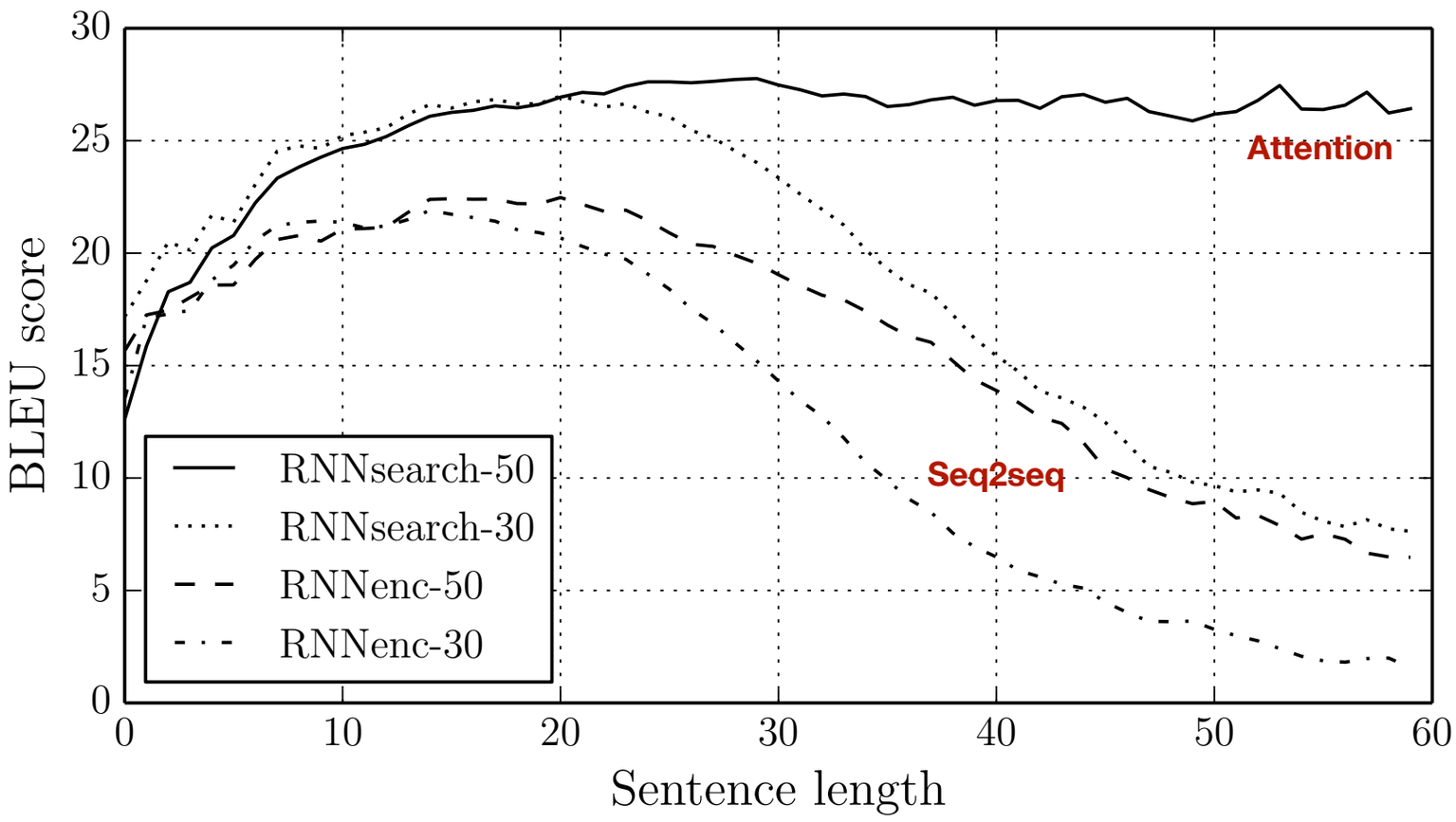


Word Embeddings!

During training, the loss function is evaluated at the decoder by the prediction loss (cross entropy) of the next word given the current word and the decoder RNN state, always feeding the decoder RNN with the true word (teacher forcing).

NMT with Seq2Seq: Limitations

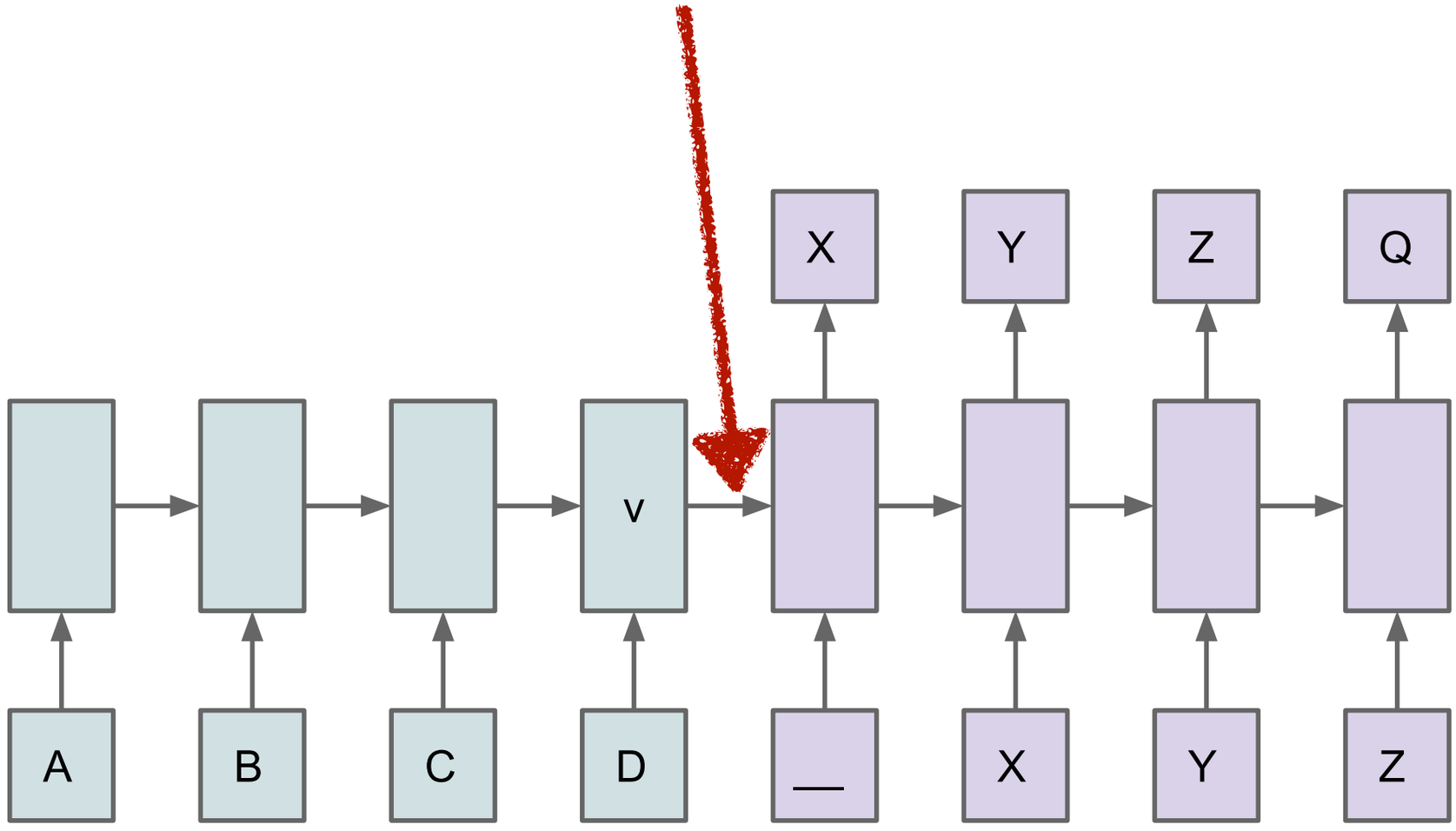
- Fixed Size Embeddings are easily overwhelmed by long inputs or long outputs
- BLEU (bilingual evaluation understudy) is an algorithm for evaluating the **quality of text which has been machine-translated** from one natural language to another.



1. Sutskever, I., et al. “Sequence to Sequence Learning with Neural Networks.” *NIPS (2014)*
2. Bahdanau, D., et al. “Neural Machine Translation by Jointly Learning to Align and Translate.” *ICLR (2015)*

Seq2Seq: The issue with long inputs

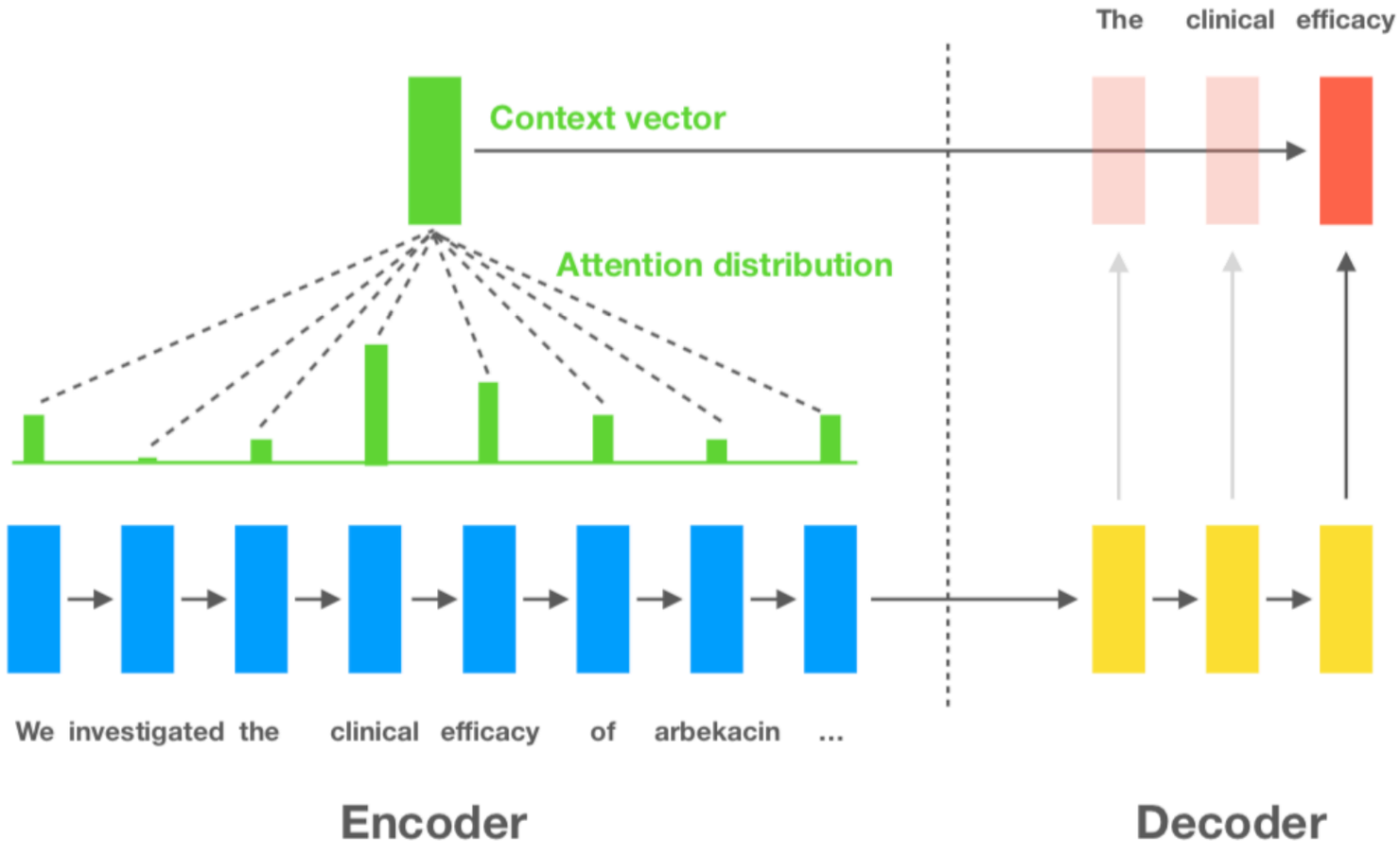
- Same embedding informs the entire output
- Needs to capture all the information about the input regardless of its length



Is there a better way to pass the information from encoder to the decoder ?

Seq2Seq with attention mechanisms

- Attention mechanisms have been shown to enhance the performance of seq2seq architectures **across almost all tasks**
- We allow the decoder to **“attend” to different parts of the source sentence** at each step of the output generation
- Each word that is generated by the decoder will be conditioned on a **unique weighted representation** of the source words



Seq2Seq with attention mechanisms

Published as a conference paper at ICLR 2015

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

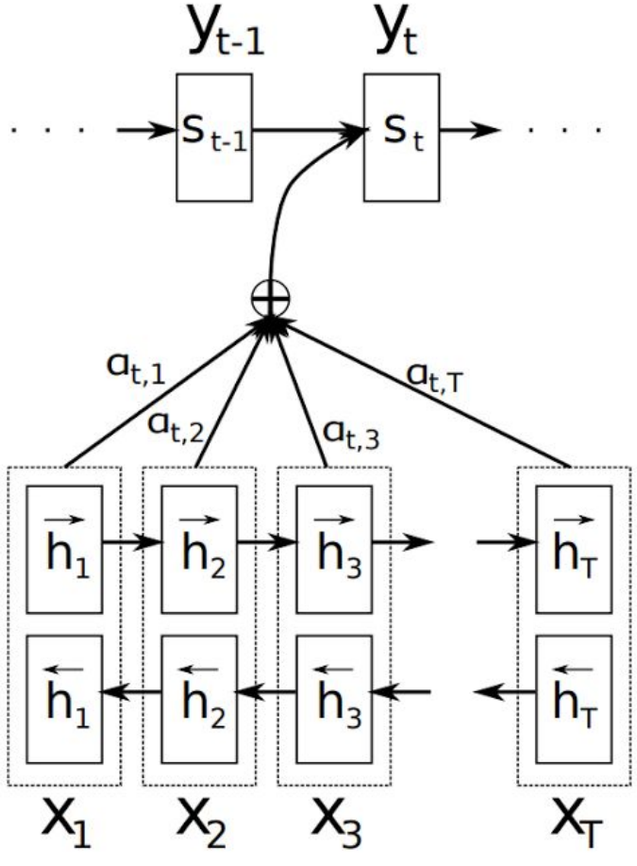
2015

Seq2Seq - Attention models

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{ik}\right)},$$

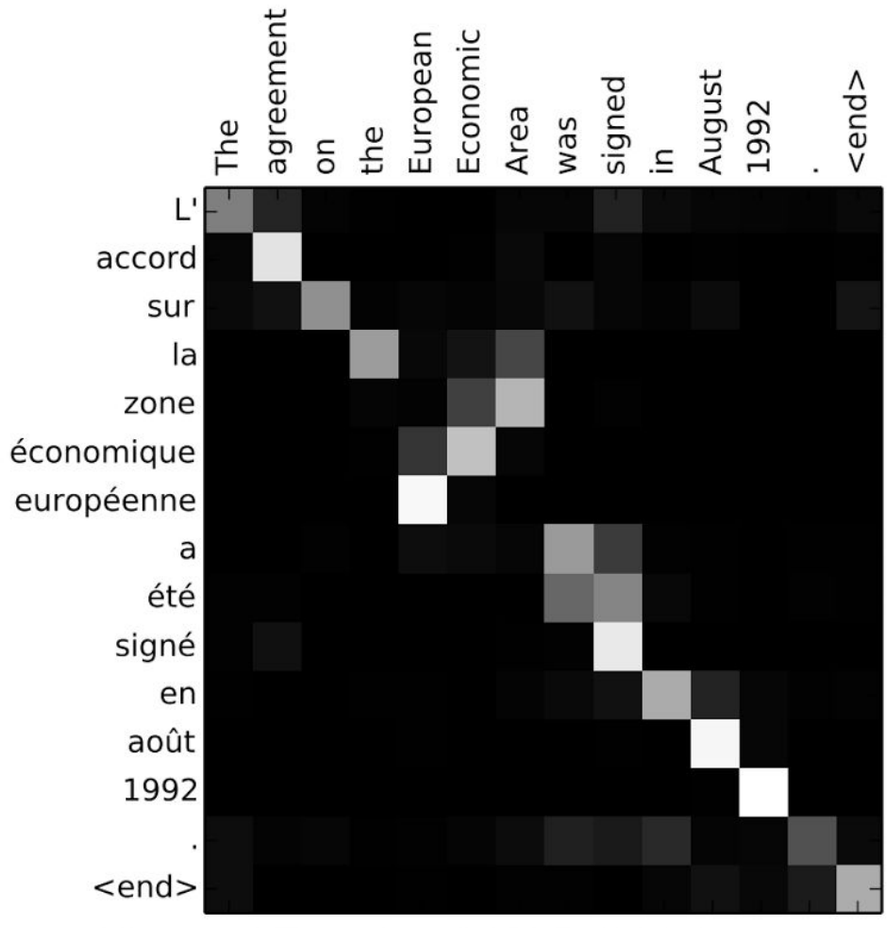
$$e_{ij} = a(s_{i-1}, h_j)$$



NEURAL MACHINE TRANSLATION
BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

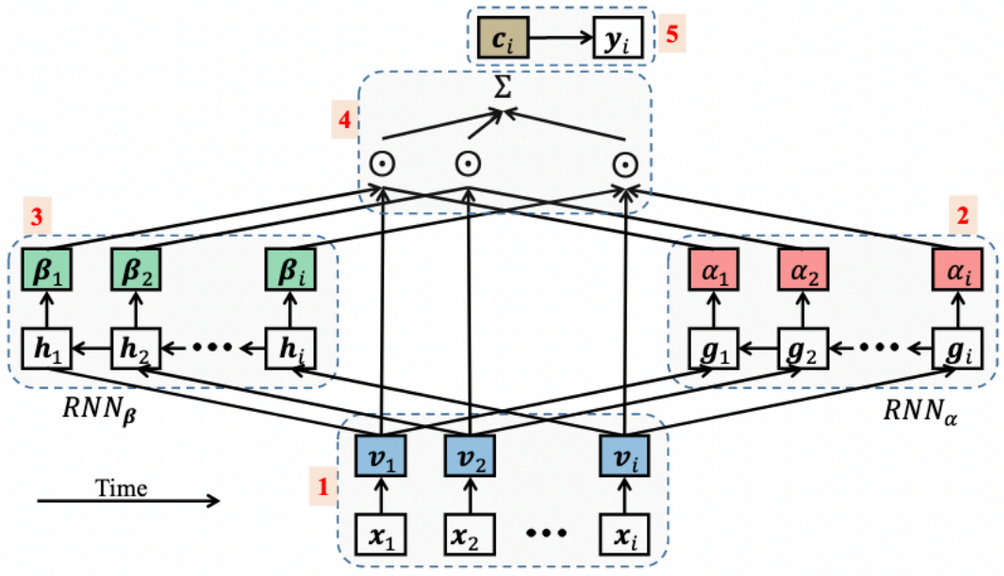


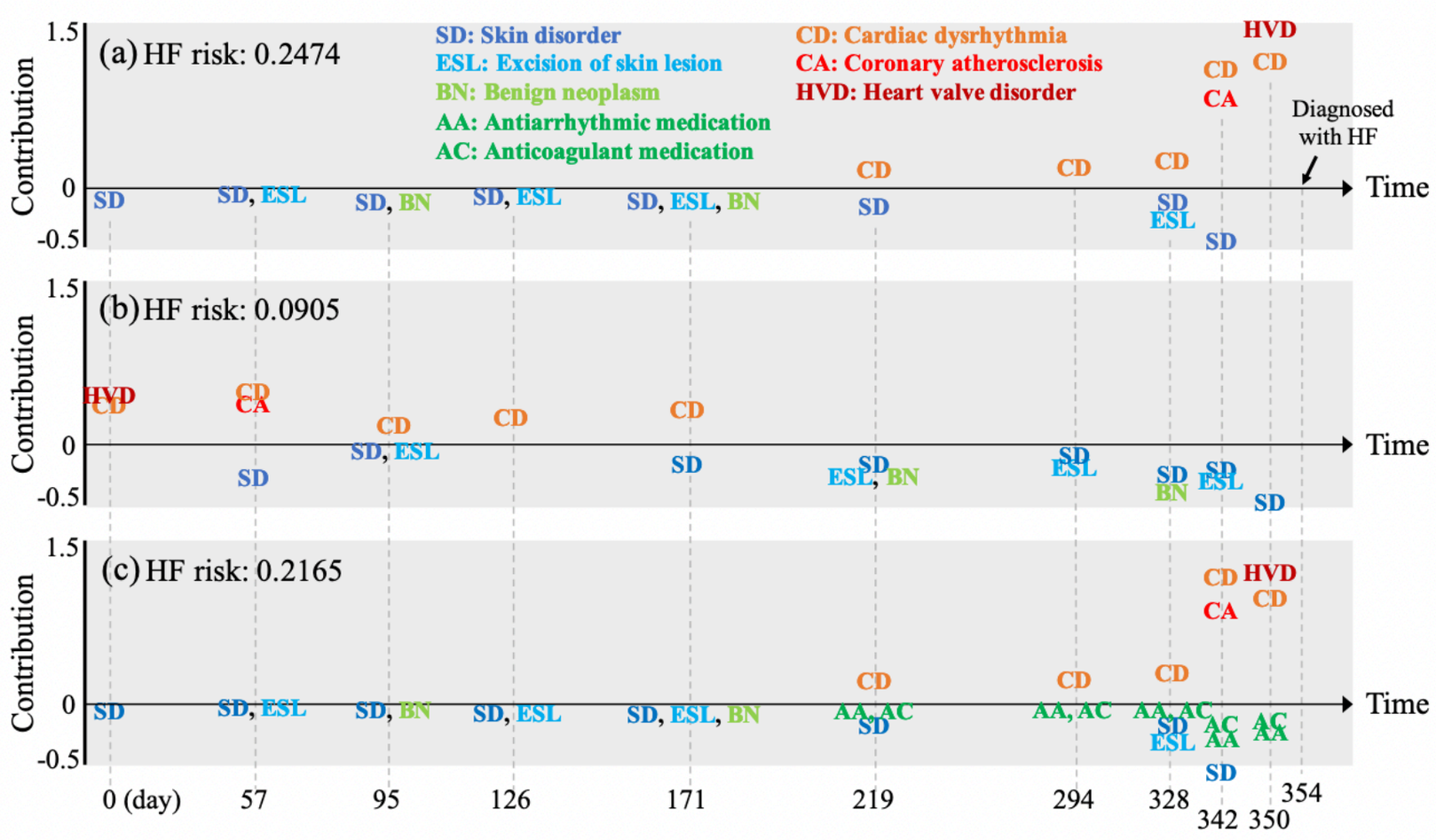
Bahdanau, D., et al. “Neural Machine Translation by Jointly Learning to Align and Translate.” *ICLR (2015)*

Attentive RNNs in health

RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism

Edward Choi*, **Mohammad Taha Bahadori***, **Joshua A. Kulas***,
Andy Schuetz†, **Walter F. Stewart**†, **Jimeng Sun***
* Georgia Institute of Technology † Sutter Health
{mp2893,bahadori,jkulas3}@gatech.edu,
{schueta1,stewarwf}@sutterhealth.org, jsun@cc.gatech.edu





Attentive RNNs in health

Attentive State-Space Modeling of Disease Progression

Ahmed M. Alaa
ECE Department
University of California, Los Angeles
ahmedmalaa@ucla.edu

Mihaela van der Schaar
University of Cambridge, and
University of California, Los Angeles
mv472@cam.ac.uk

$$p_\theta(\vec{\boldsymbol{x}}_T, \vec{\boldsymbol{z}}_T) = \prod_{t=1}^{T} \underbrace{p_\theta(x_t \mid z_t)}_{\textbf{Emission}} \underbrace{p_\theta(z_t \mid \vec{\boldsymbol{x}}_{t-1}, \vec{\boldsymbol{z}}_{t-1})}_{\textbf{Transition}},$$

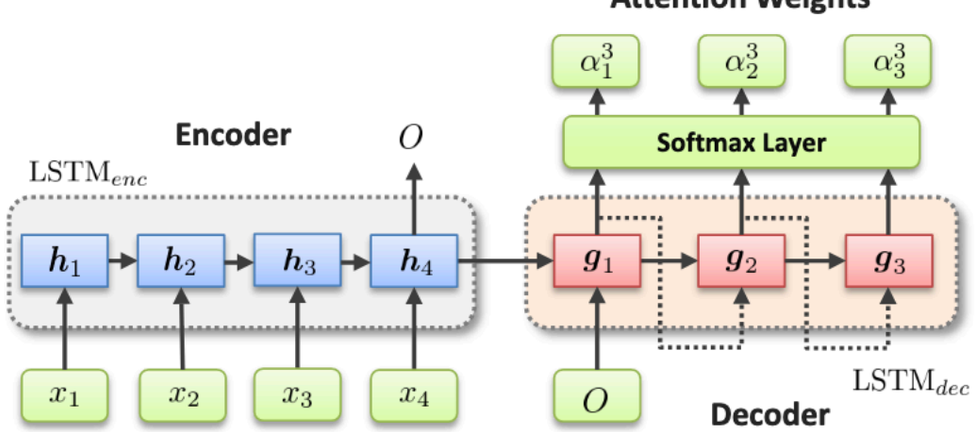


Figure 2: Seq2Seq architecture for the attention mechanism $\boldsymbol{A}$.

$$p_\theta(z_t \mid \vec{\boldsymbol{x}}_{t-1}, \vec{\boldsymbol{z}}_{t-1}) = p_\theta(z_t \mid \vec{\boldsymbol{\alpha}}_t, \vec{\boldsymbol{z}}_{t-1}),$$

$$p_\theta(z_t \mid \vec{\boldsymbol{\alpha}}_t, \vec{\boldsymbol{z}}_{t-1}) = \sum_{t'=1}^{t-1} \alpha_{t'}^{t-1} \, \boldsymbol{P}(z_{t'}, z_t), \; \forall t \geq 1,$$

$$\vec{\boldsymbol{\alpha}}_t = A_t(\vec{\boldsymbol{x}}_t).$$

Attentive RNNs in health

Attentive State-Space Modeling of Disease Progression

Ahmed M. Alaa
ECE Department
University of California, Los Angeles
ahmedmalaa@ucla.edu

Mihaela van der Schaar
University of Cambridge, and
University of California, Los Angeles
mv472@cam.ac.uk

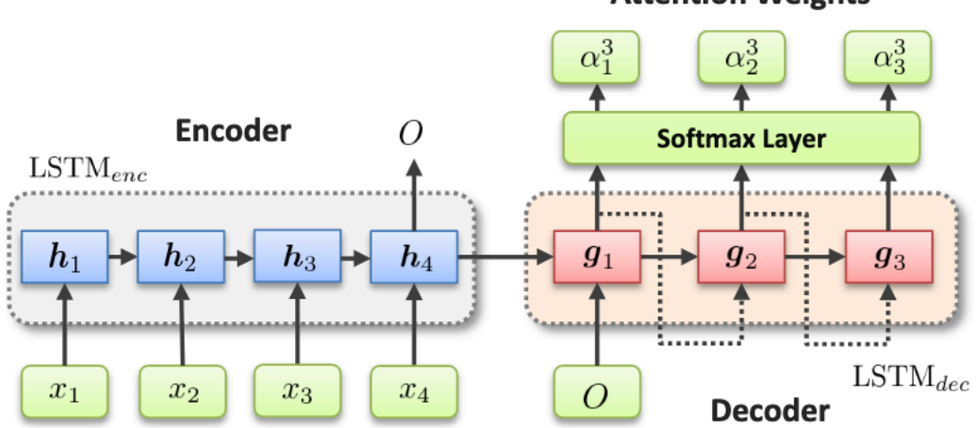


Figure 2: Seq2Seq architecture for the attention mechanism $\boldsymbol{A}$.

$$p_\theta(\vec{\boldsymbol{x}}_T, \vec{\boldsymbol{z}}_T) = \prod_{t=1}^{T} \underbrace{p_\theta(x_t \mid z_t)}_{\textbf{Emission}} \underbrace{p_\theta(z_t \mid \vec{\boldsymbol{x}}_{t-1}, \vec{\boldsymbol{z}}_{t-1})}_{\textbf{Transition}},$$

$$p_\theta(z_t \mid \vec{\boldsymbol{x}}_{t-1}, \vec{\boldsymbol{z}}_{t-1}) = p_\theta(z_t \mid \vec{\boldsymbol{\alpha}}_t, \vec{\boldsymbol{z}}_{t-1}),$$

$$p_\theta(z_t \mid \vec{\boldsymbol{\alpha}}_t, \vec{\boldsymbol{z}}_{t-1}) = \sum_{t'=1}^{t-1} \alpha_{t'}^{t-1} \boldsymbol{P}(z_{t'}, z_t), \forall t \geq 1,$$

HMM!

$$\vec{\boldsymbol{\alpha}}_t = A_t(\vec{\boldsymbol{x}}_t).$$

Transformer Networks

Transformer networks

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

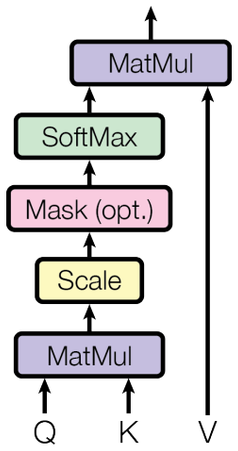
Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

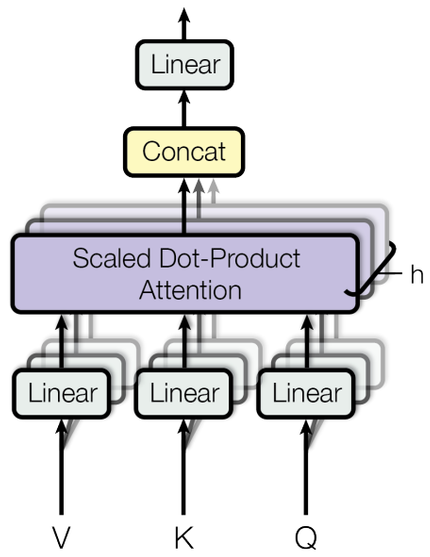
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Scaled Dot-Product Attention



Multi-Head Attention



Dec 2017

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output.

Self-attention: an introduction

A series of videos on transformers

Lennart Svensson

- We will use **slides** from Lennart Svensson, a great teacher in Chalmers (Sweden)
- Excellent **series of videos** on the topic

BERT

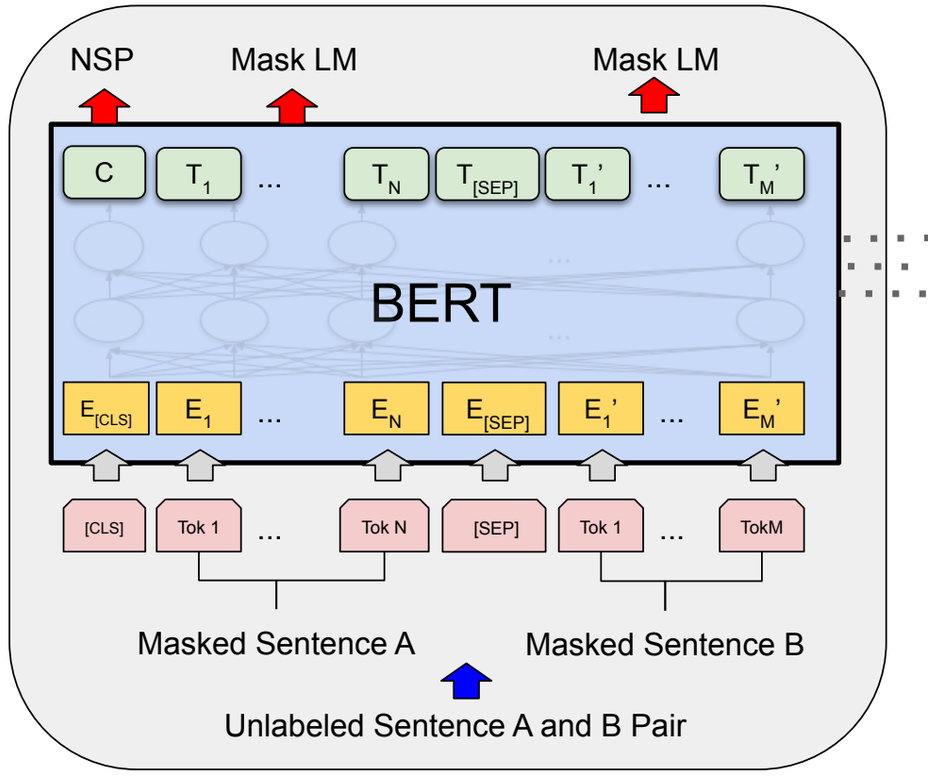
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Lennart slides here!

May 2019

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.



Pre-training

Article | Open Access | Published: 28 April 2020

BEHRT: Transformer for Electronic Health Records

Yikuan Li, Shishir Rao ✉, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi & Gholamreza Salimi-Khorshidi

Scientific Reports **10**, Article number: 7155 (2020) | Cite this article

Volume 21 Supplement 2

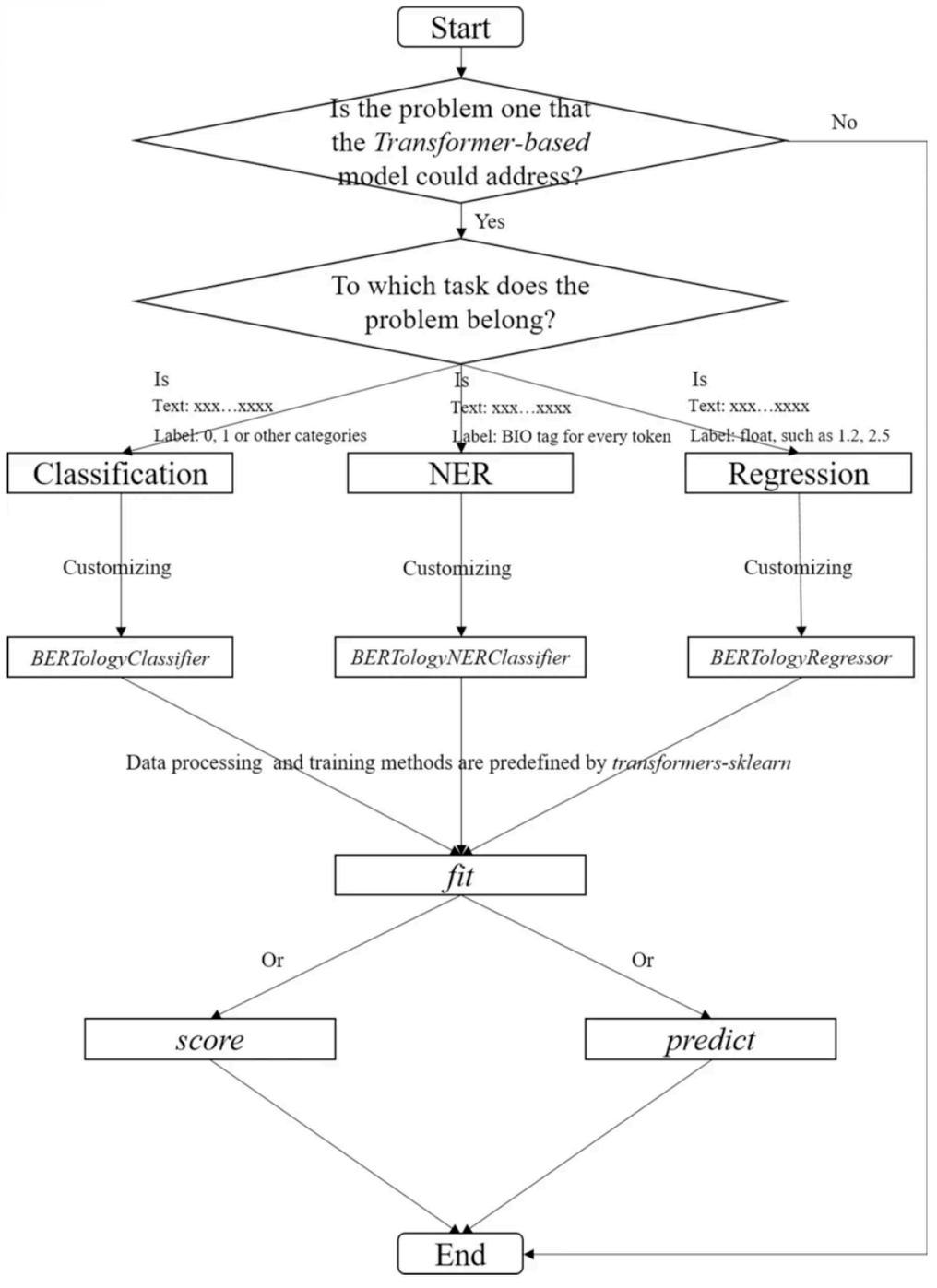
Health Big Data and Artificial Intelligence

Software | Open Access | Published: 30 July 2021

Transformers-sklearn: a toolkit for medical language understanding with *transformer-based* models

Feihong Yang, Xuwen Wang, Hetong Ma & Jiao Li

BMC Medical Informatics and Decision Making **21**, Article number: 90 (2021) | Cite this article



Publicly Available Clinical BERT Embeddings

Emily Alsentzer
Harvard-MIT
Cambridge, MA
emilya@mit.edu

John R. Murphy
MIT CSAIL
Cambridge, MA
jrmurphy@mit.edu

Willie Boag
MIT CSAIL
Cambridge, MA
wboag@mit.edu

Wei-Hung Weng
MIT CSAIL
Cambridge, MA
ckbjimmy@mit.edu

Di Jin
MIT CSAIL
Cambridge, MA
jindi15@mit.edu

Tristan Naumann
Microsoft Research
Redmond, WA
tristan@microsoft.com

Matthew B. A. McDermott
MIT CSAIL
Cambridge, MA
mmd@mit.edu

Review > J Biomed Inform. 2022 Feb;126:103982. doi: 10.1016/j.jbi.2021.103982.
Epub 2021 Dec 31.

AMMU: A survey of transformer-based biomedical pretrained language models

Katikapalli Subramanyam Kalyan [1], Ajit Rajasekharan [2], Sivanesan Sangeetha [3]

Affiliations + expand

PMID: 34974190 DOI: 10.1016/j.jbi.2021.103982

JOURNAL ARTICLE

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

Yanrong Ji, Zhihan Zhou, Han Liu ✉, Ramana V Davuluri ✉ Author Notes

Bioinformatics, Volume 37, Issue 15, 1 August 2021, Pages 2112–2120, https://doi.org/10.1093/bioinformatics/btab083

Published: 04 February 2021 **Article history** ▾

Article | Open Access | Published: 04 October 2021

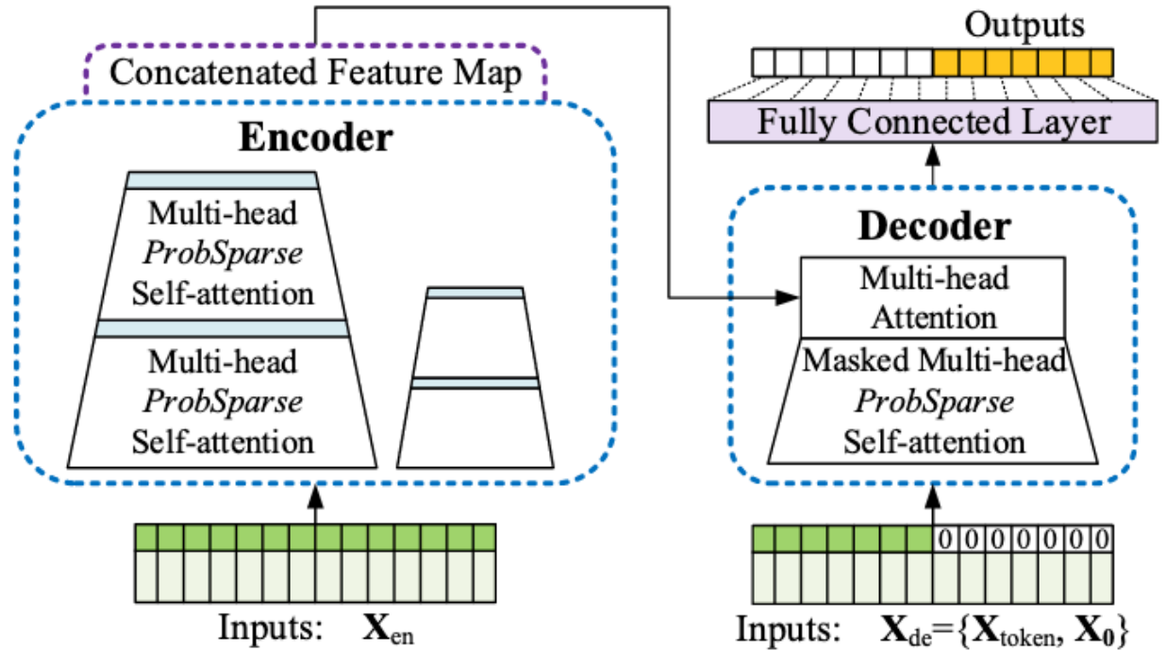
Effective gene expression prediction from sequence by integrating long-range interactions

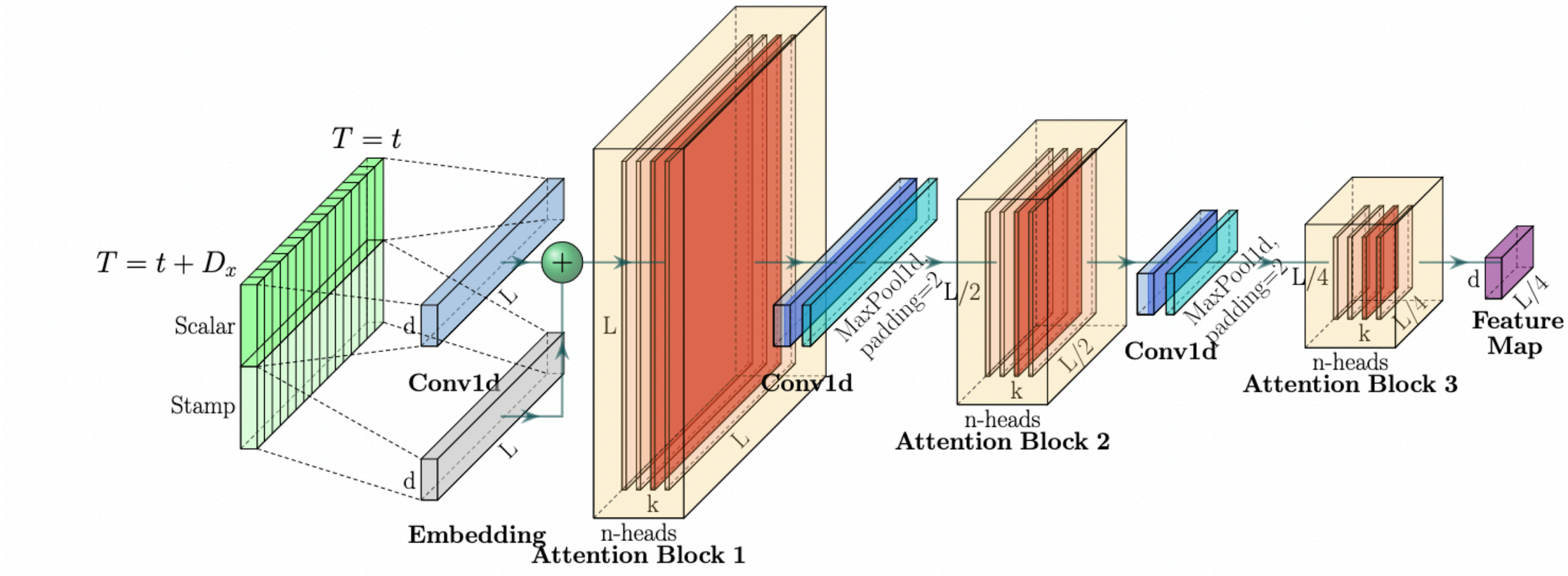
Žiga Avsec ✉, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli ✉ & David R. Kelley ✉

Nature Methods **18**, 1196–1203 (2021) | Cite this article

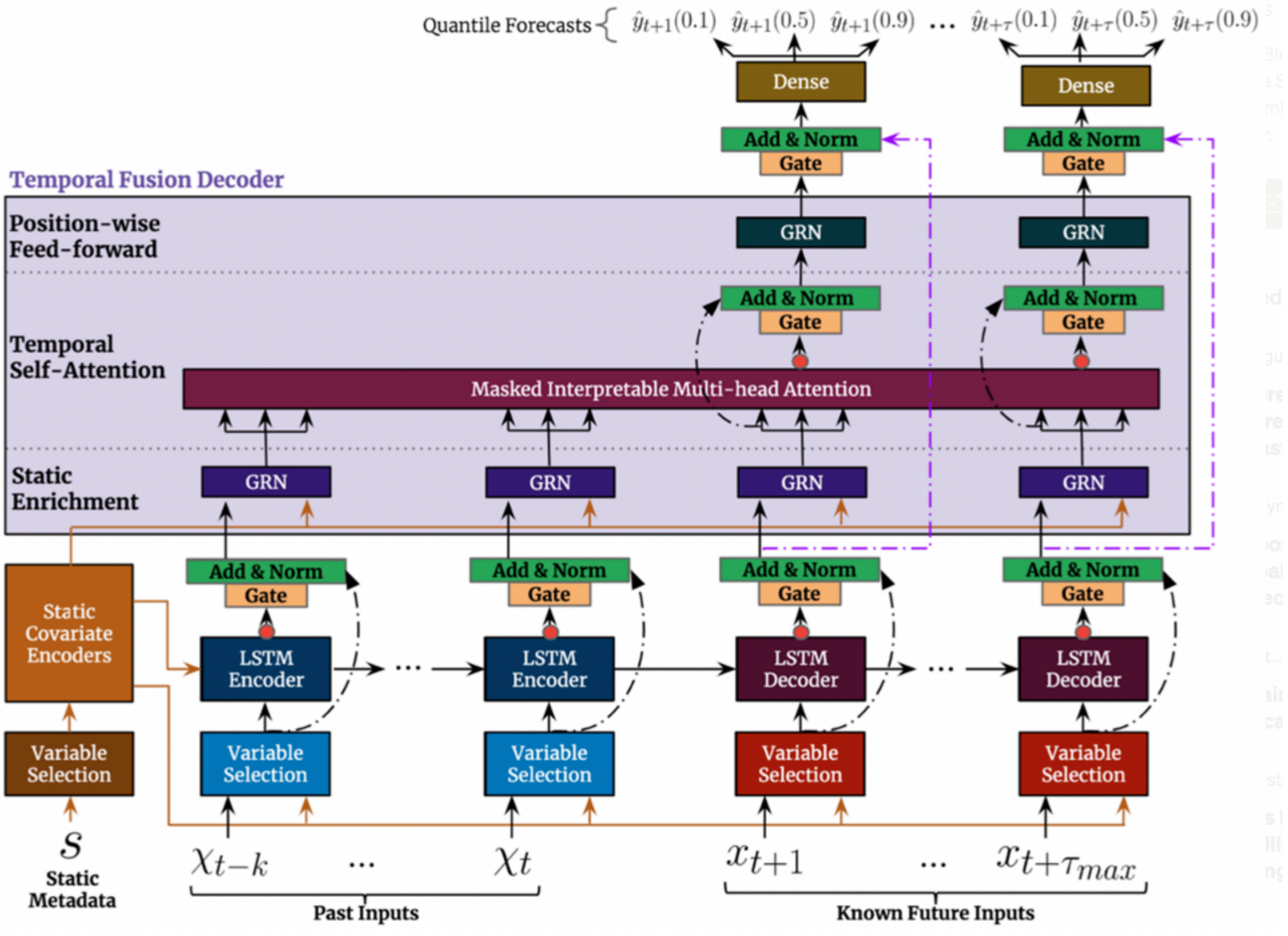
Transformer Networks for time-series forecasting

Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting





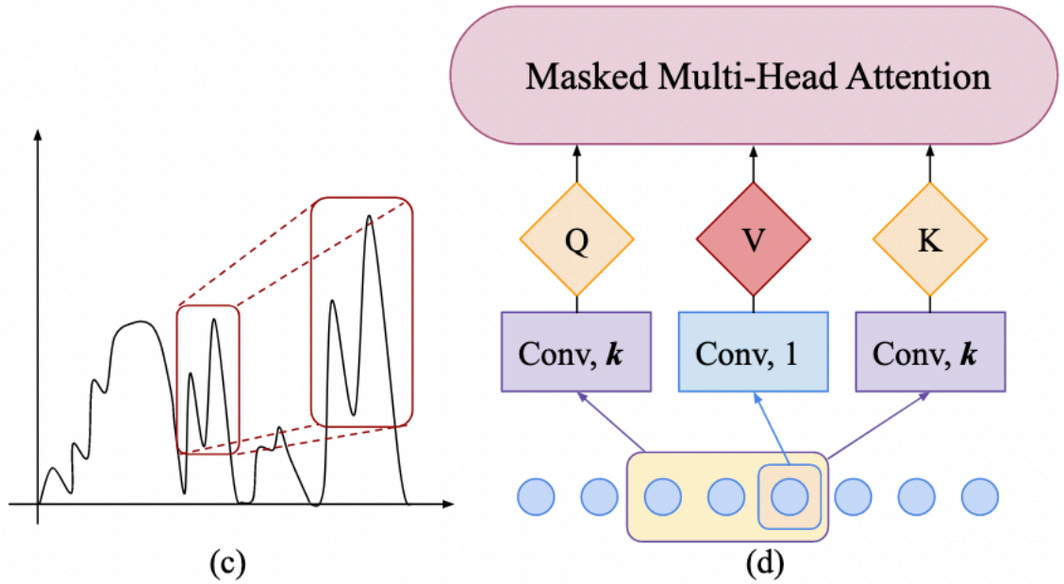
Temporal fusion transformer (Google Research)

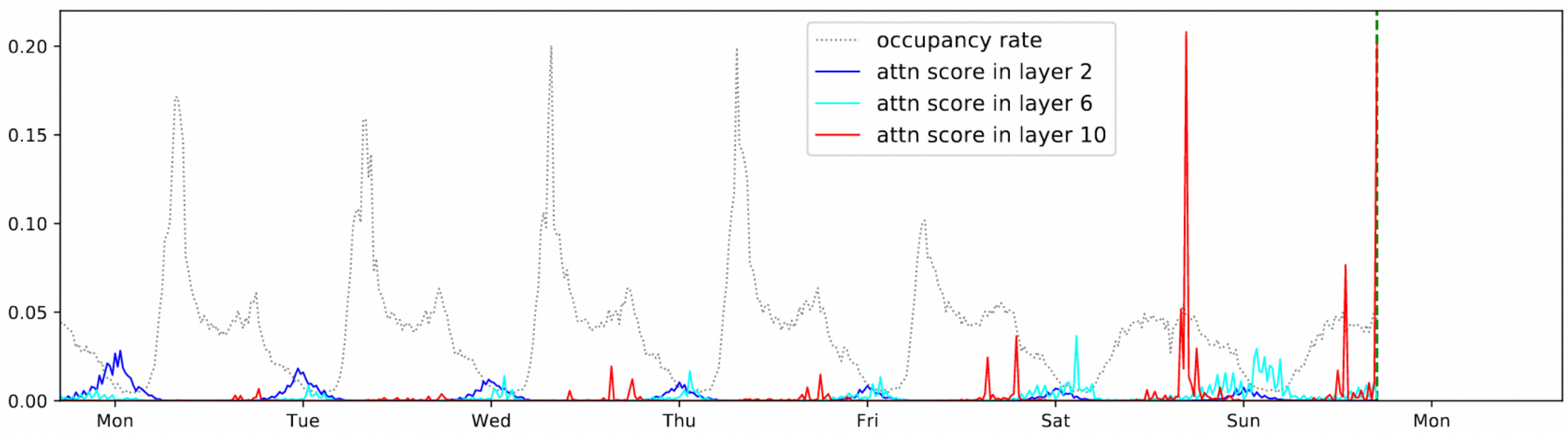


Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting

Shiyang Li shiyangli@ucsb.edu
Xiaoyong Jin x_jin@ucsb.edu
Yao Xuan yxuan@ucsb.edu
Xiyou Zhou xiyou@ucsb.edu
Wenhu Chen wenhuchen@ucsb.edu
Yu-Xiang Wang yuxiangw@cs.ucsb.edu
Xifeng Yan xyan@cs.ucsb.edu

University of California, Santa Barbara





> Comput Biol Med. 2022 Sep;148:105922. doi: 10.1016/j.compbiomed.2022.105922.
Epub 2022 Aug 2.

Static-Dynamic coordinated Transformer for Tumor Longitudinal Growth Prediction

Hexi Wang [1], Ning Xiao [1], Jina Zhang [1], Wanting Yang [1], Yulan Ma [1], Yao Suo [1], Juanjuan Zhao [2], Yan Qiang [1], Jianhong Lian [3], Qianqian Yang [4]

RESEARCH-ARTICLE OPEN ACCESS

Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series

Authors: Sindhu Tipirneni, Chandan K. Reddy Authors Info & Claims

ACM Transactions on Knowledge Discovery from Data, Volume 16, Issue 6 • December 2022 • Article No.: 105, pp 1–17 • https://doi.org/10.1145/3516367

Another revolution: **Vision Transformer**

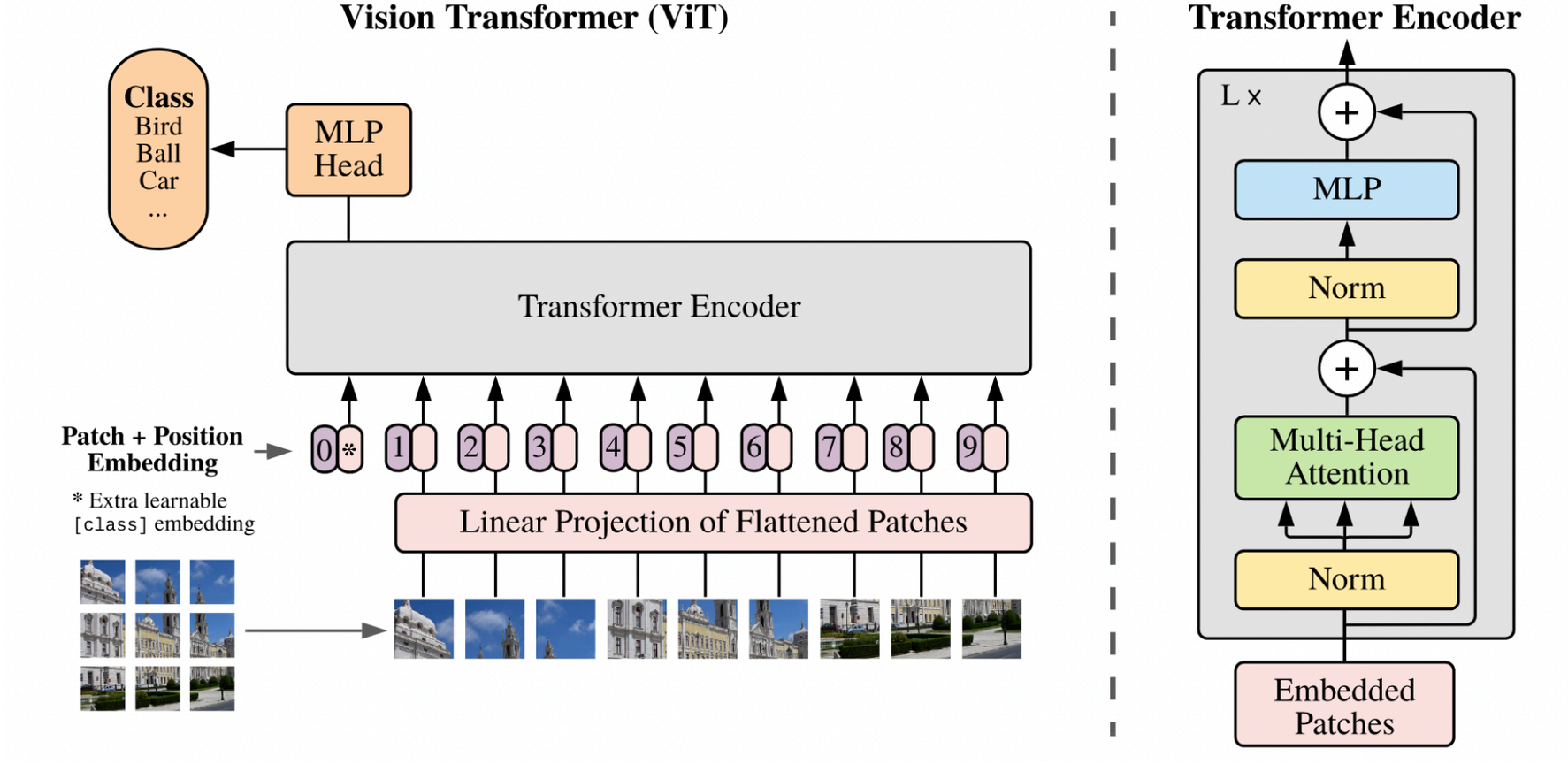
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com



Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron[1,2] Hugo Touvron[1,3] Ishan Misra[1] Hervé Jegou[1]
Julien Mairal[2] Piotr Bojanowski[1] Armand Joulin[1]

[1] Facebook AI Research [2] Inria* [3] Sorbonne University

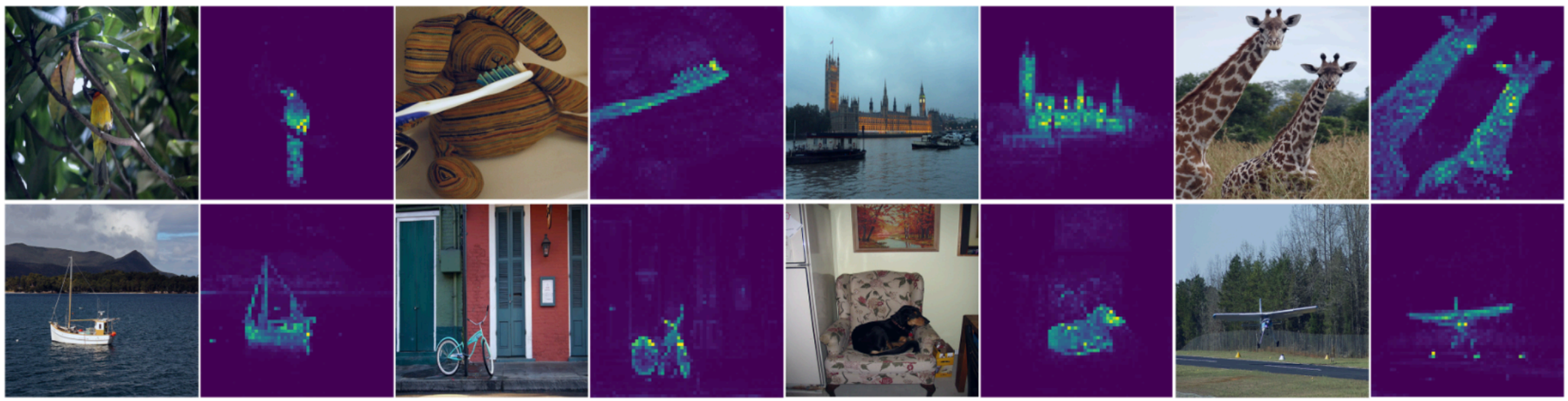


Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look at the self-attention of the `[CLS]` token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

Vision Transformer-based recognition of diabetic retinopathy grade

Jianfang Wu [1], Ruo Hu [1], Zhenghong Xiao [1], Jiaxu Chen [2], Jingwei Liu [3]

Affiliations + expand

PMID: 34693536
DOI: 10.1002/mp.15312

Vision-Language Transformer for Interpretable Pathology Visual Question Answering

Usman Naseem, Matloob Khushi, Jinman Kim

PMID: 35358054
DOI: 10.1109/JBHI.2022.3163751

A vision transformer for emphysema classification using CT images

Yanan Wu [1 2], Shouliang Qi [1 2], Yu Sun [1], Shuyue Xia [3], Yudong Yao [4], Wei Qian [5]

Affiliations + expand

PMID: 34826824
DOI: 10.1088/1361-6560/ac3dc8

Vision Transformer for femur fracture classification

Leonardo Tanzi [1], Andrea Audisio [2], Giansalvo Cirrincione [3], Alessandro Aprato [2], Enrico Vezzetti [4]

Affiliations + expand

PMID: 35469638
DOI: 10.1016/j.injury.2022.04.013