Umělá inteligence a strojové učení

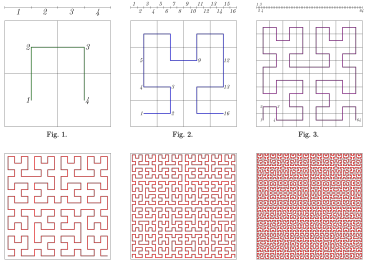
Michal Hradiš: sequences and language

Sequence processing

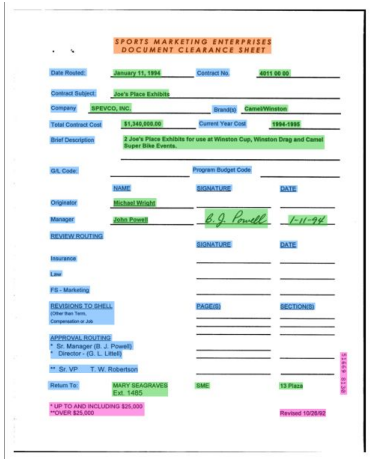
TEXT

This tool allows you to visualize the token s of token ization models of the various Google Cloud Ve are also count e d , and hover ing over them will code of this application is available on Gi t hub .

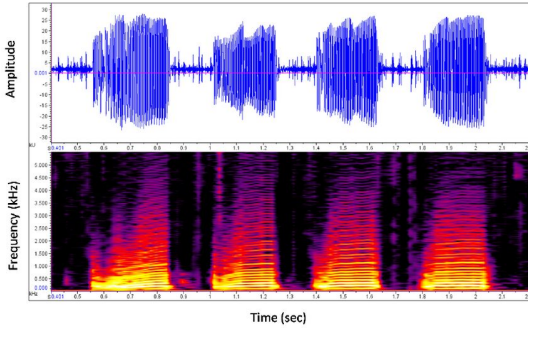
Image



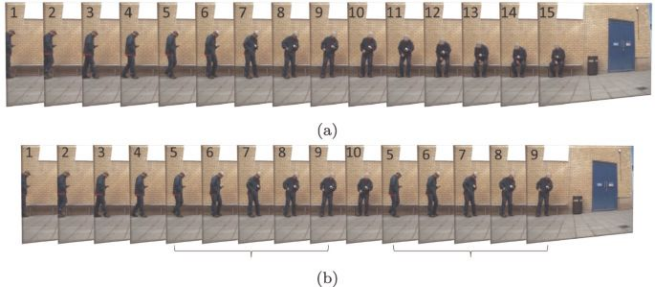
Documents



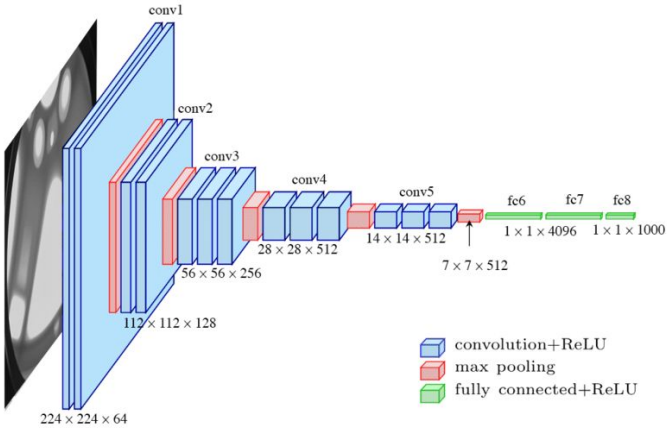
Sound

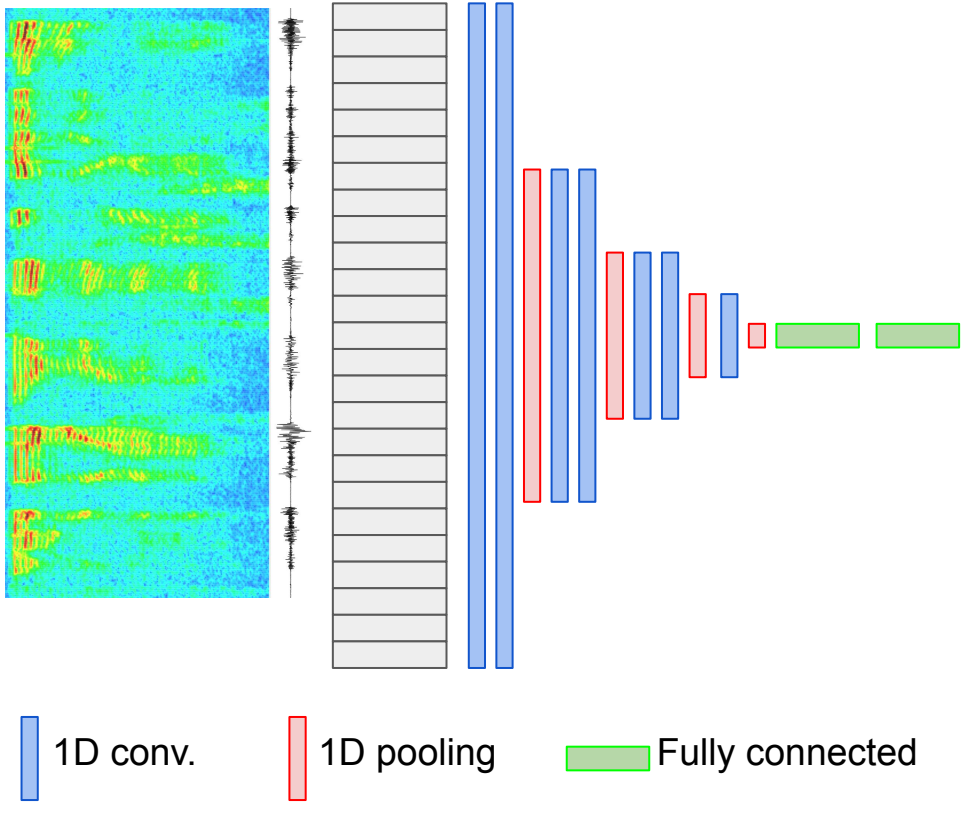


Video



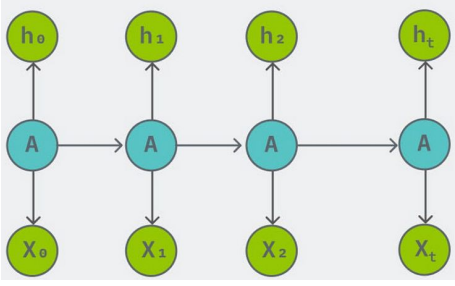
Sequences - similar to images - just 1D



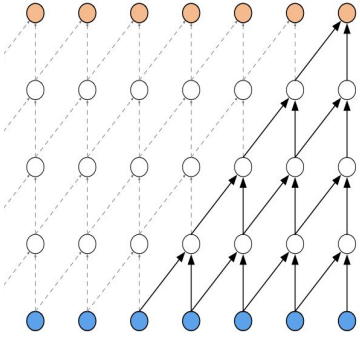


Sequence processing - communictaion

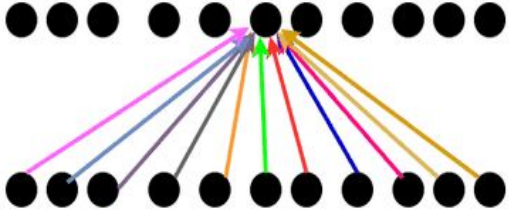
Recurrent



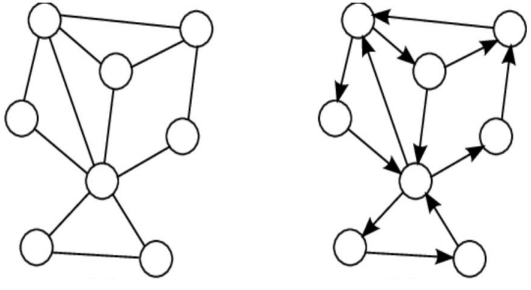
Convolution



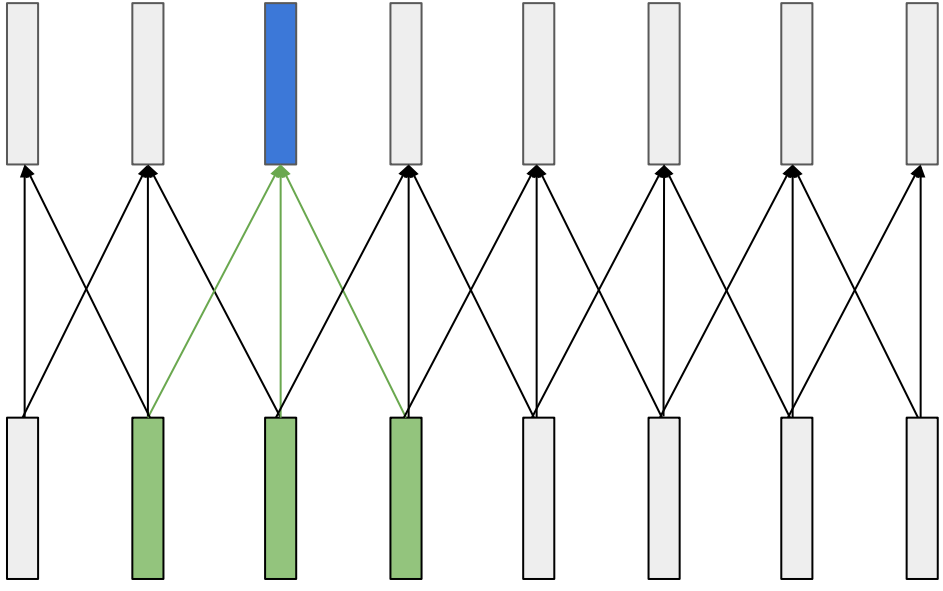
Attention



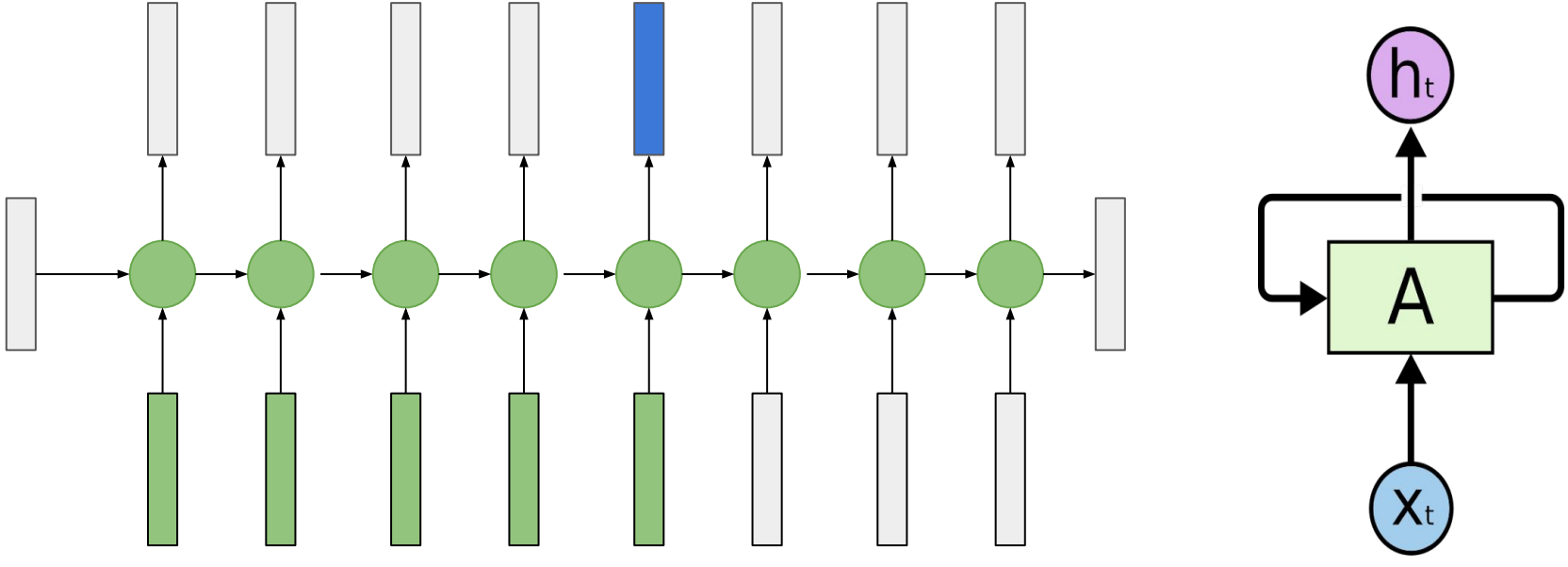
Graph networks



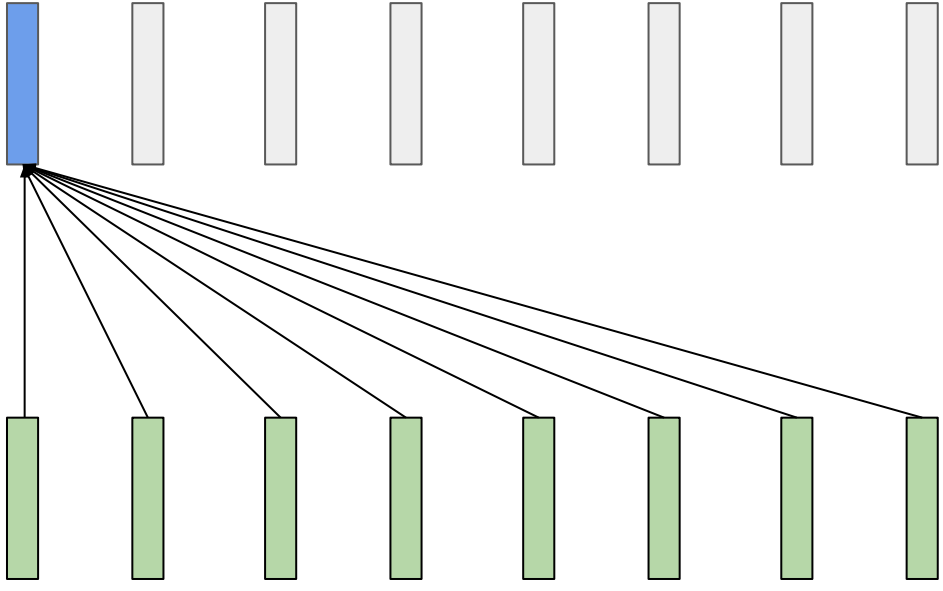
Conv. layers



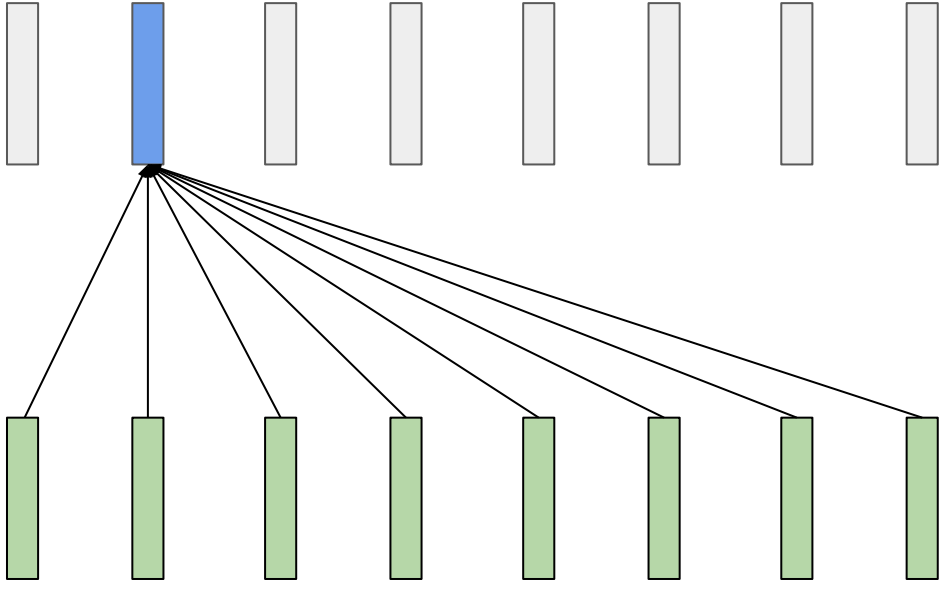
Recurrent layers



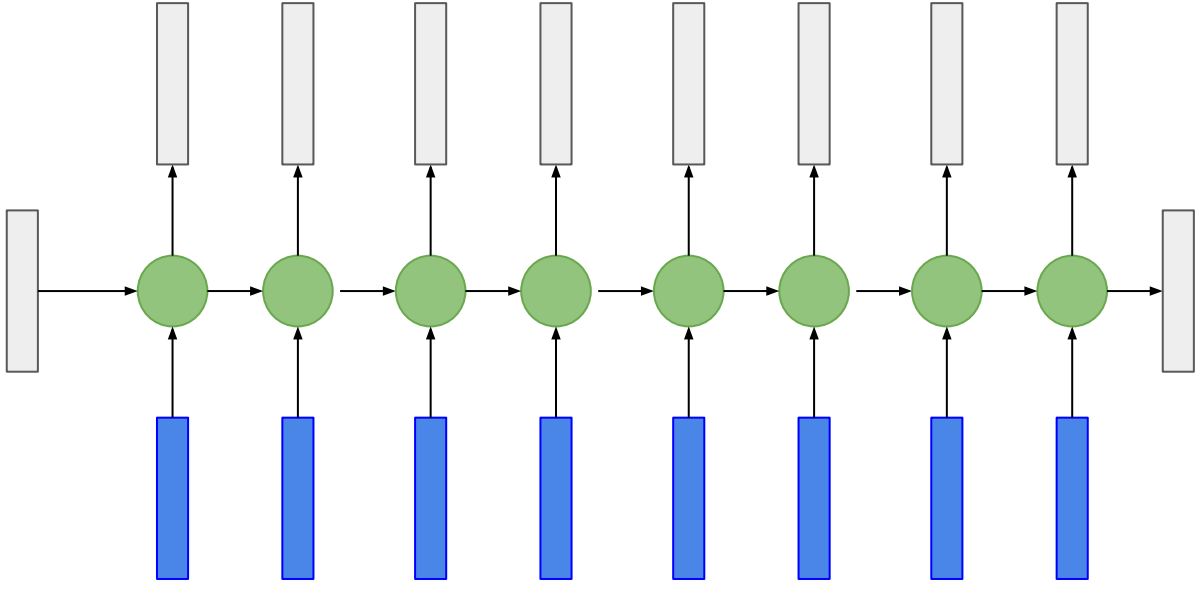
Attention layers



Attention layers

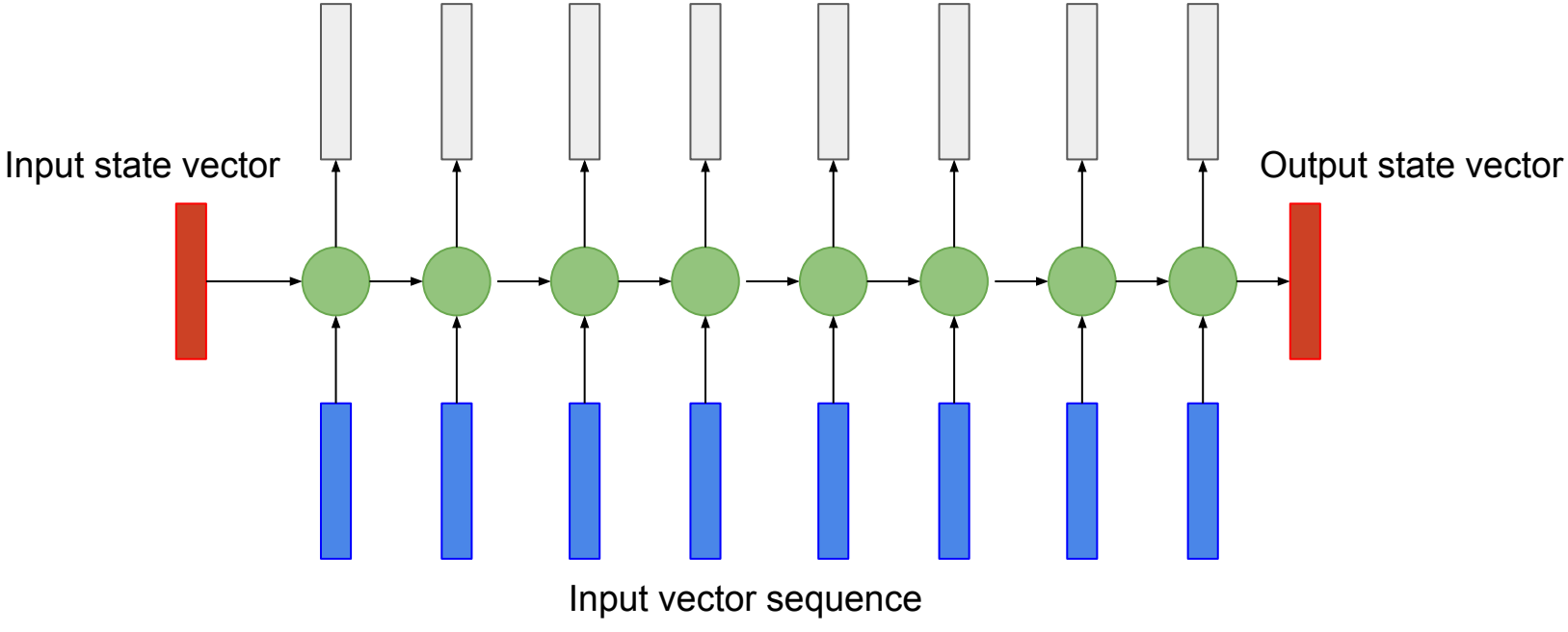


Recurrent layers

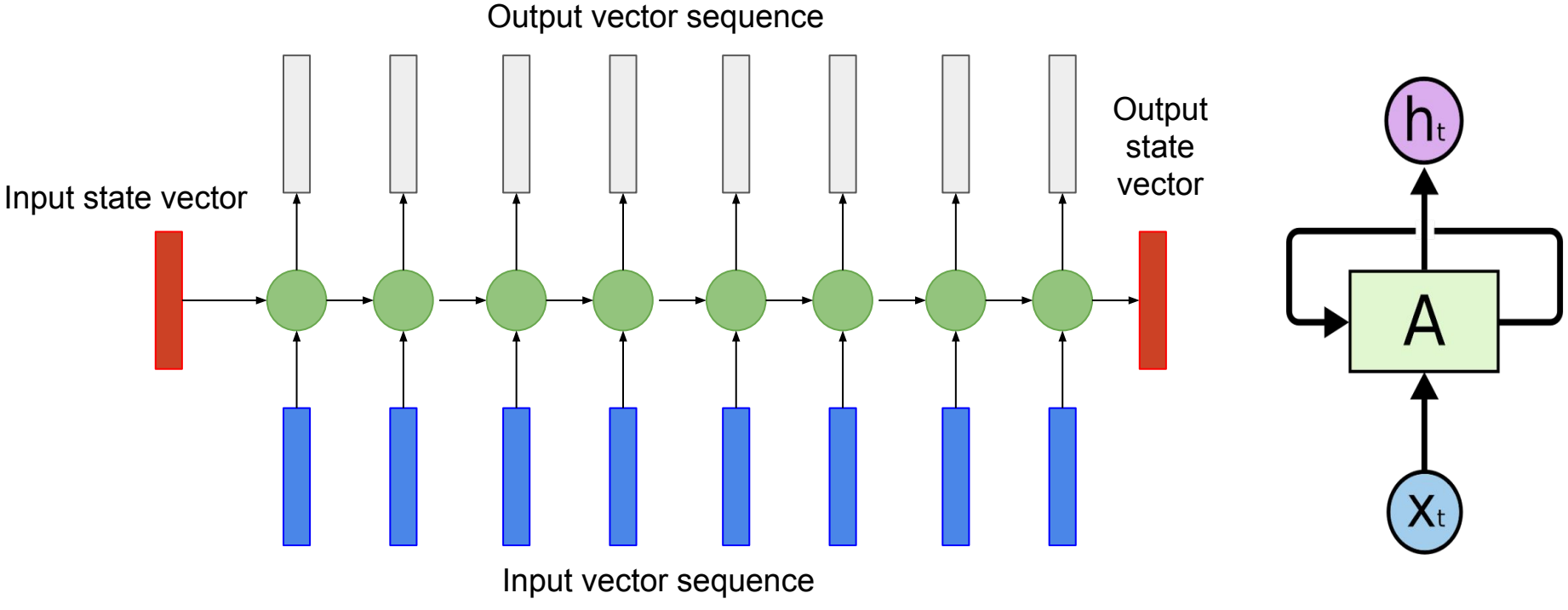


Input vector sequence

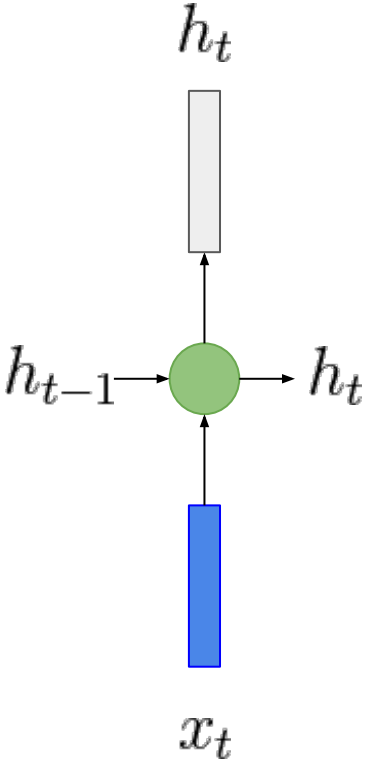
Recurrent layers



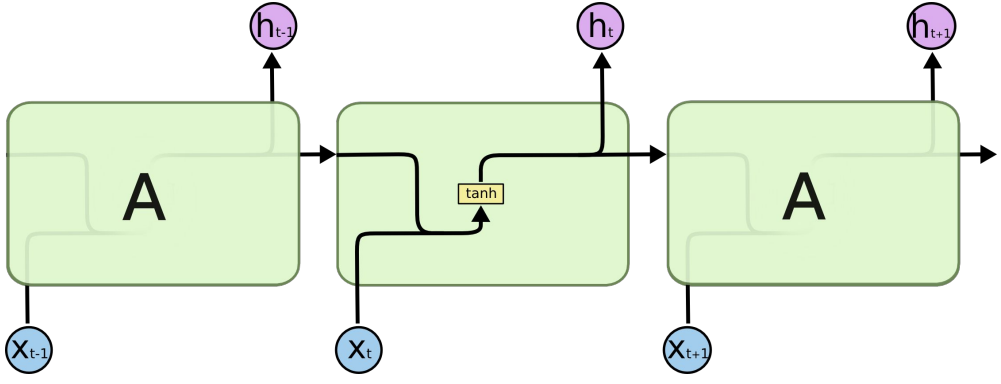
Recurrent layers



Vanila RNN

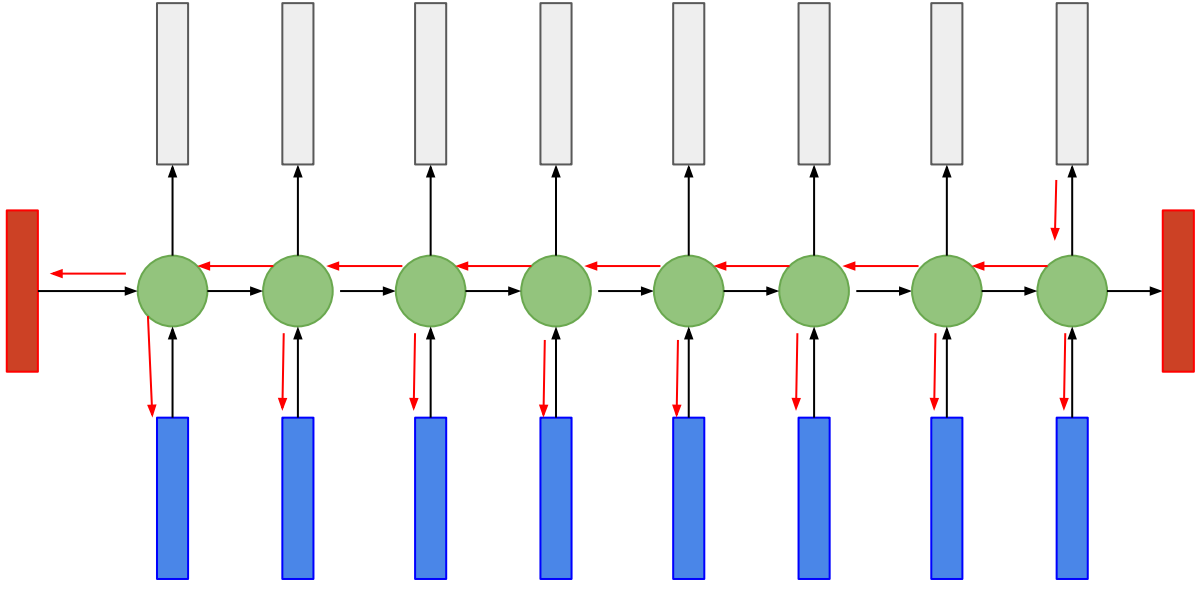


$$h_t = tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

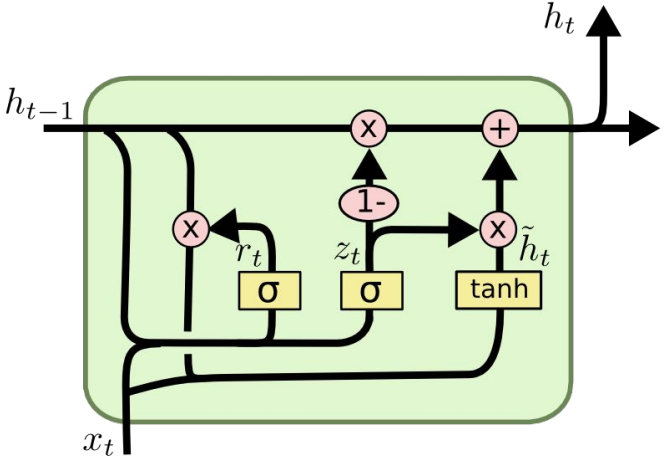


Christopher Olah: Understanding LSTM Networks.
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Recurrent layers - training



GRU - Gated Recurrent Unit /// nebo LSTM



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

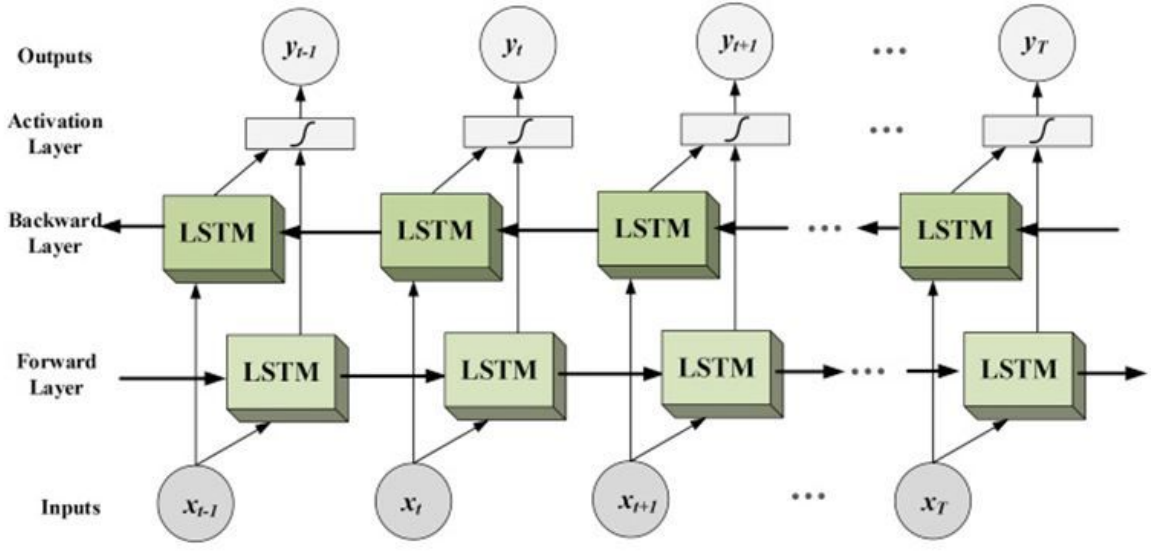
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

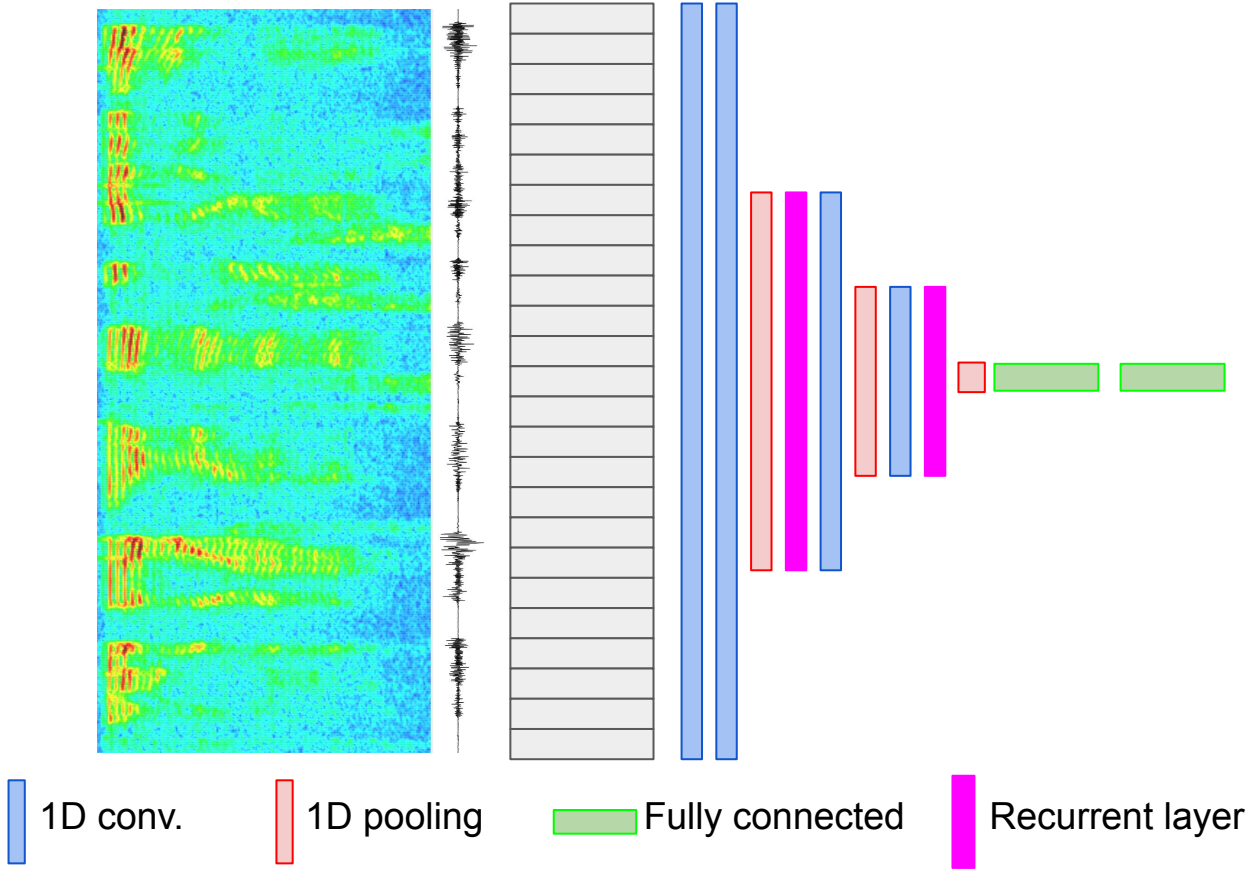
Christopher Olah: Understanding LSTM Networks.
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Bidirectional recurrent layer

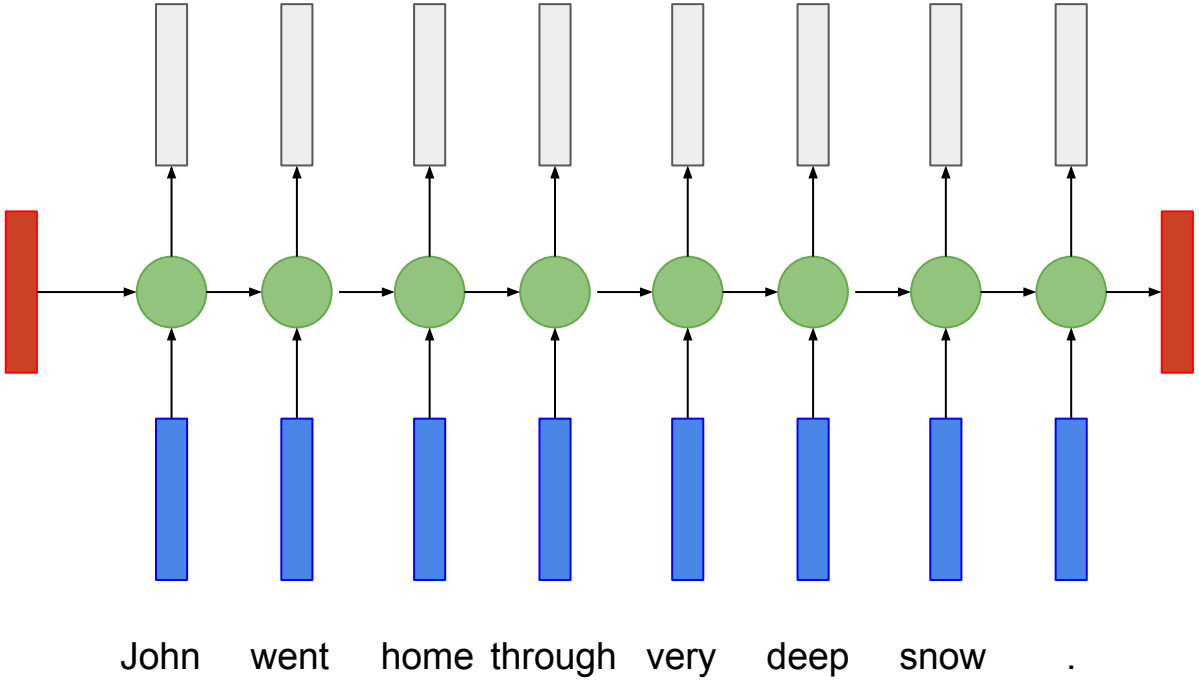


Deep Dive into Bidirectional LSTM
https://www.i2tutorials.com/technology/deep-dive-into-bidirectional-lstm/

Mix convolution and recurrent layers



Text as input?



One-hot-encoding

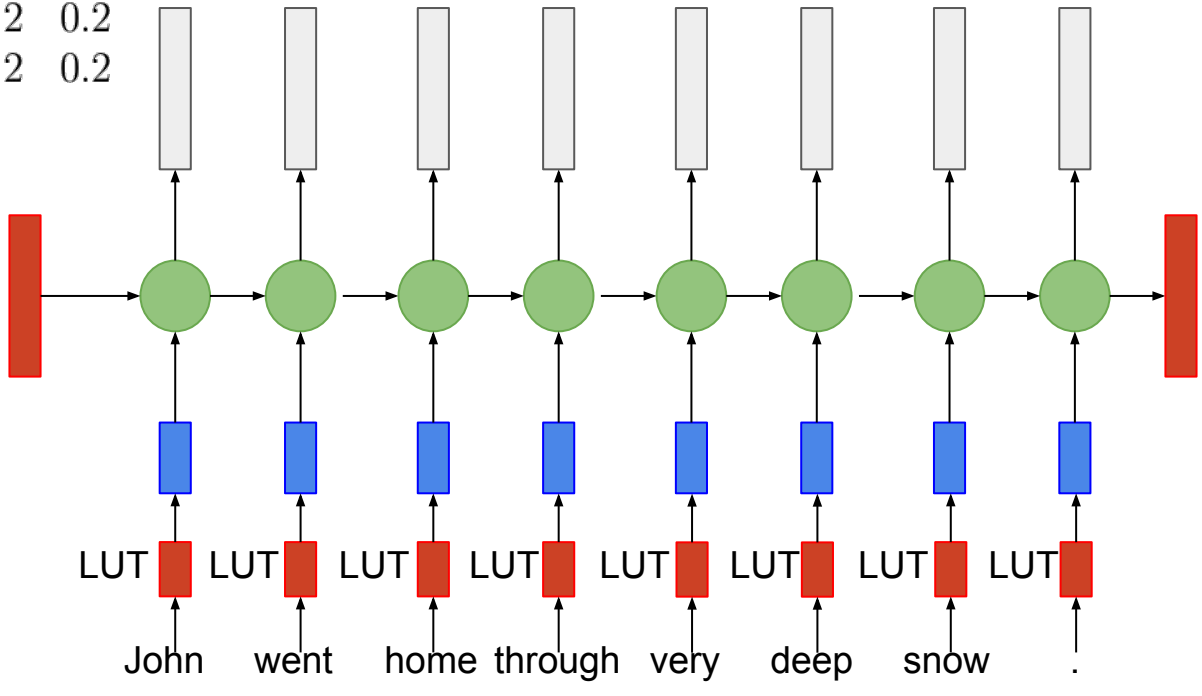
John

$$x_t = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$Wx_t \quad \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.3 & 0.2 & 1.2 & 0.2 & 1.0 & 0.0 & 0.2 & 0.2 \\ 0.8 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.3 & 0.2 & 1.2 & 0.2 & 1.0 & 0.0 & 0.2 & 0.2 \\ 0.8 & 0.2 & 0.1 & 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 1.2 \\ 0.2 \\ 0.2 \\ 1.2 \\ 0.1 \end{bmatrix}$$

Word embedding - LUT

snow	0.2	0.2	0.2	0.2	0.2	0.2	0.2
went	0.2	1.2	0.2	1.0	0.0	0.2	0.2
john	0.2	0.2	0.2	0.2	0.2	0.2	0.2
deep	0.2	0.2	0.2	0.2	0.2	0.2	0.2
very	0.2	1.2	0.2	1.0	0.0	0.2	0.2
home	0.2	0.1	0.2	0.2	0.2	0.2	0.2

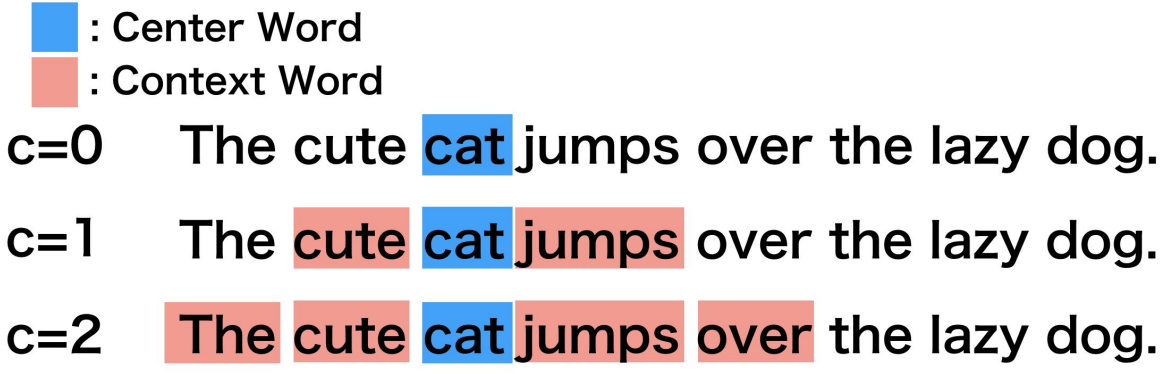


Word embeddings

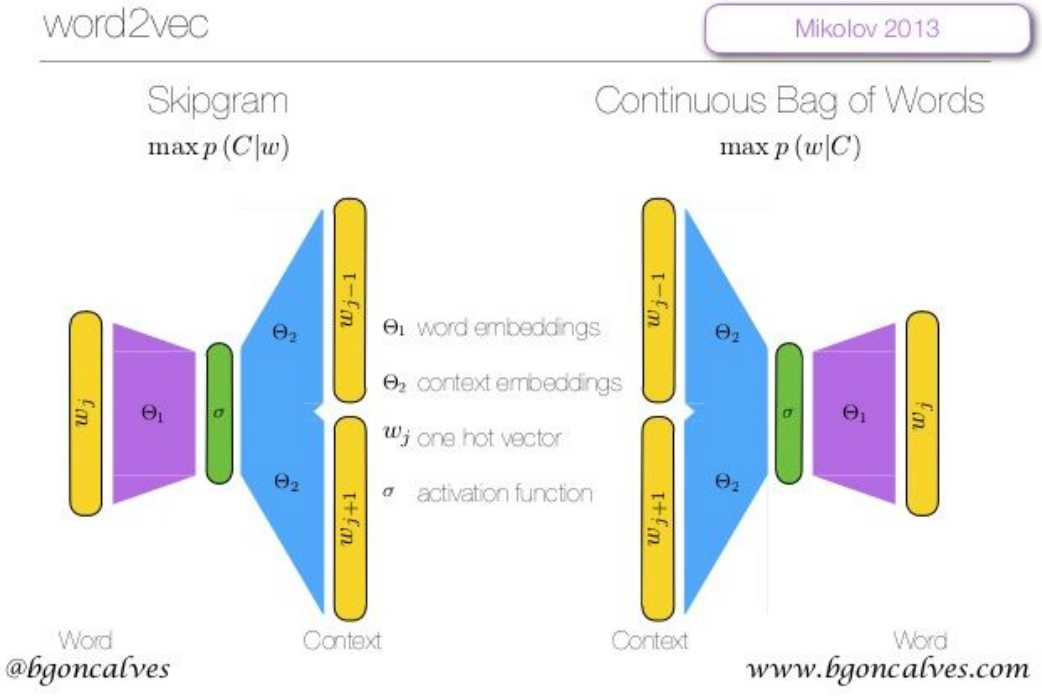
monarch [0.1, 0.5, 0, -2.5, …, -0.2, 3.0]

“monarch” close to “king”
“monarch” different to “driver”

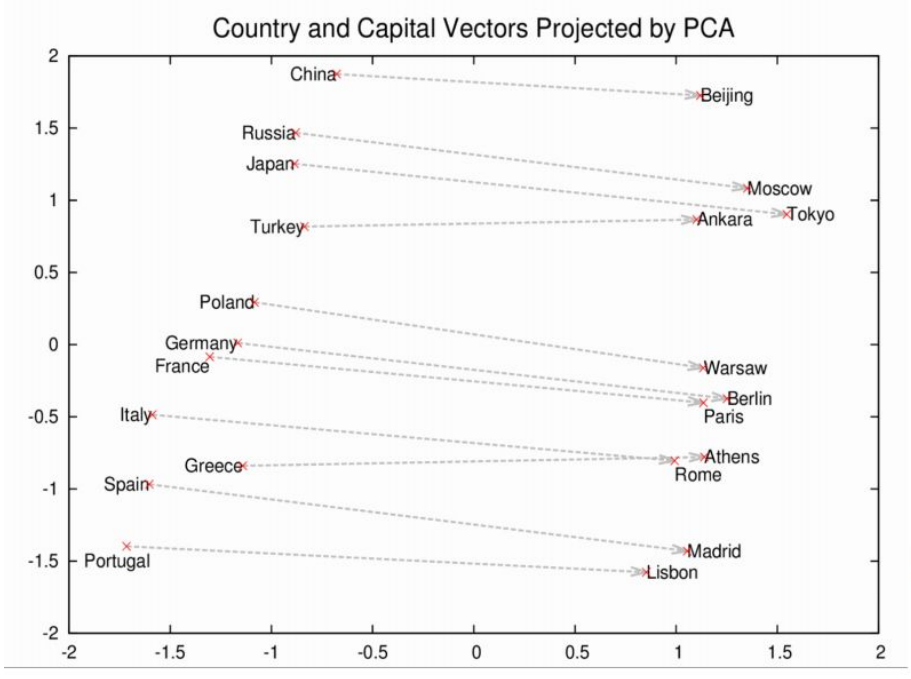
Similar words appear in similar contexts

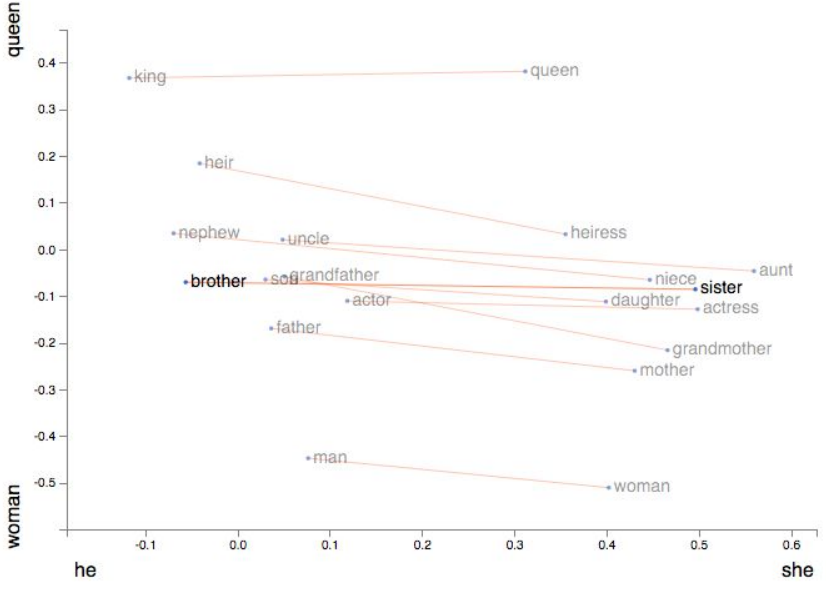


Word2Vec


word2vec
Mikolov 2013
Skipgram
$\max p(C|w)$
Continuous Bag of Words
$\max p(w|C)$
w_j
Θ_1
σ
Θ_2
w_{j-1}
Θ_2
w_{j+1}
Θ_1 word embeddings
Θ_2 context embeddings
w_j one hot vector
σ activation function
Word
Context
Context
Word
@bgoncalves
www.bgoncalves.com

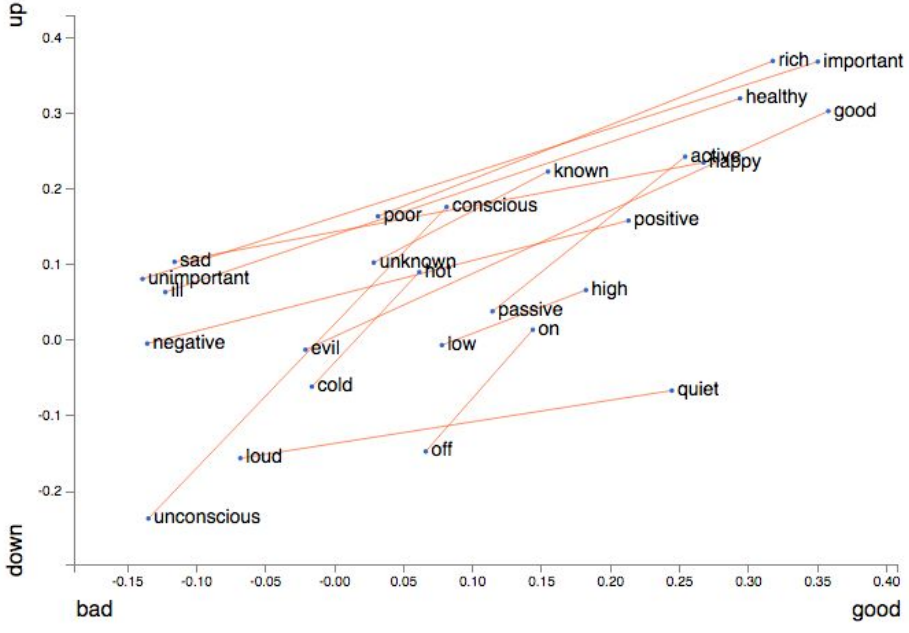
Word Vector math





https://p.migdal.pl/2017/01/06/king-man-woman-queen-why.html

Word Vector math



https://lamyiowce.github.io/word2viz/

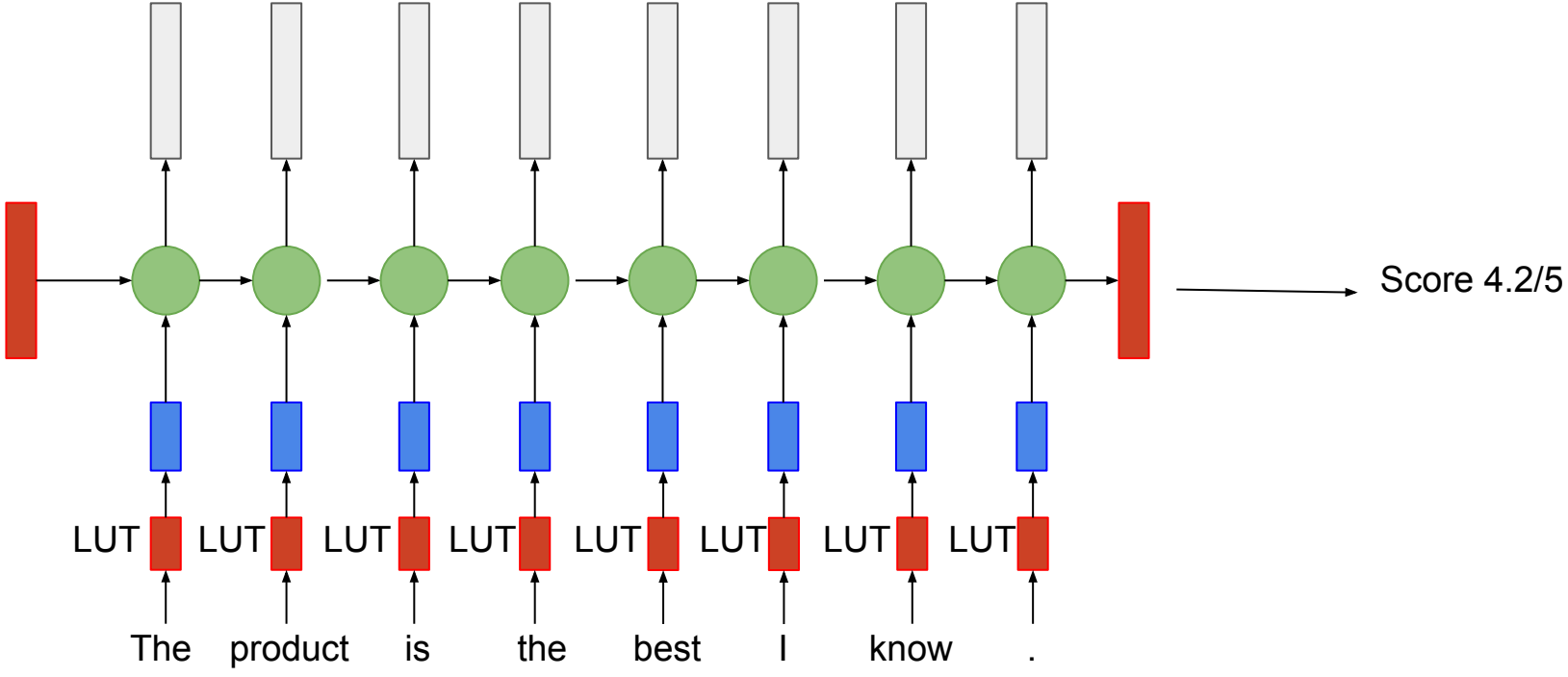
fastText

Alternative Gensim

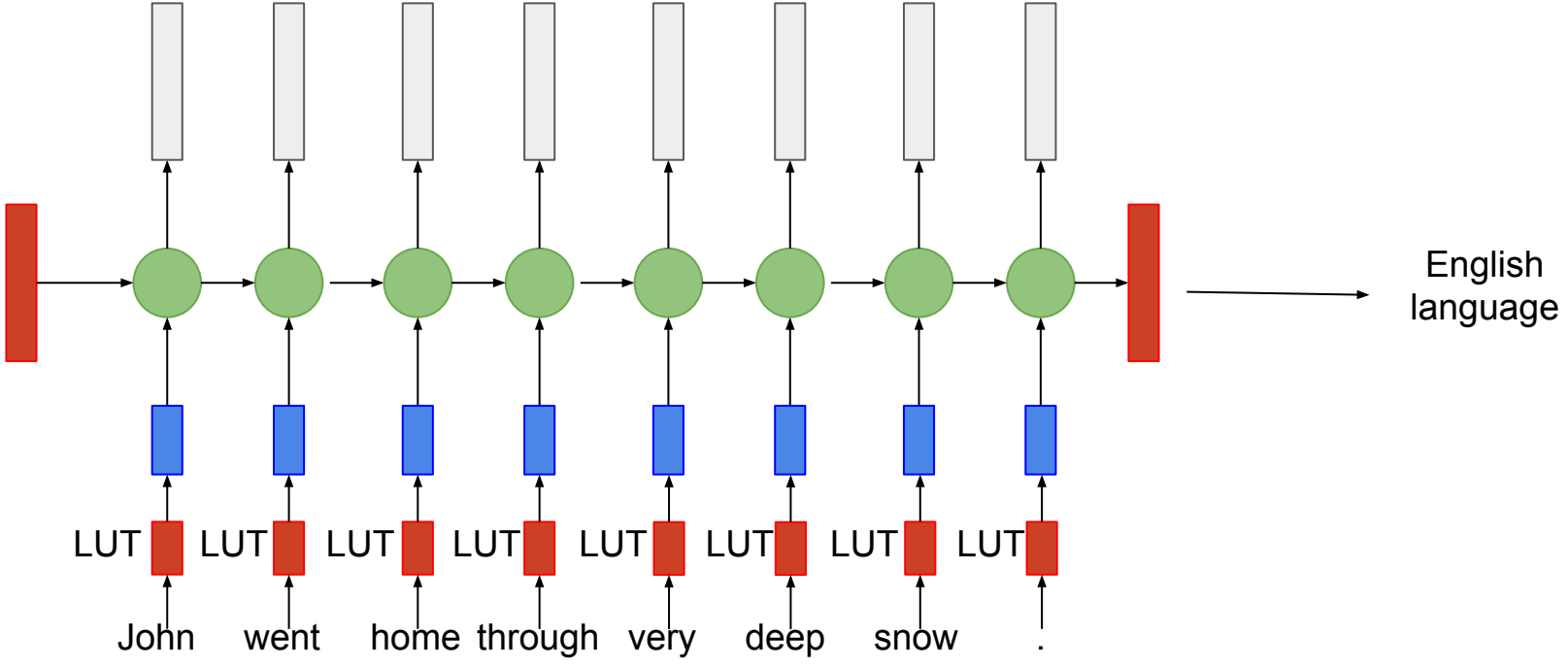
- Each word is represented as a bag of character n-grams in addition to the word itself
- Word vectors for 157 languages
- Aligned word vectors for 44 languages (Czech)
  - Joulin et al.: Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion, 2018.
- Others: WordPiece embeddings (Wu et al., 2016)

https://fasttext.cc/

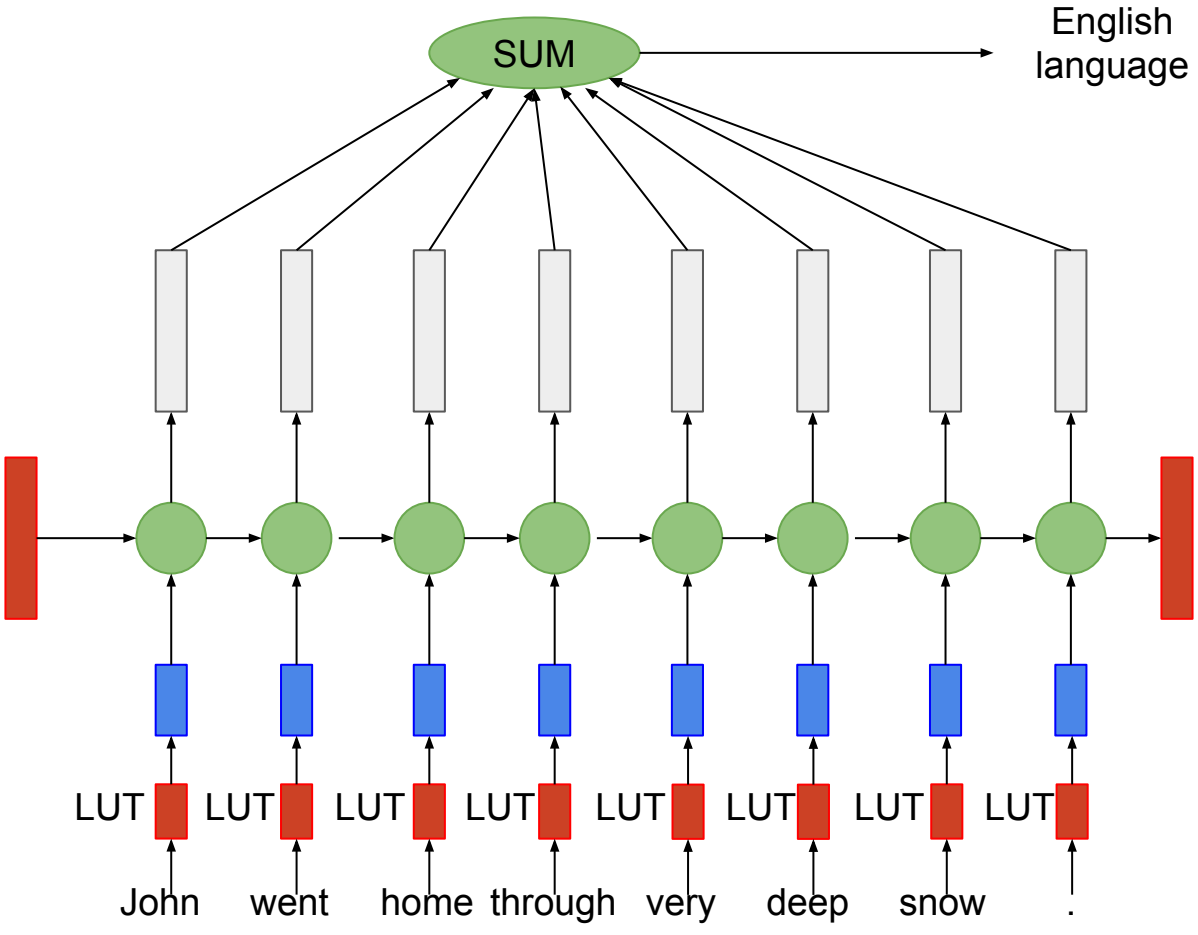
Regression


LUT
LUT
LUT
LUT
LUT
LUT
LUT
LUT
The
product
is
the
best
I
know
.
Score 4.2/5

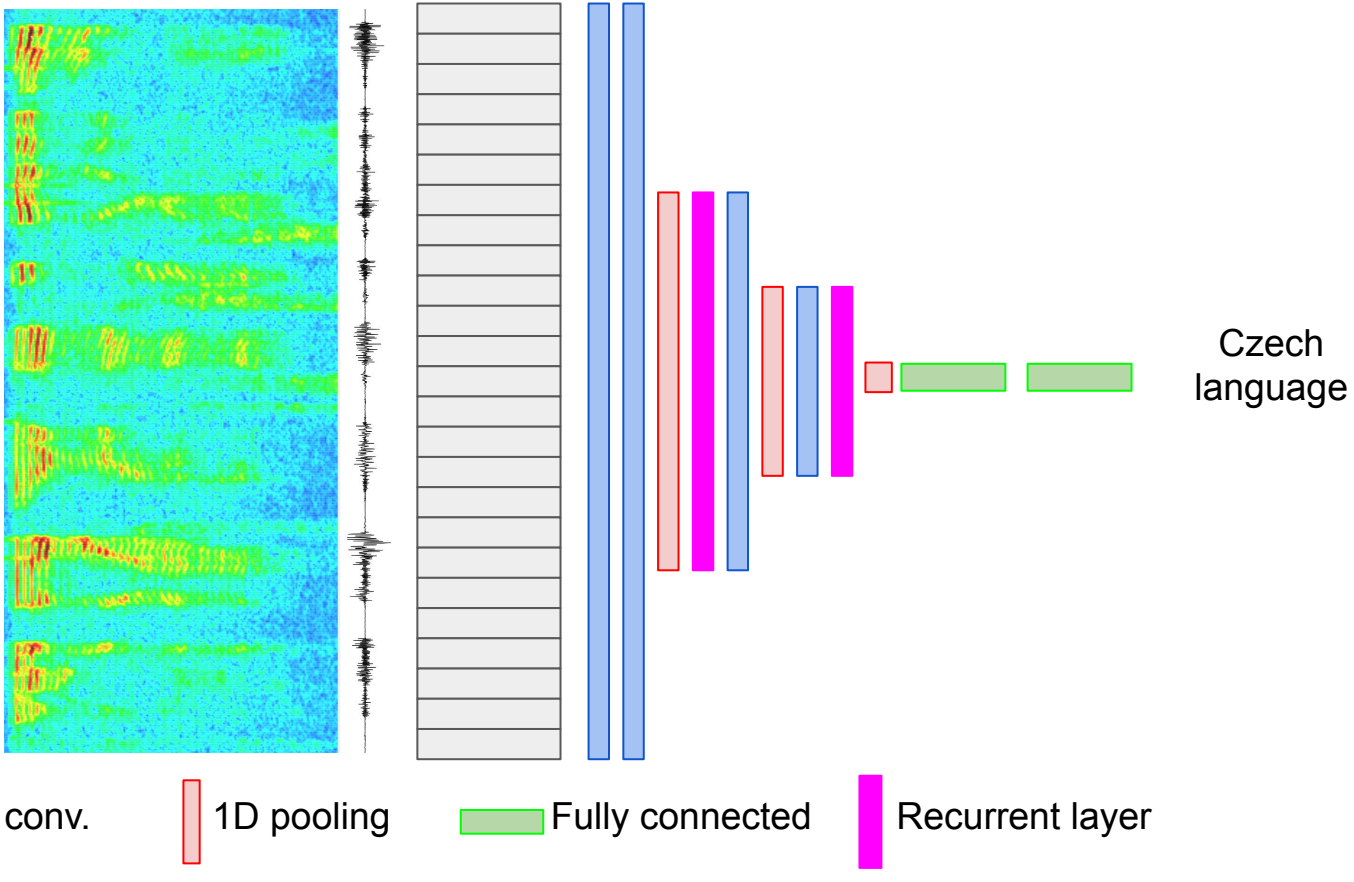
Classification



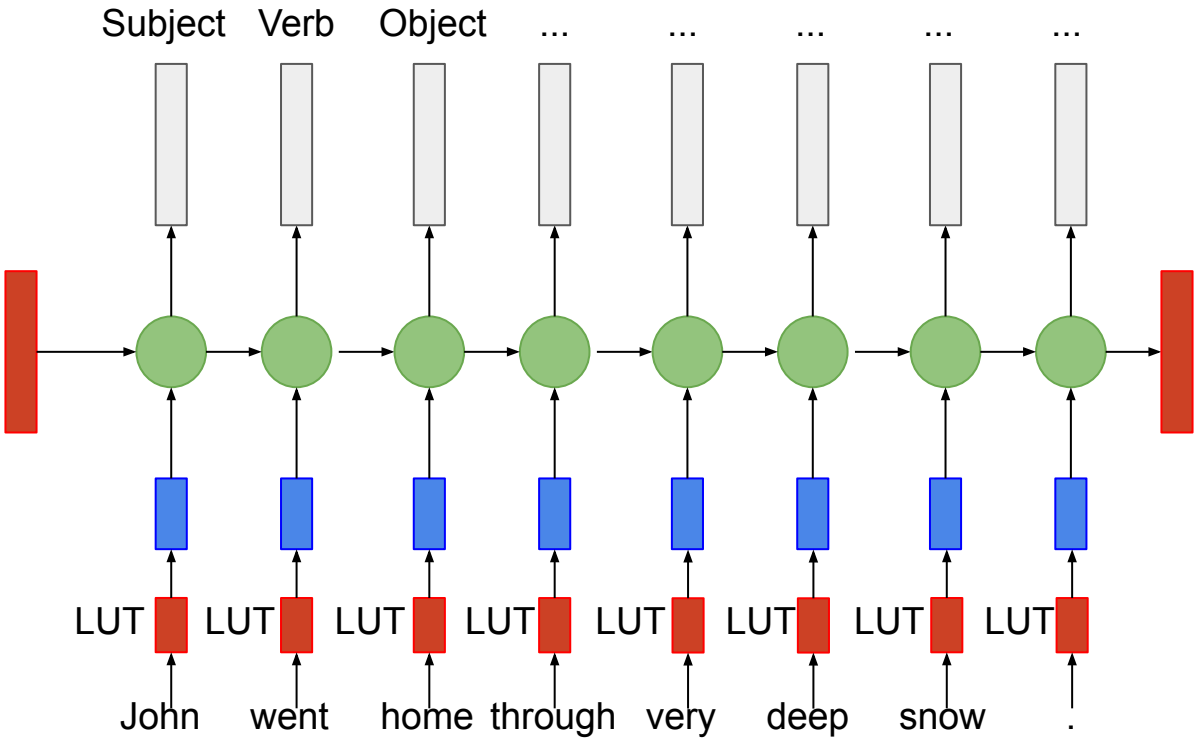
Classification



Classification



Word tagging



Reading Comprehension (https://openai.com/blog/better-language-models/)

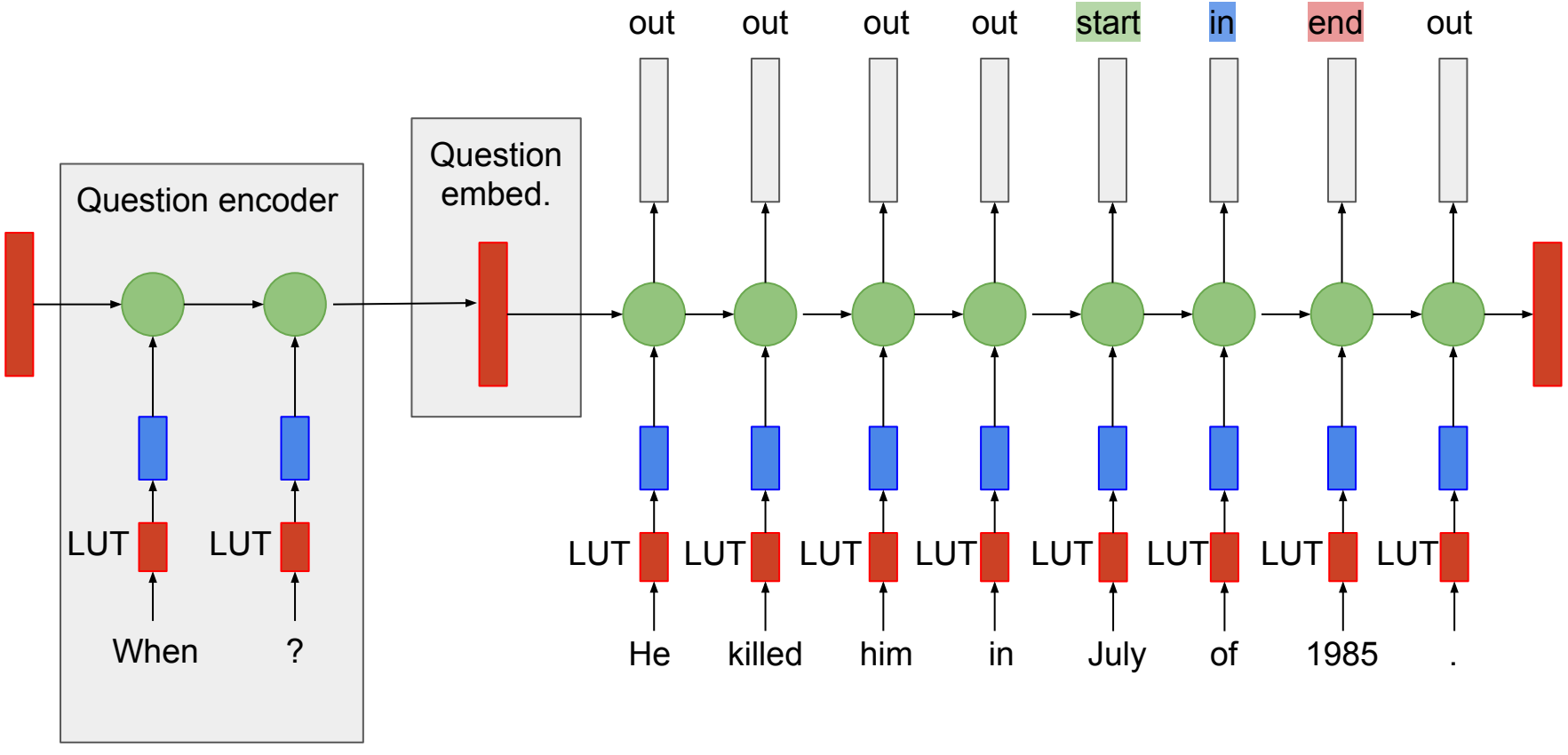
The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of "one world, one dream". Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the "Journey of Harmony", lasted 129 days and carried the torch 137,000 km (85,000 mi) – the longest distance of any Olympic torch relay since the tradition was started ahead of the 1936 Summer Olympics.

After being lit at the birthplace of the Olympic Games in Olympia, Greece on March 24, the torch traveled to the Panathinaiko Stadium in Athens, and then to Beijing, arriving on March 31. From Beijing, the torch was following a route passing through six continents. The torch has visited cities along the Silk Road, symbolizing ancient links between China and the rest of the world. The relay also included an ascent with the flame to the top of Mount Everest on the border of Nepal and Tibet, China from the Chinese side, which was closed specially for the event.

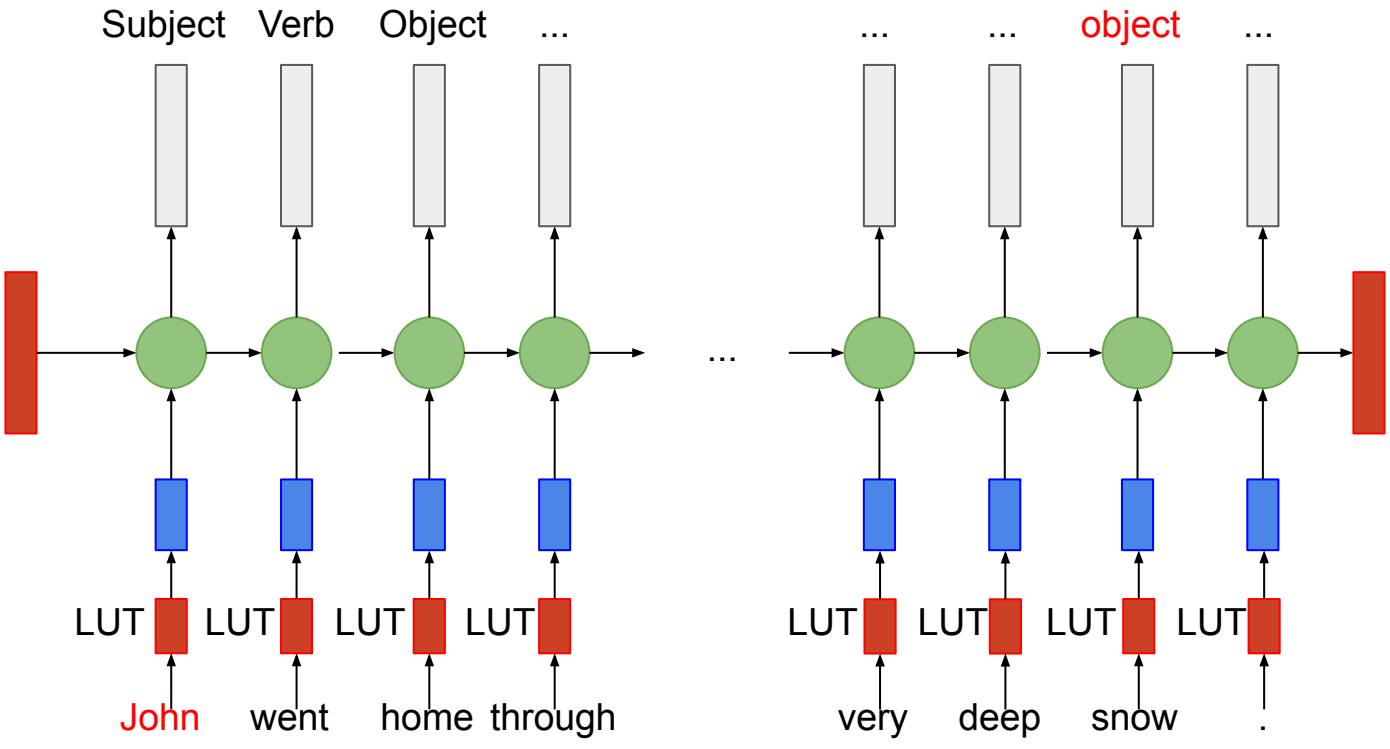
Q: What was the theme?
A: "one world, one dream".
Q: What was the length of the race?
A: 137,000 km
Q: Was it larger than previous ones?
A: No (wrong?)

Q: Where did the race begin?
A: Olympia, Greece
Q: Is there anything notable about that place?
A: birthplace of Olympic Games
Q: Where did they go after?
A: Athens

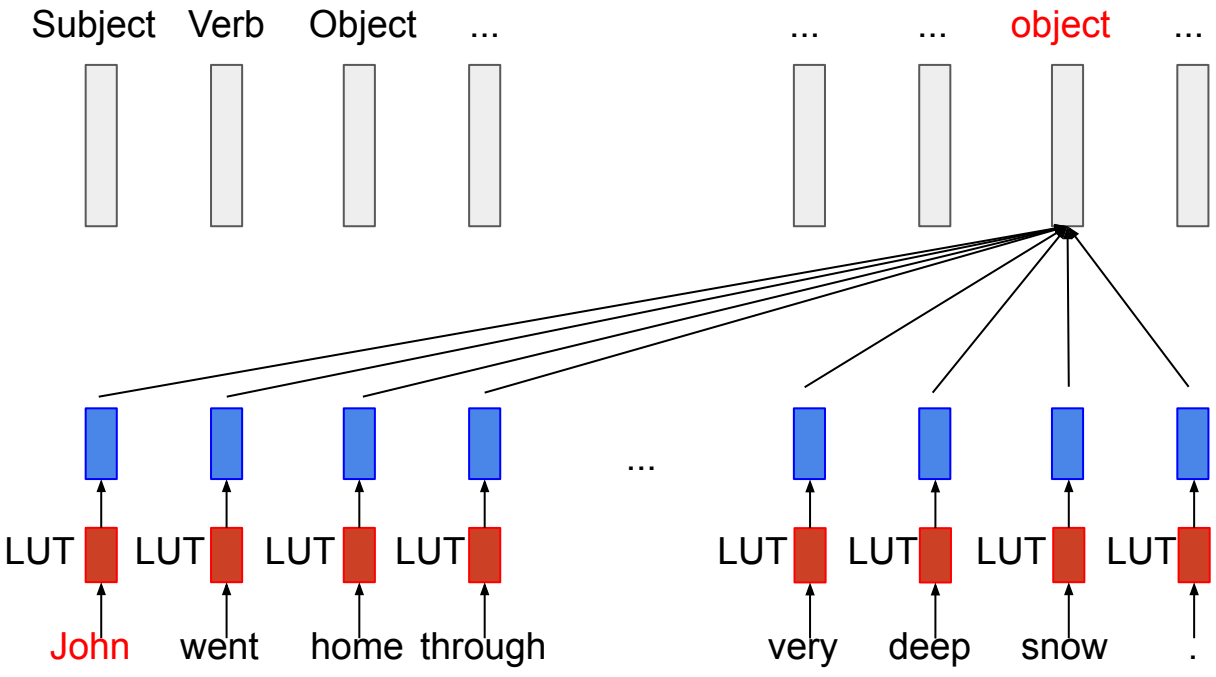
Reading Comprehension



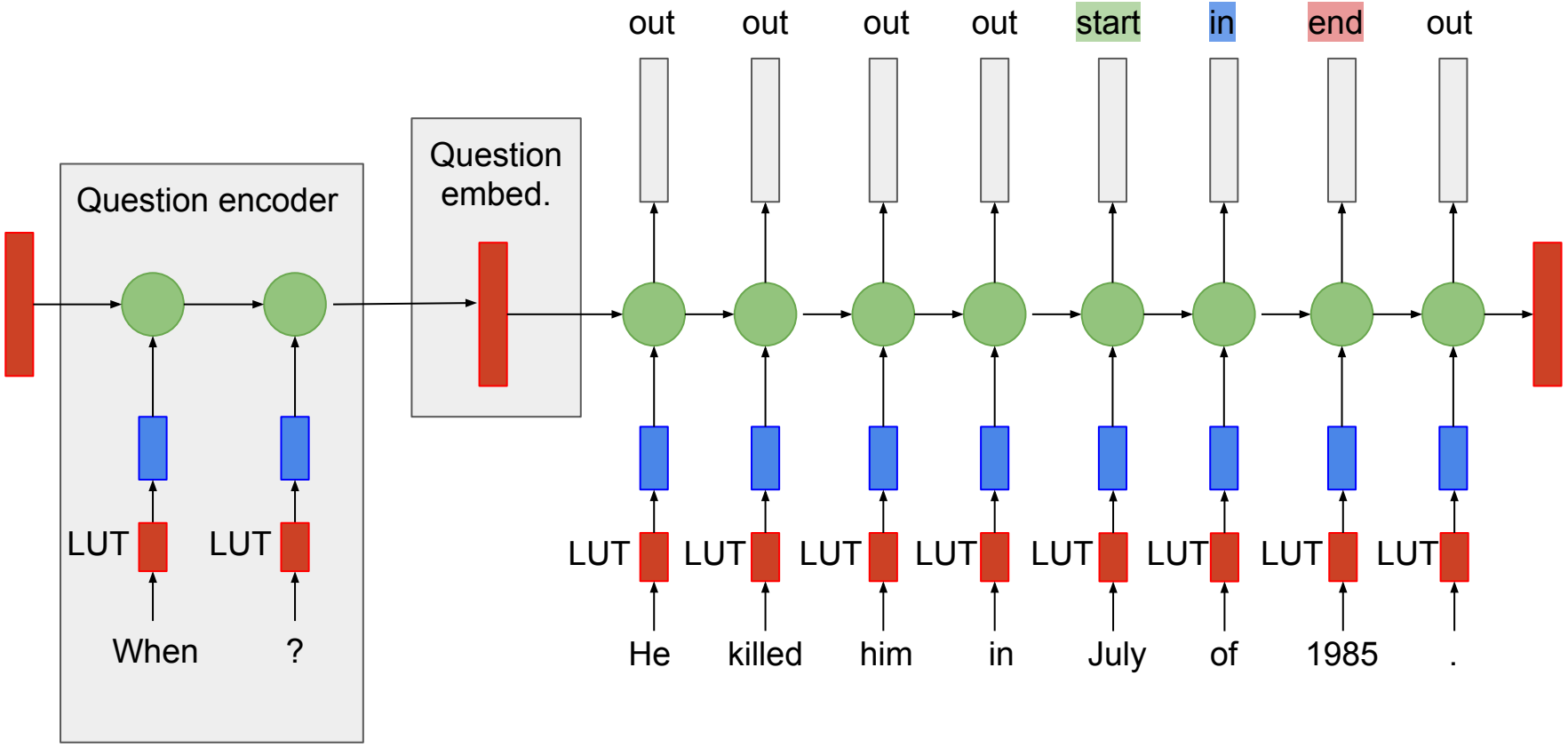
Learn long distance dependencies?



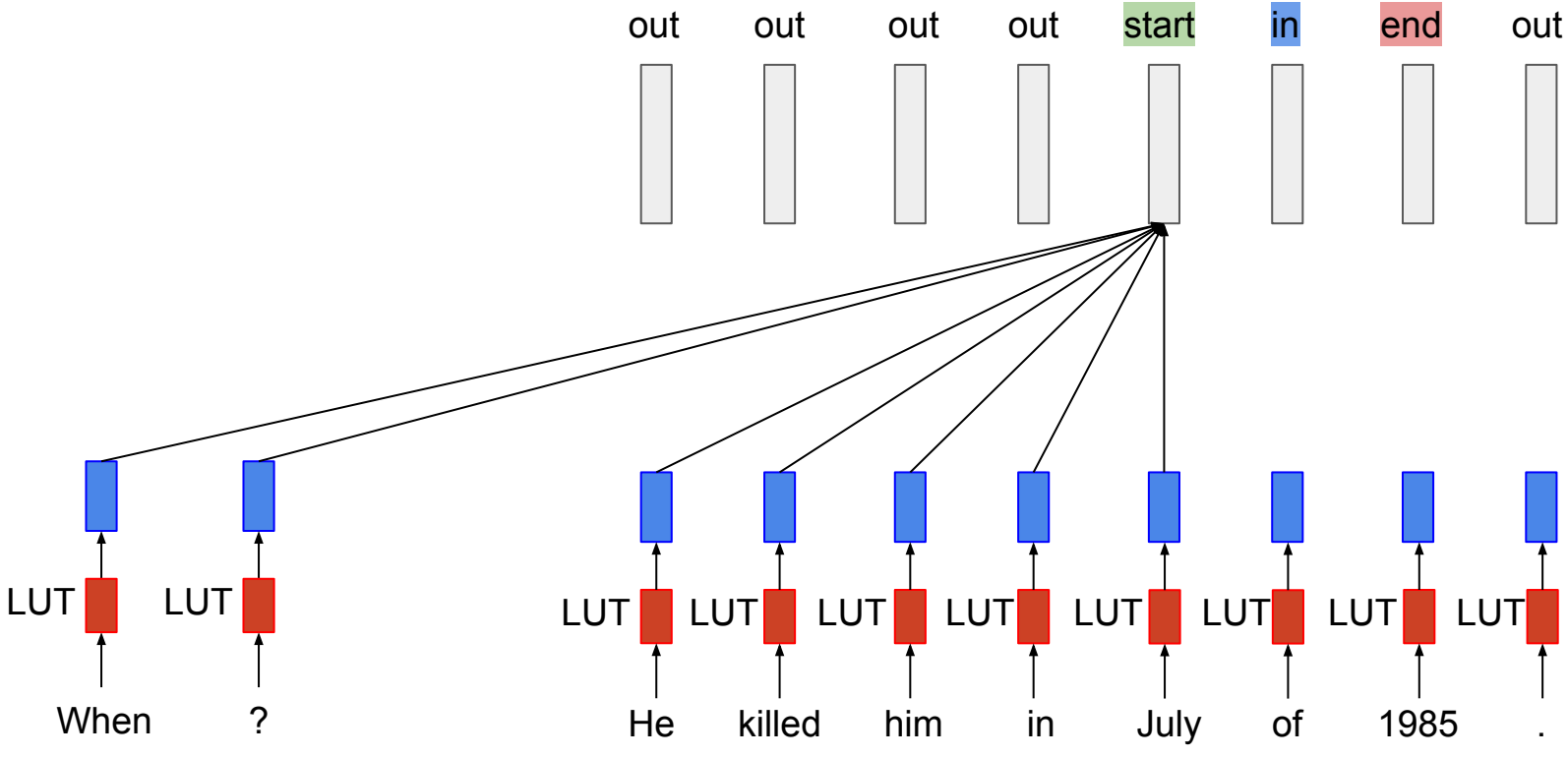
Attention?


Subject
Verb
Object
...
...
...
object
...
LUT
LUT
LUT
LUT
...
LUT
LUT
LUT
LUT
John
went
home
through
very
deep
snow
.

Attention - long distance dependencies

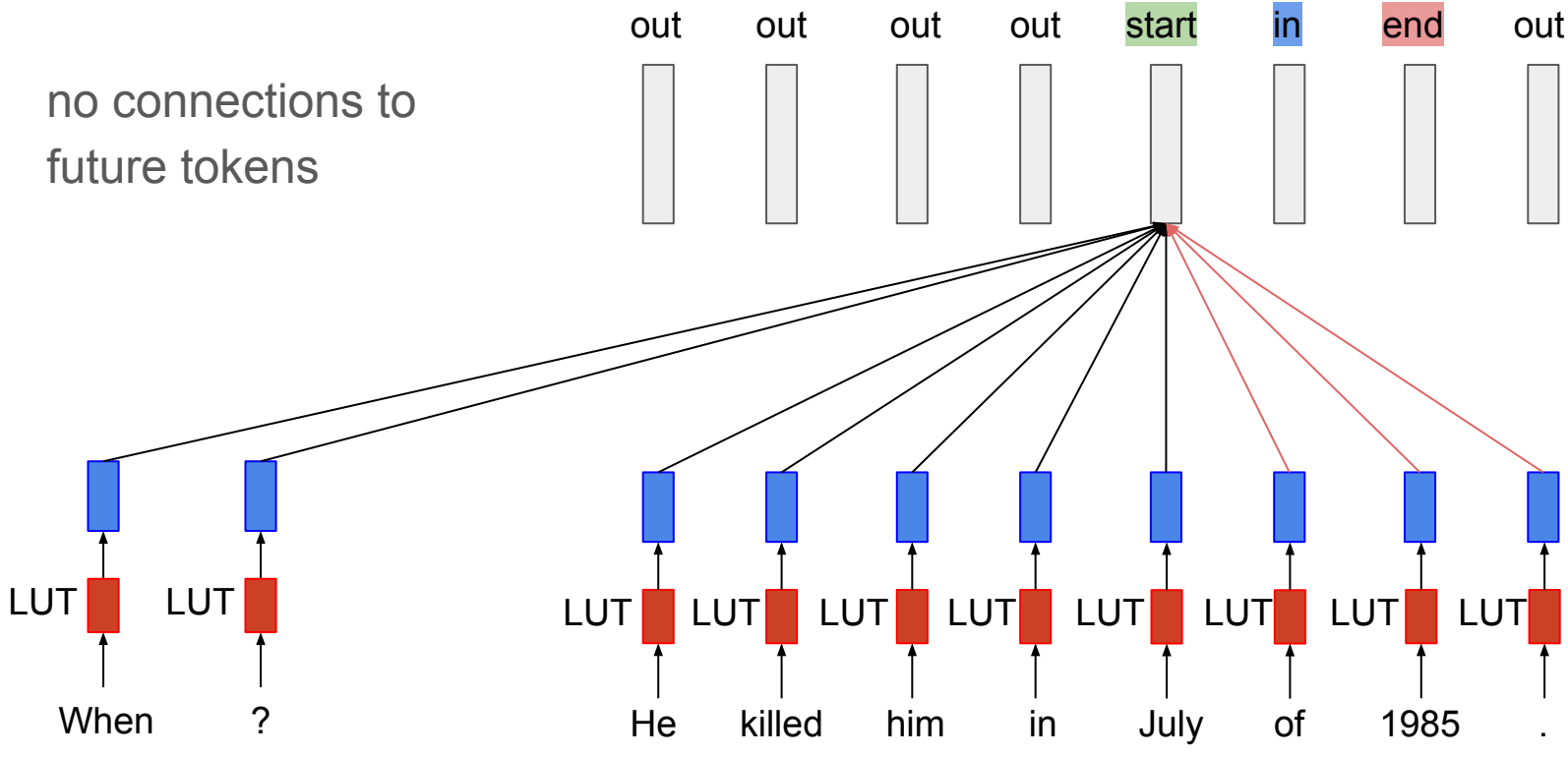


Attention - long distance dependencies



Attention - causal attention mask / auto-regresive / causal models

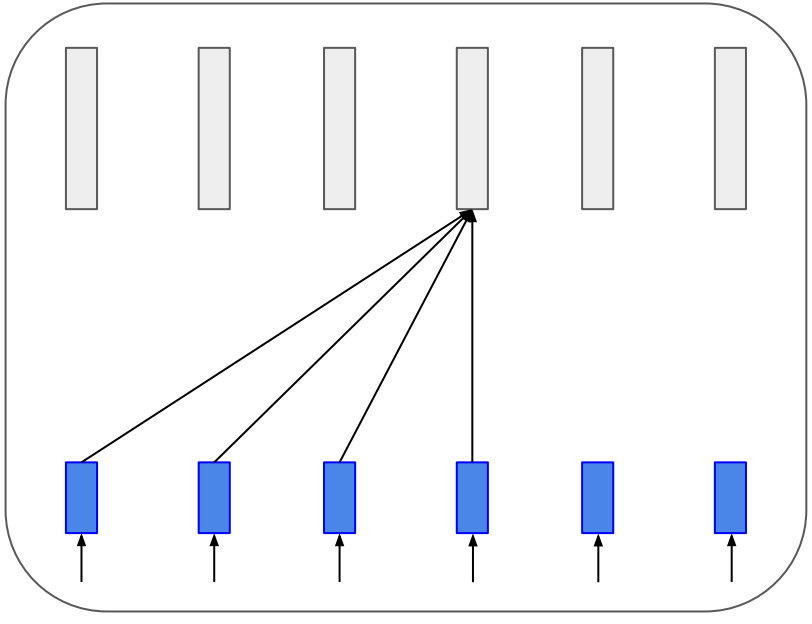
- no connections to future tokens



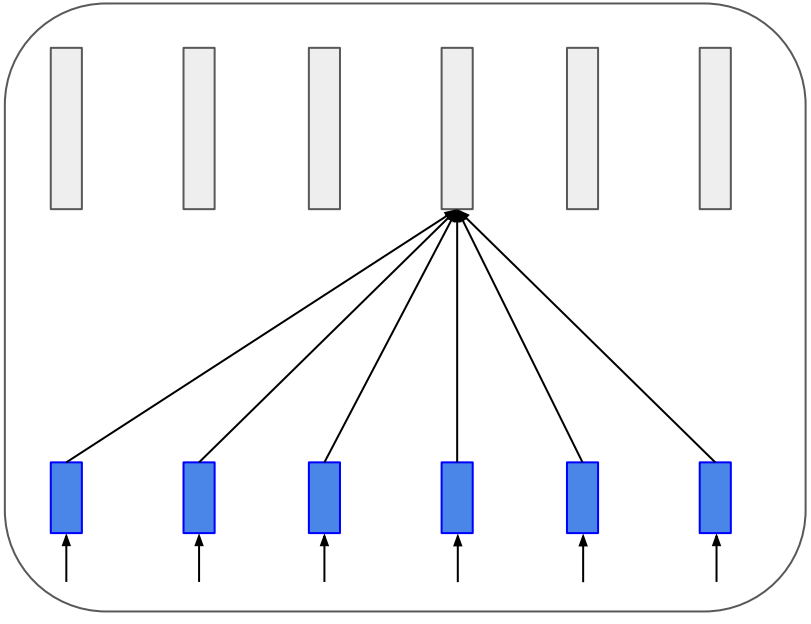
Causal vs. full attention

- set future attention weights to 0/-inf

Transformer Decoder
causal

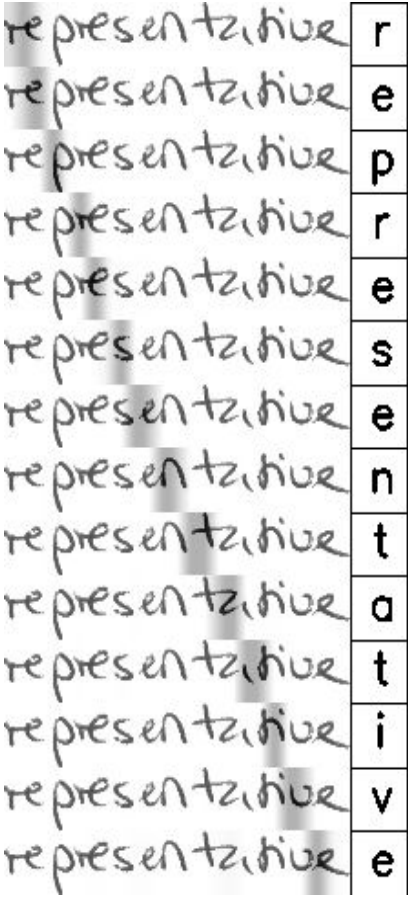


Transformer Encoder
full



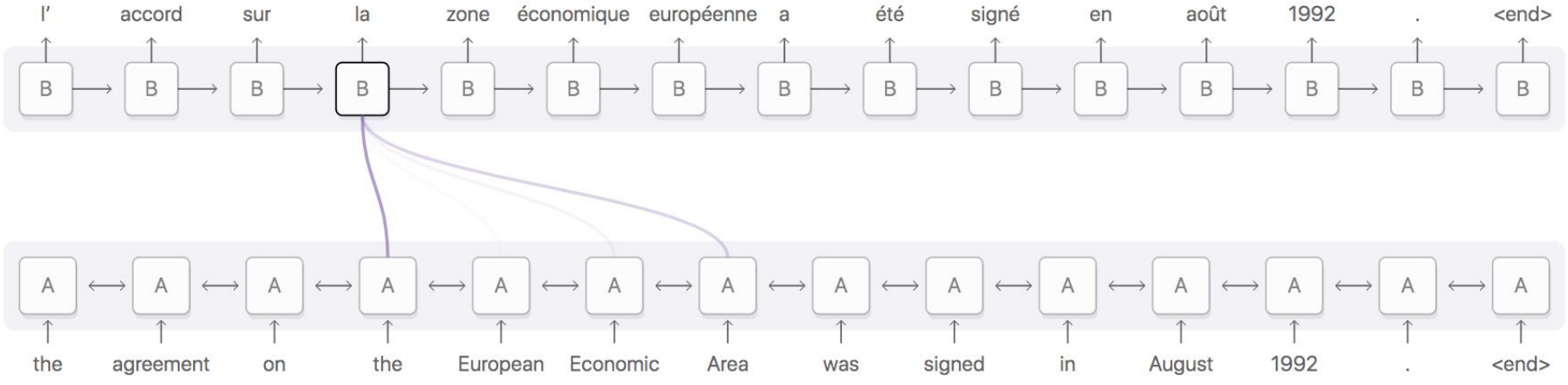
Attention - text transcription

- Generate characters sequentially
- Learn to access information useful for each character
-



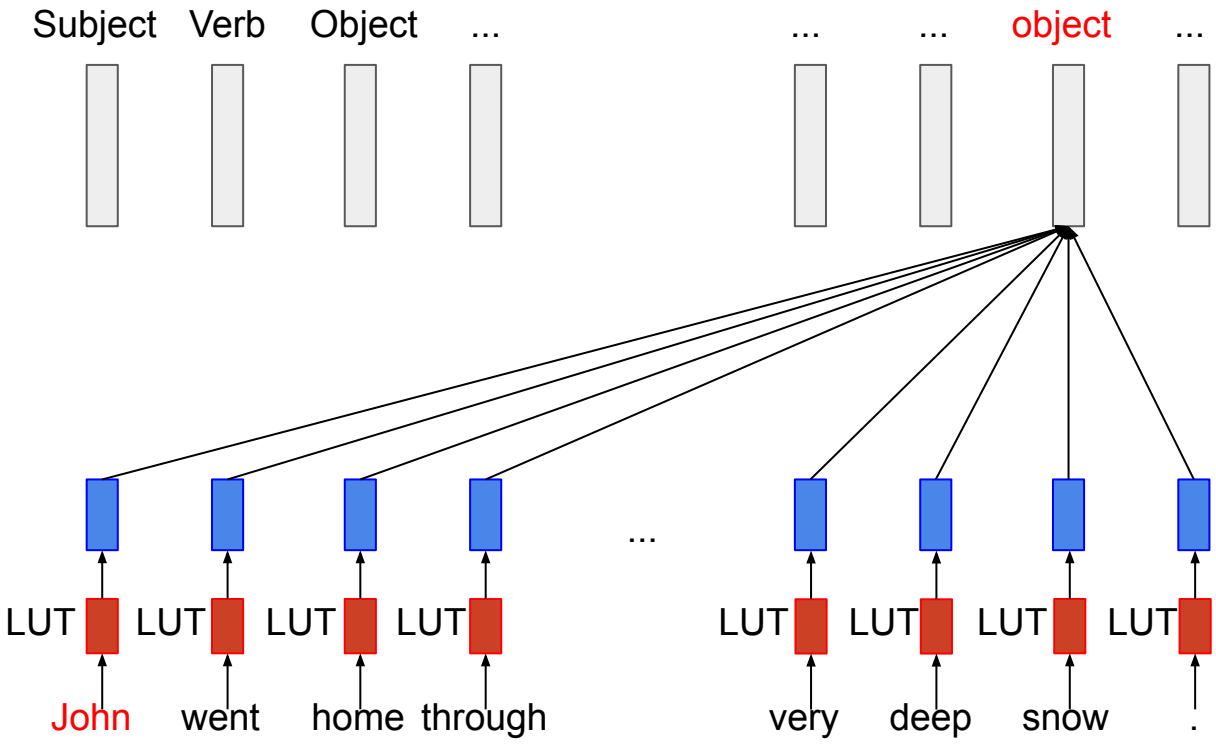
Kang et al.: Convolve, Attend and Spell: An Attention-based Sequence-to-Sequence Model for Handwritten Word Recognition. GCPR 2018.

Attention

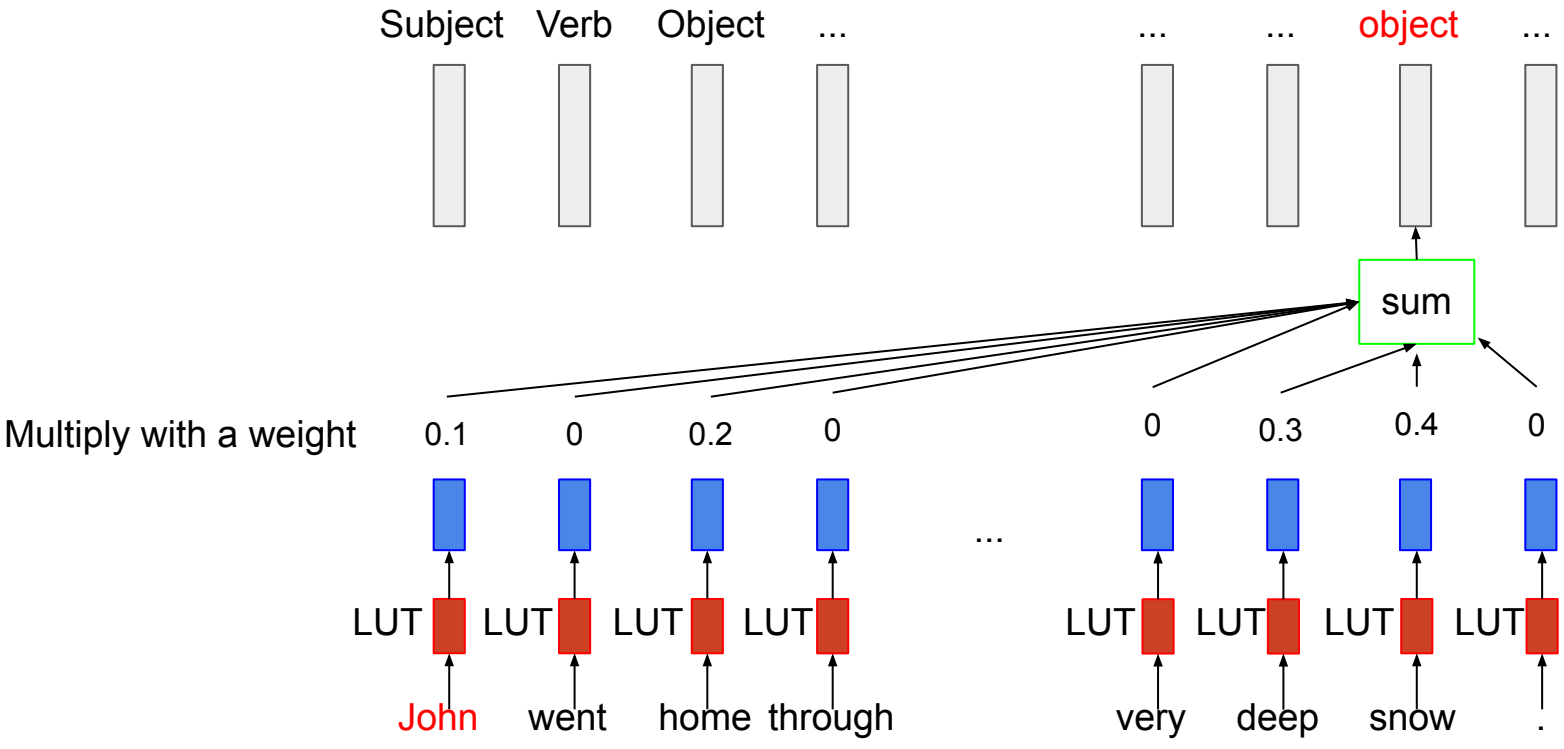


Olah, Carter: Attention and Augmented Recurrent Neural Networks, 2016.
https://distill.pub/2016/augmented-rnns/

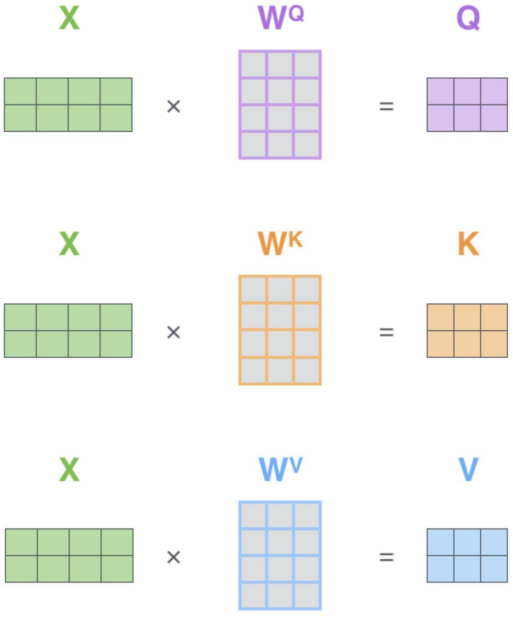
Attention - mechanism

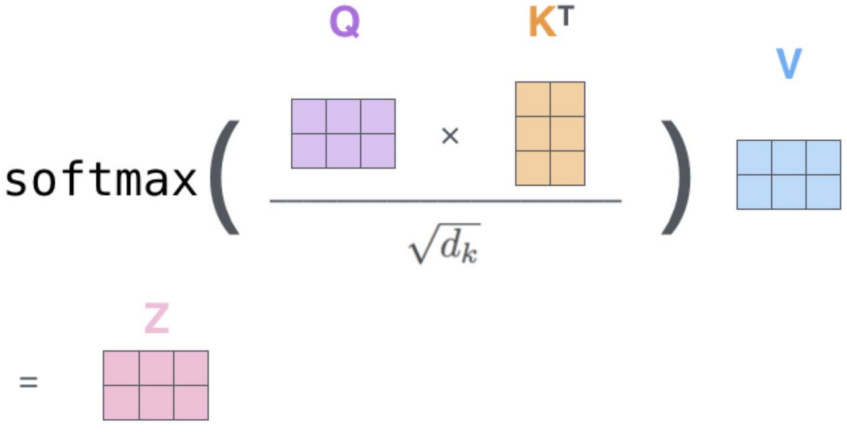


Attention - mechanism



Attention - computations

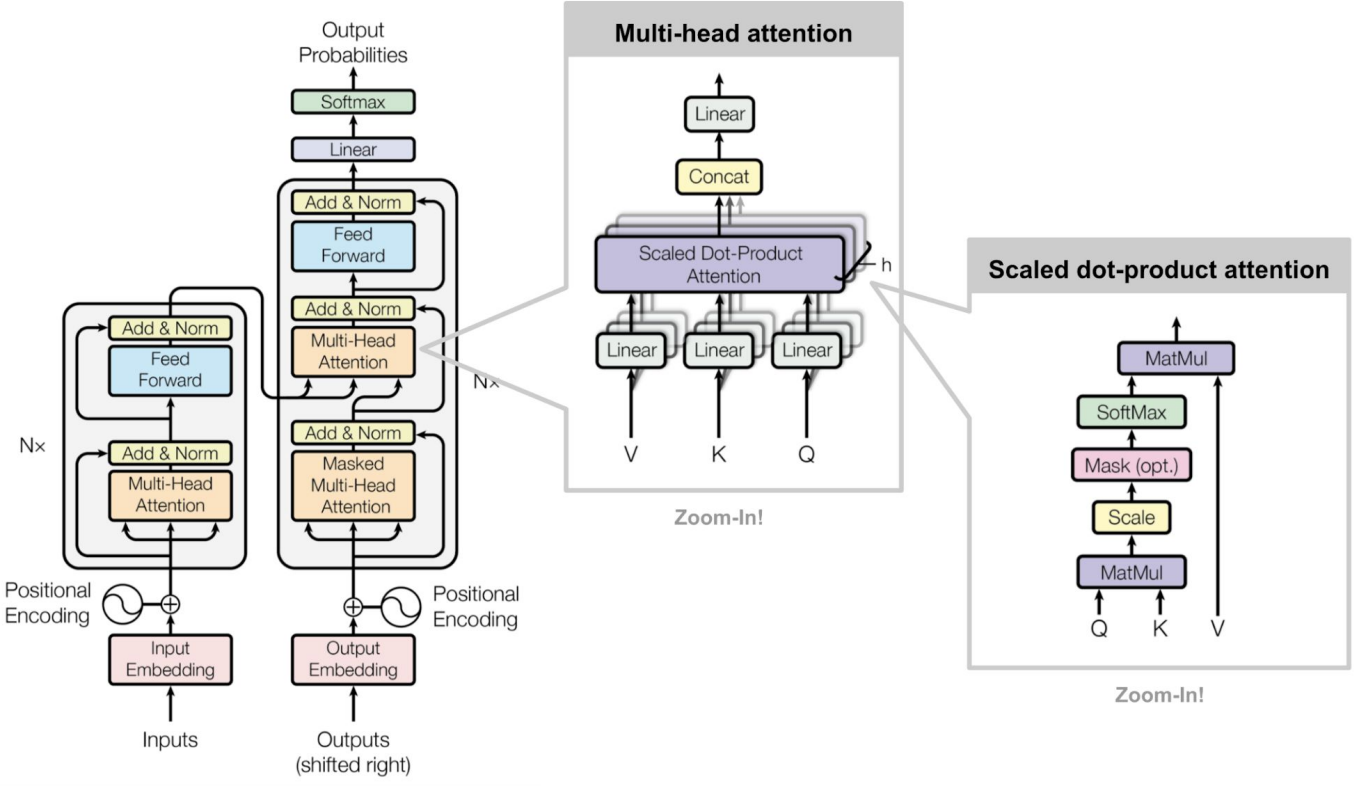




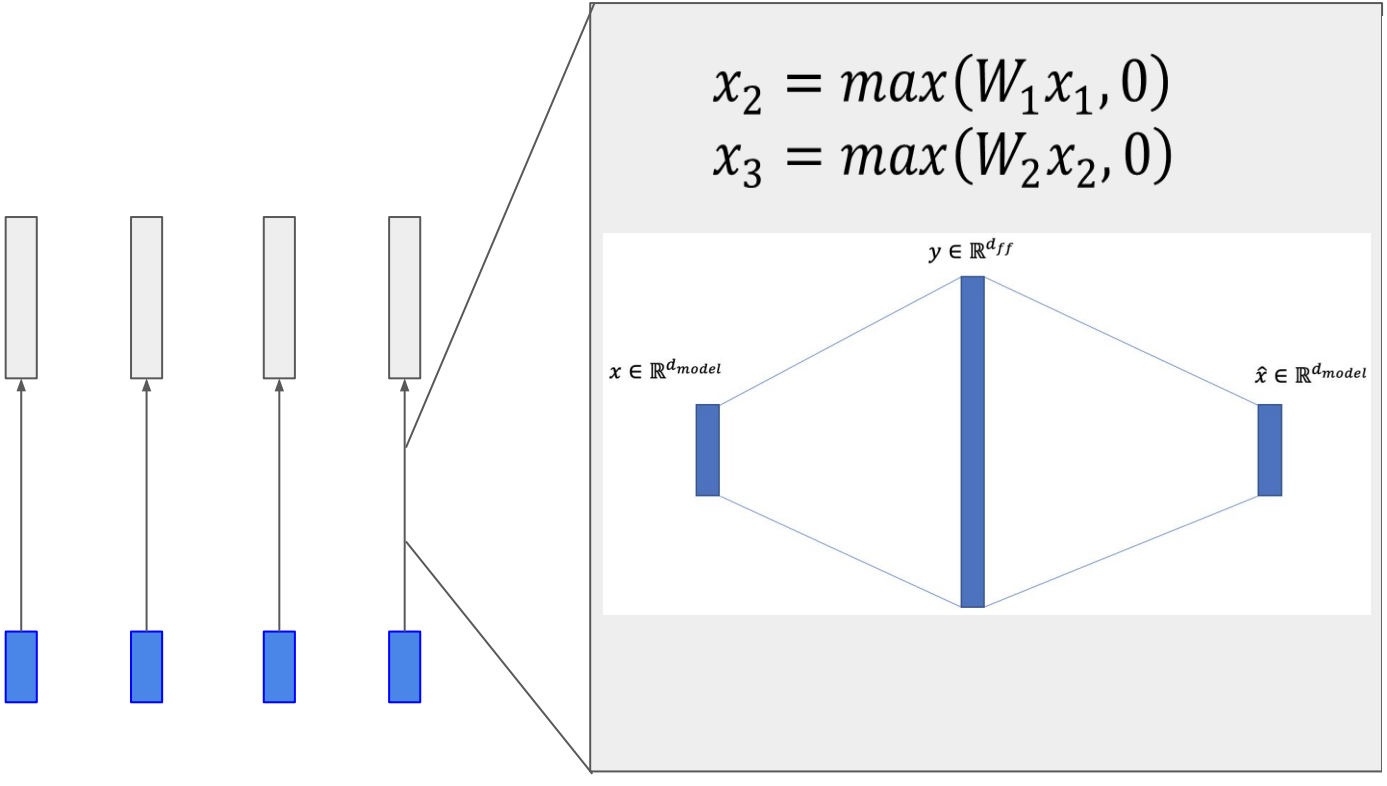
The self-attention calculation in matrix form

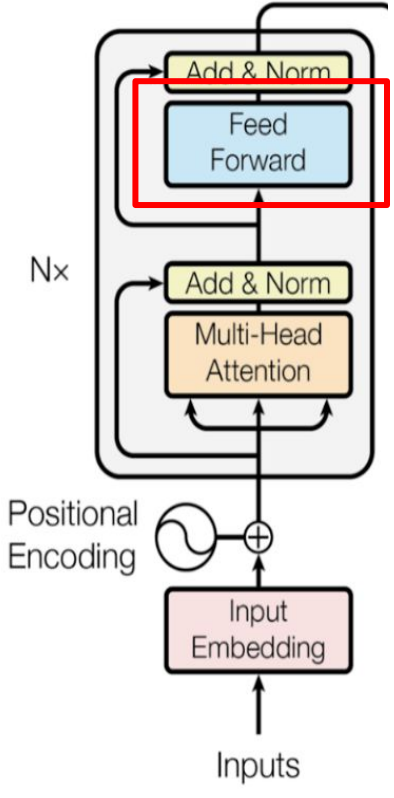
Attention Is All You Need (Vaswani et al, NIPS 2017)

Tool: [Tensor2tensor]

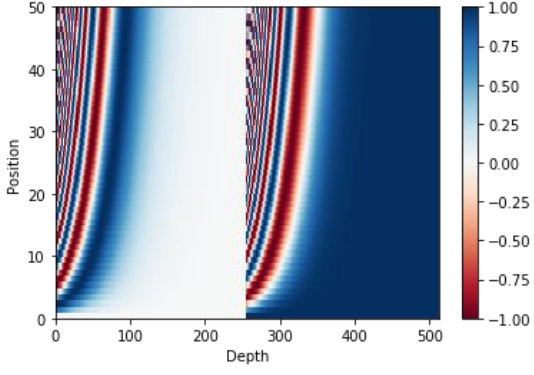


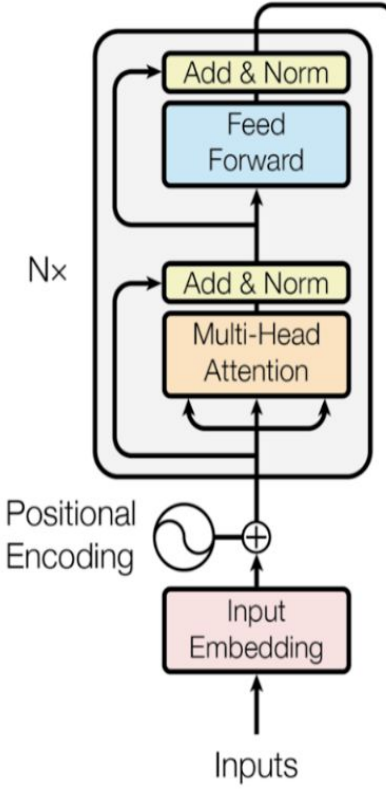
MLP - Feed Forward





Transformer - positional encoding

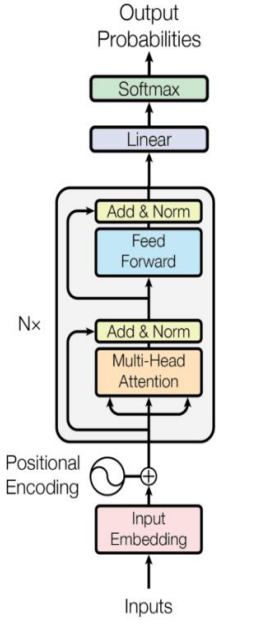




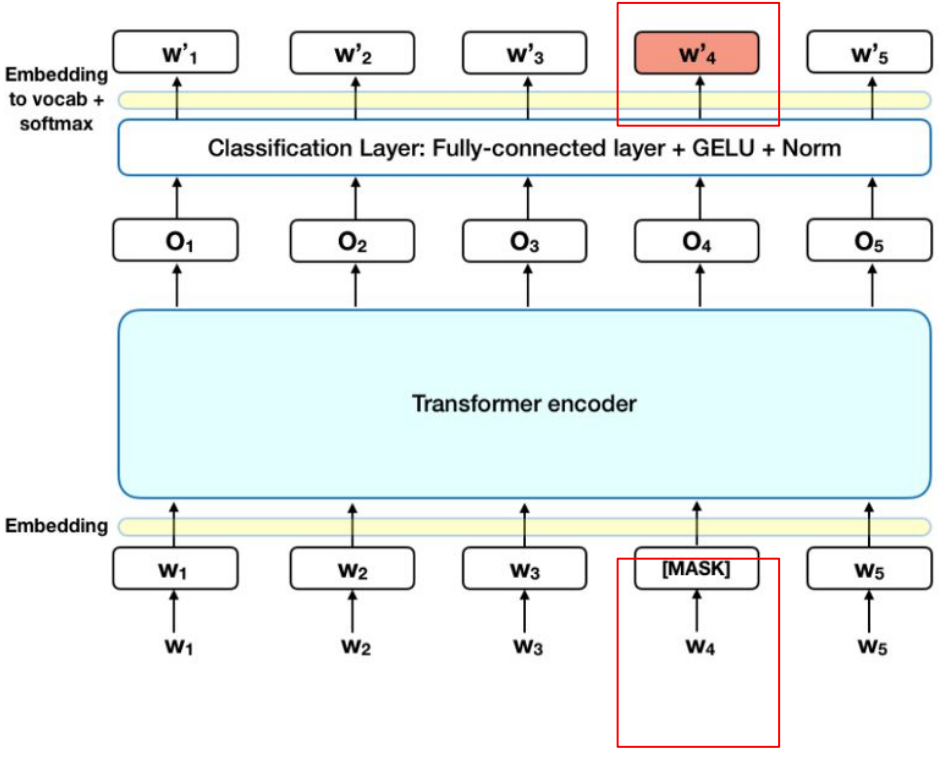
BERT - Bidirectional Encoder Representations from Transformers

Trained to fill in masked words in a sentence.

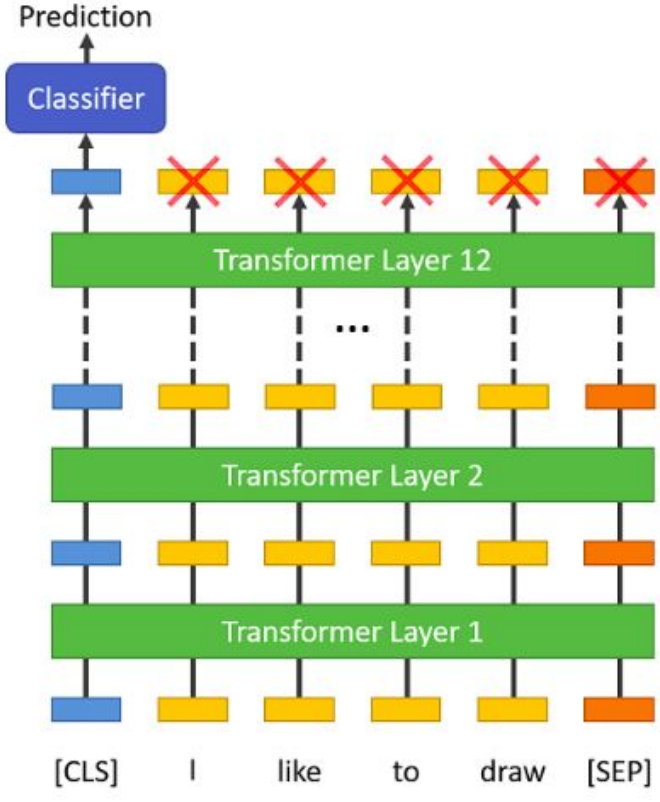
Fine-tuned for specific “downstream” tasks in text processing / NLP.



Encoder-only



BERT - fine-tune for downstream tasks



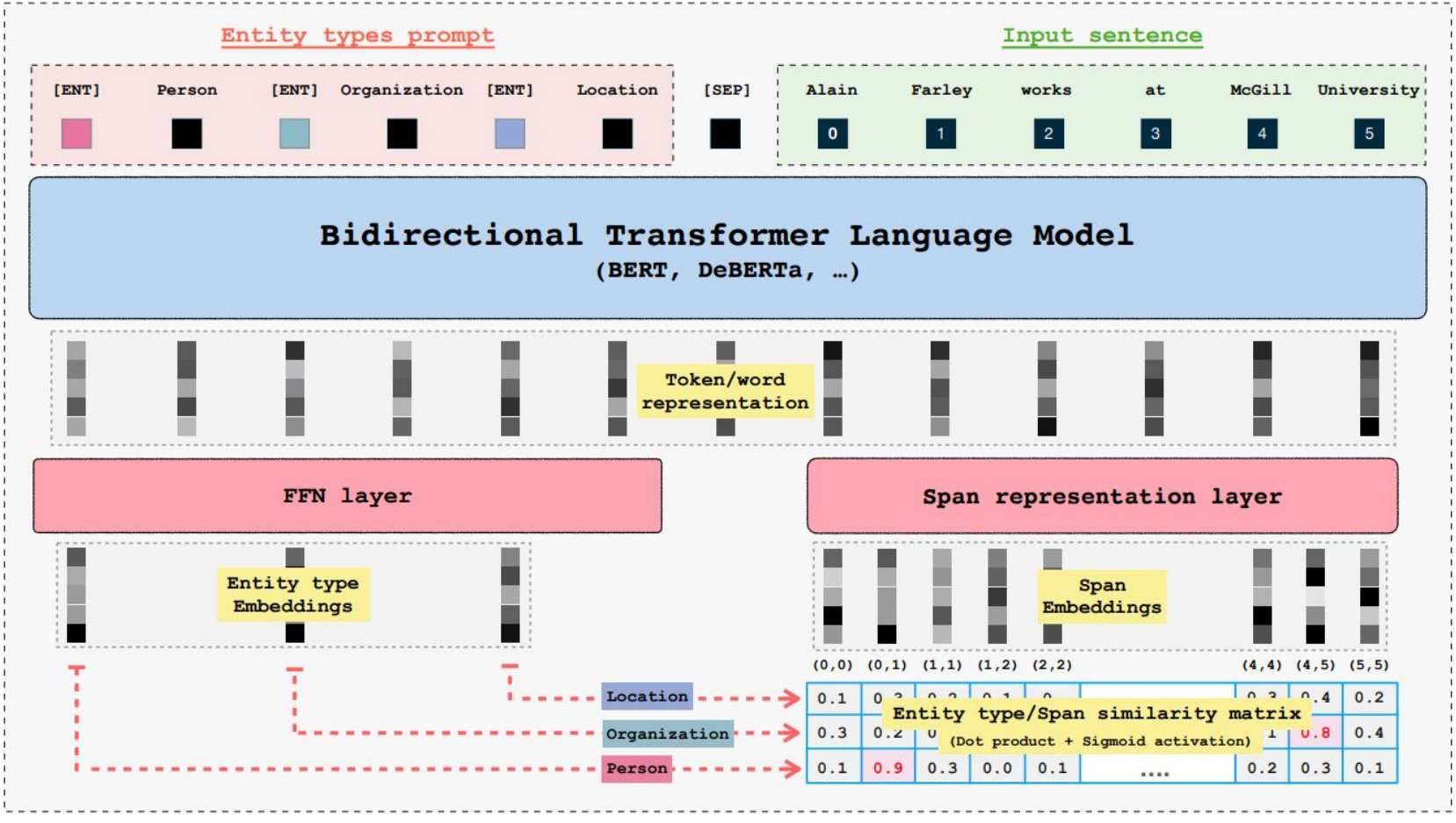
Reading Comprehension (https://openai.com/blog/better-language-models/)

The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of "one world, one dream". Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the "Journey of Harmony", lasted 129 days and carried the torch 137,000 km (85,000 mi) – the longest distance of any Olympic torch relay since the tradition was started ahead of the 1936 Summer Olympics.

After being lit at the birthplace of the Olympic Games in Olympia, Greece on March 24, the torch traveled to the Panathinaiko Stadium in Athens, and then to Beijing, arriving on March 31. From Beijing, the torch was following a route passing through six continents. The torch has visited cities along the Silk Road, symbolizing ancient links between China and the rest of the world. The relay also included an ascent with the flame to the top of Mount Everest on the border of Nepal and Tibet, China from the Chinese side, which was closed specially for the event.

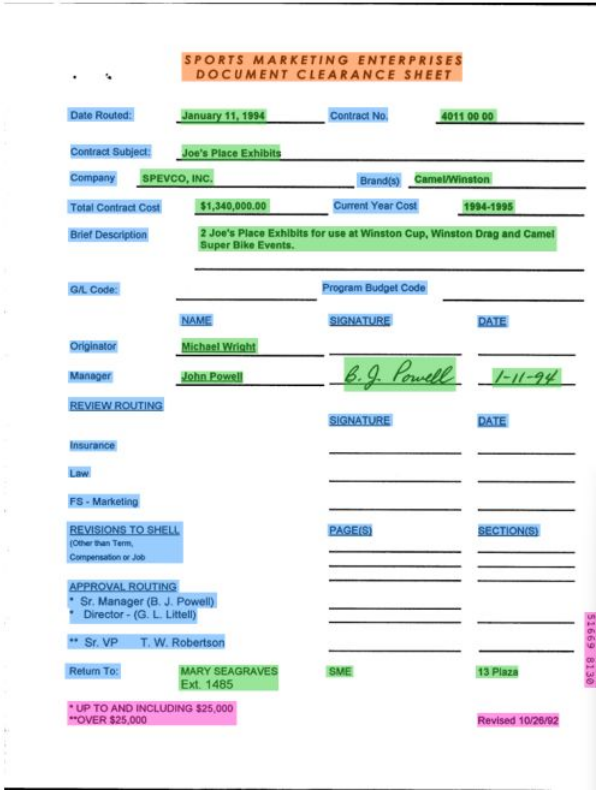
Q: What was the theme?
A: "one world, one dream".
Q: What was the length of the race?
A: 137,000 km
Q: Was it larger than previous ones?
A: No (wrong?)

Q: Where did the race begin?
A: Olympia, Greece
Q: Is there anything notable about that place?
A: birthplace of Olympic Games
Q: Where did they go after?
A: Athens



Zaratiana et al.:GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer, 2023.

Multi-modal


SPORTS MARKETING ENTERPRISES
DOCUMENT CLEARANCE SHEET

Date Routed: January 11, 1994 Contract No. 4011 00 00

Contract Subject: Joe's Place Exhibits

Company SPEVCO, INC. Brand(s) Camel/Winston

Total Contract Cost $1,340,000.00 Current Year Cost 1994-1995

Brief Description 2 Joe's Place Exhibits for use at Winston Cup, Winston Drag and Camel Super Bike Events.

G/L Code: Program Budget Code

NAME SIGNATURE DATE

Originator Michael Wright

Manager John Powell B. J. Powell 1-11-94

REVIEW ROUTING SIGNATURE DATE

Insurance

Law

FS - Marketing

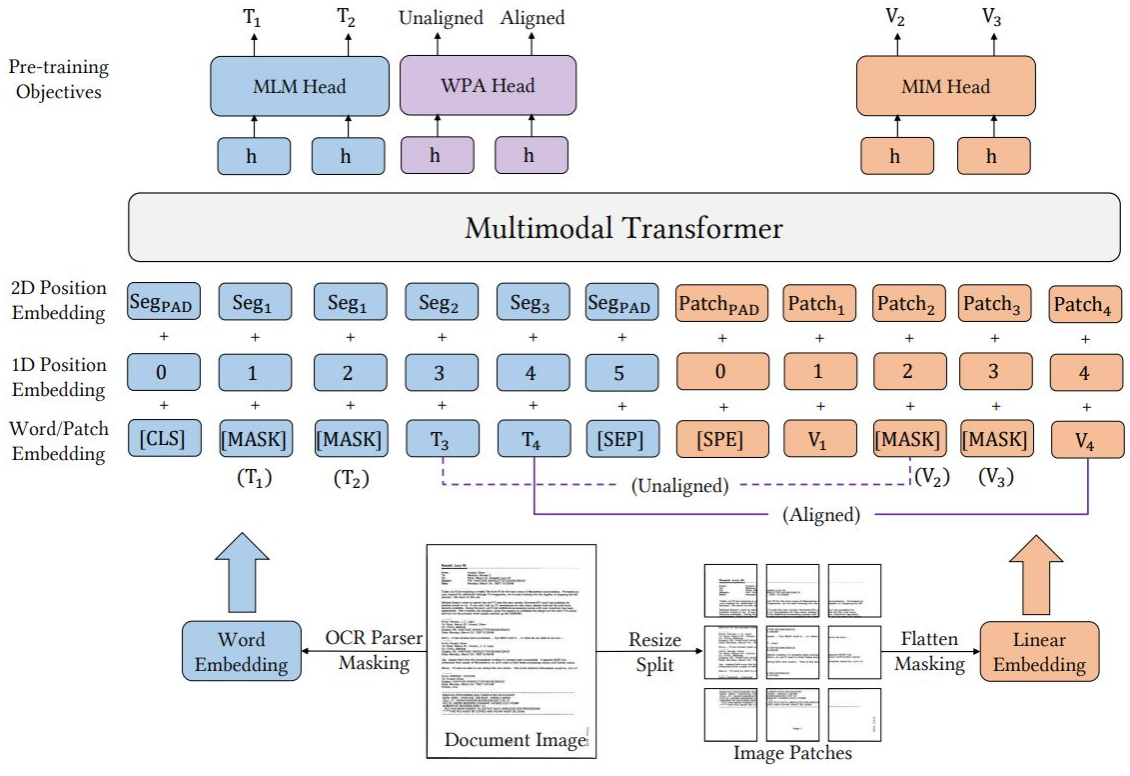
REVISIONS TO SHELL (Other than Term, Compensation or Job) PAGE(S) SECTION(S)

APPROVAL ROUTING
* Sr. Manager (B. J. Powell)
* Director - (G. L. Littell)

** Sr. VP T. W. Robertson

Return To: MARY SEAGRAVES Ext. 1485 SME 13 Plaza

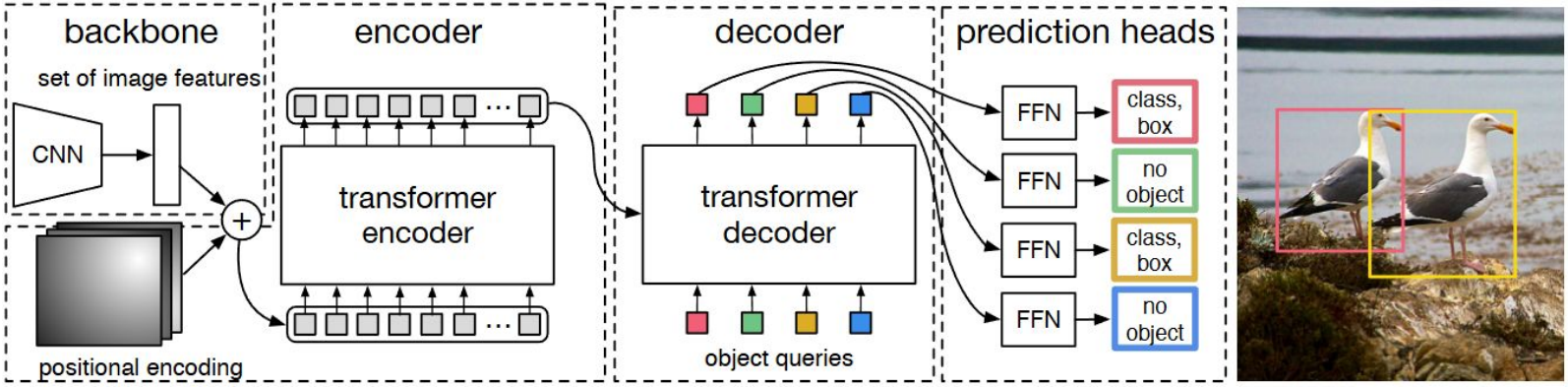
* UP TO AND INCLUDING $25,000
**OVER $25,000 Revised 10/26/92



Huang et al.: LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking

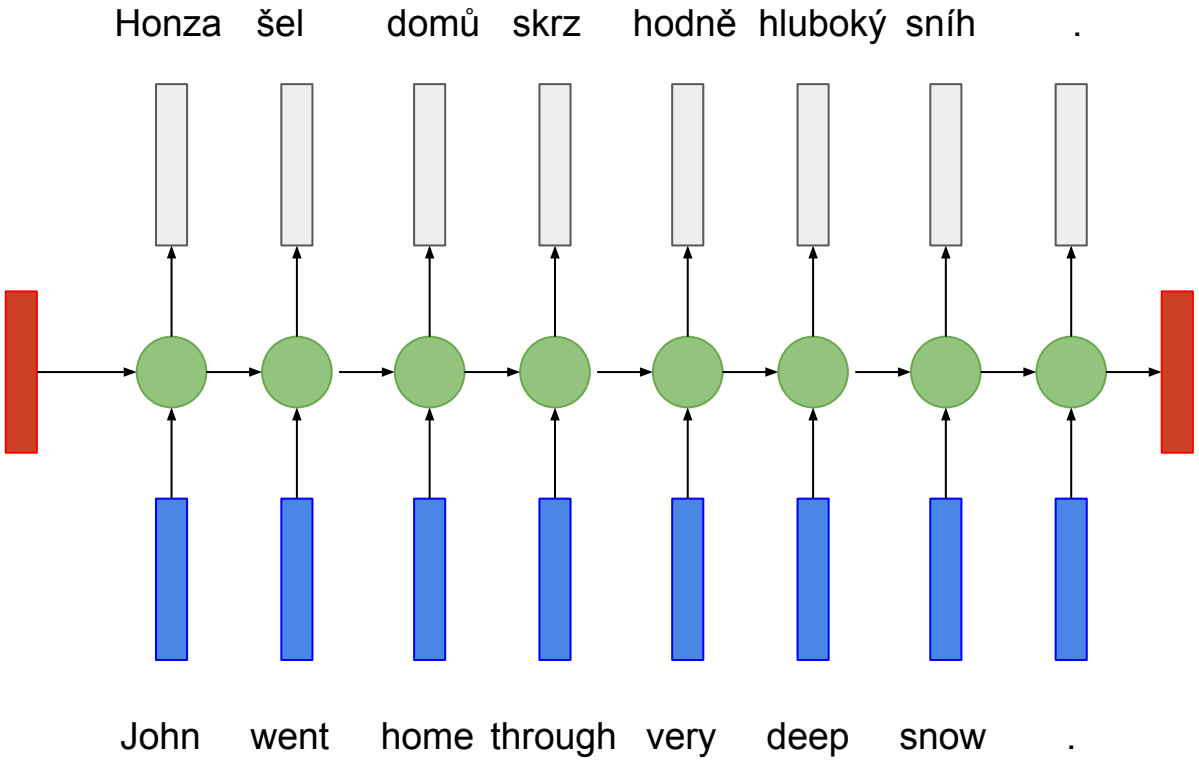
Detection transformer (DETR)

- Directly predict all the bounding boxes w/o NMS
- Encoder-decoder
- Transformer based decoder



Carion, Nicolas, et al. "End-to-end object detection with transformers." *European conference on computer vision.* Cham: Springer International Publishing, 2020.

Sequence to sequence models? Translation?



Honza šel domů hodně hlubokým sněhem.

Auto-regresive factorization (generative seq. models)

P(sentence | ~~John went home through very deep snow~~)

$$P(w_1, w_2, w_3, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots P(w_n|w_{n-1}, w_{n-2}, \ldots, w_1)$$

$$P(w_1, w_2, w_3, \ldots, w_n) = \prod P(w_i|w_{i-1}, w_{i-2}, \ldots, w_1)$$

Now we need a model which predicts probability of a single word given its history (prefix).

Language models

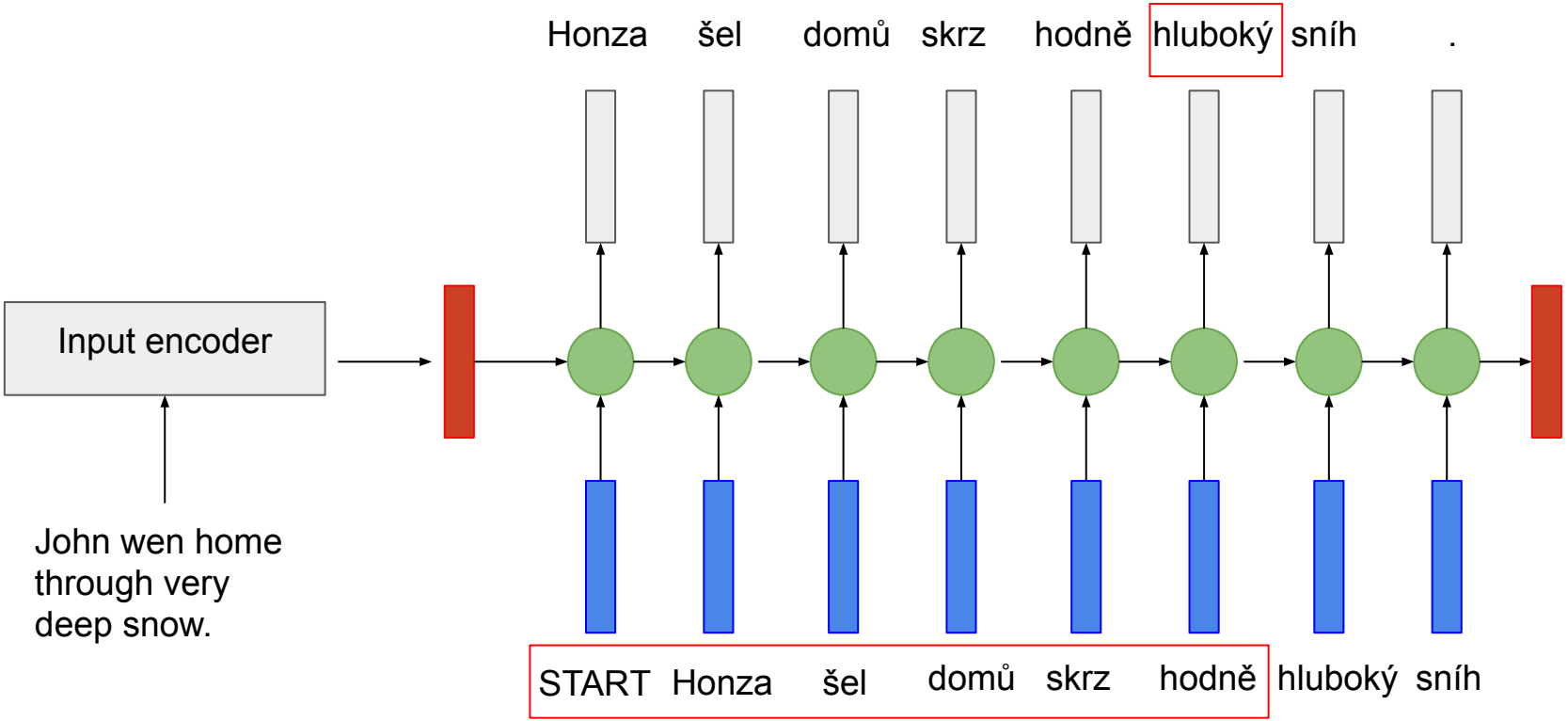
What is the next word in a sentence?

People like to visit ???????

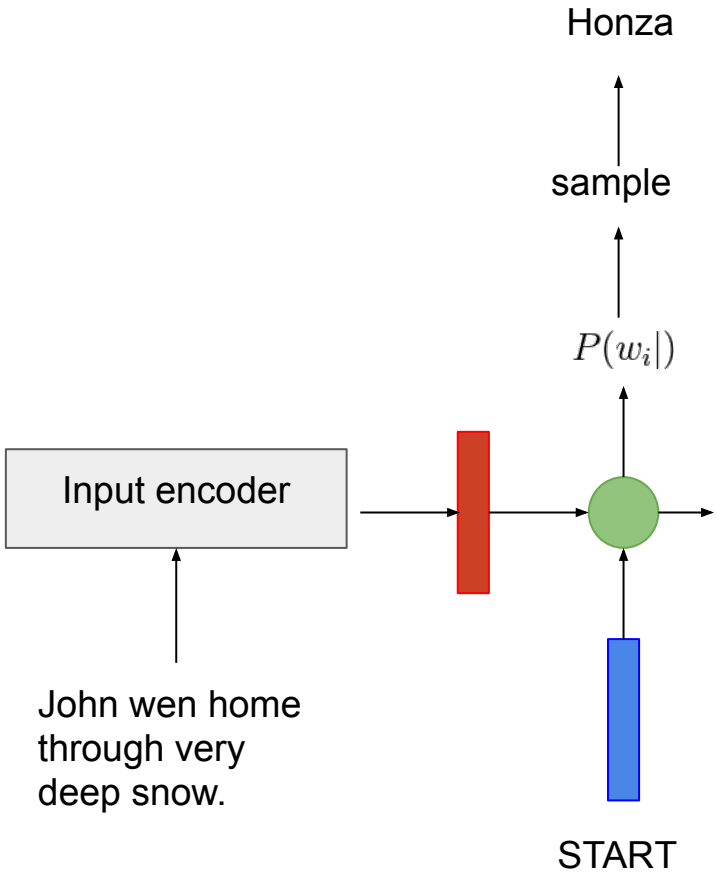
a library
a cinema
pubs
restaurants
a sauna
Prague
….
their grandma

Seq2seq - sentence probability

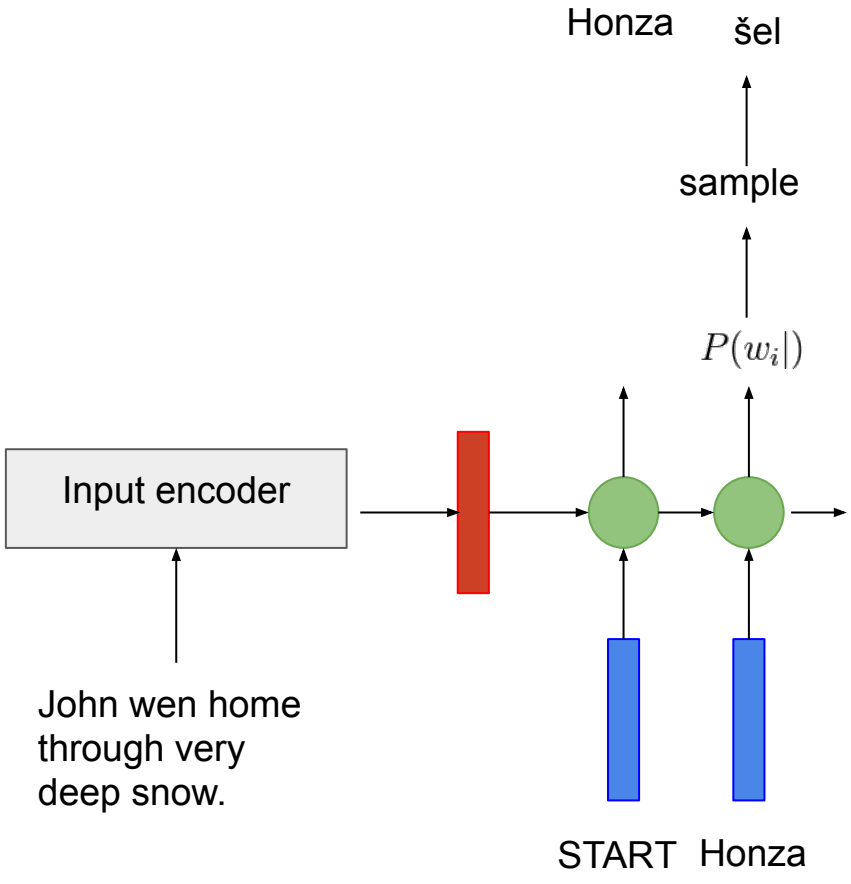
$$P(w_1, w_2, w_3, \ldots, w_n) = \prod P(w_i | w_{i-1}, w_{i-2}, \ldots, w_1)$$



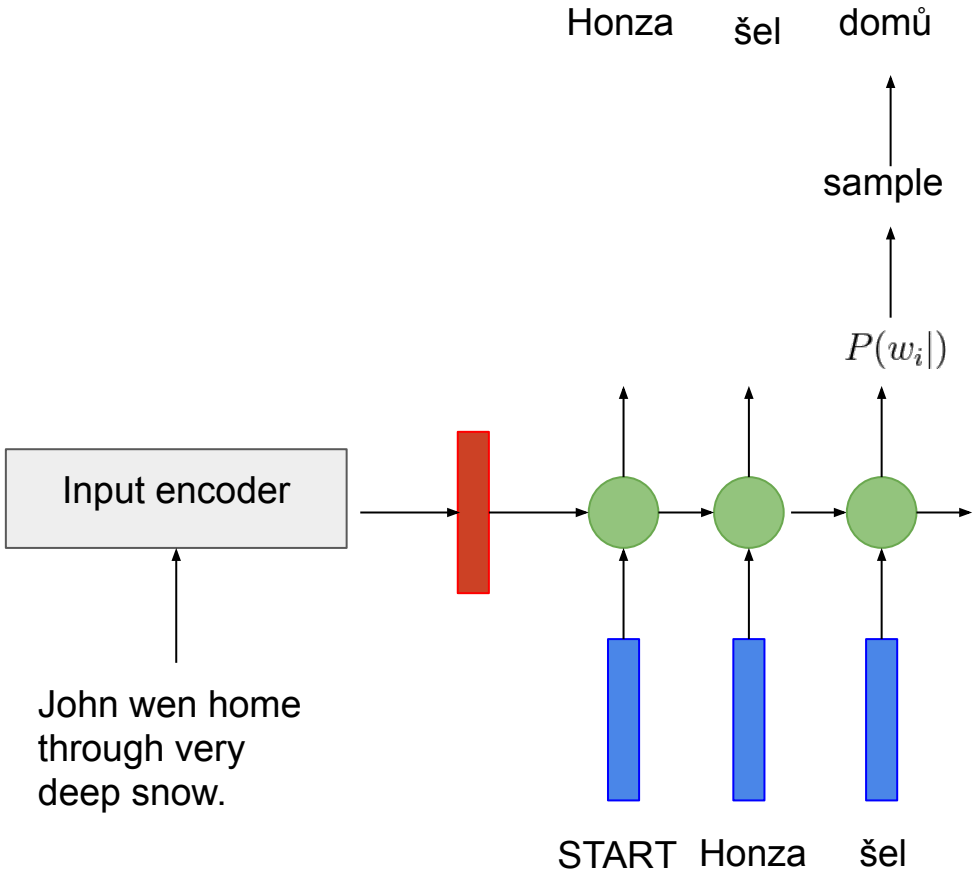
Seq2seq - sentence output sampling



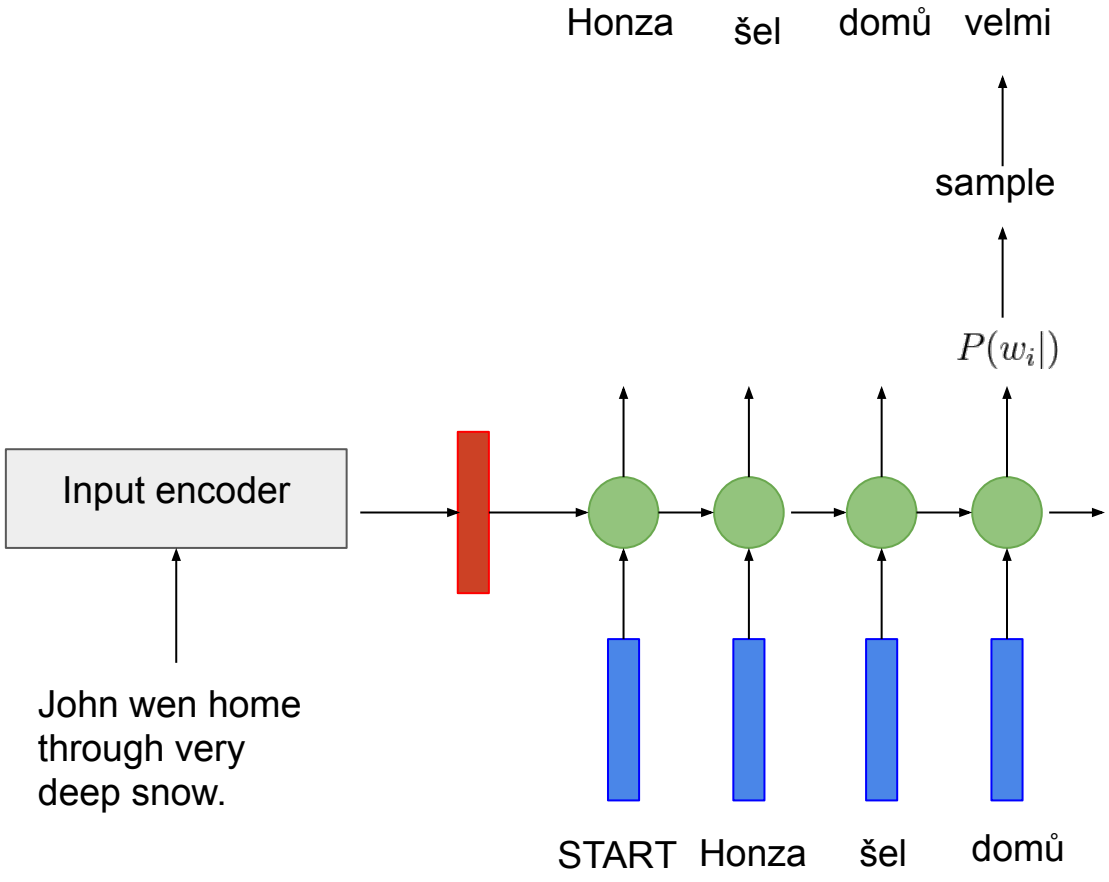
Seq2seq - sentence output sampling



Seq2seq - sentence output sampling

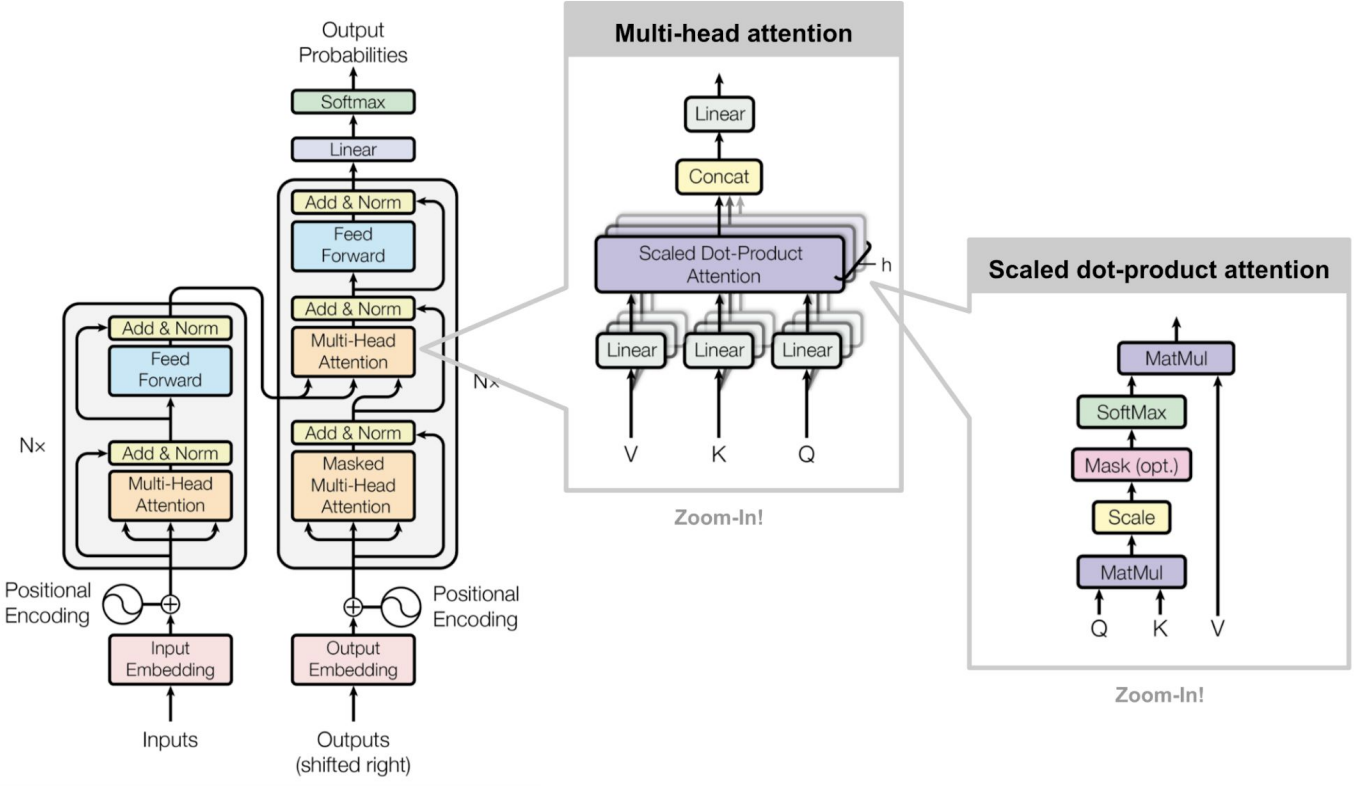


Seq2seq - sentence output sampling



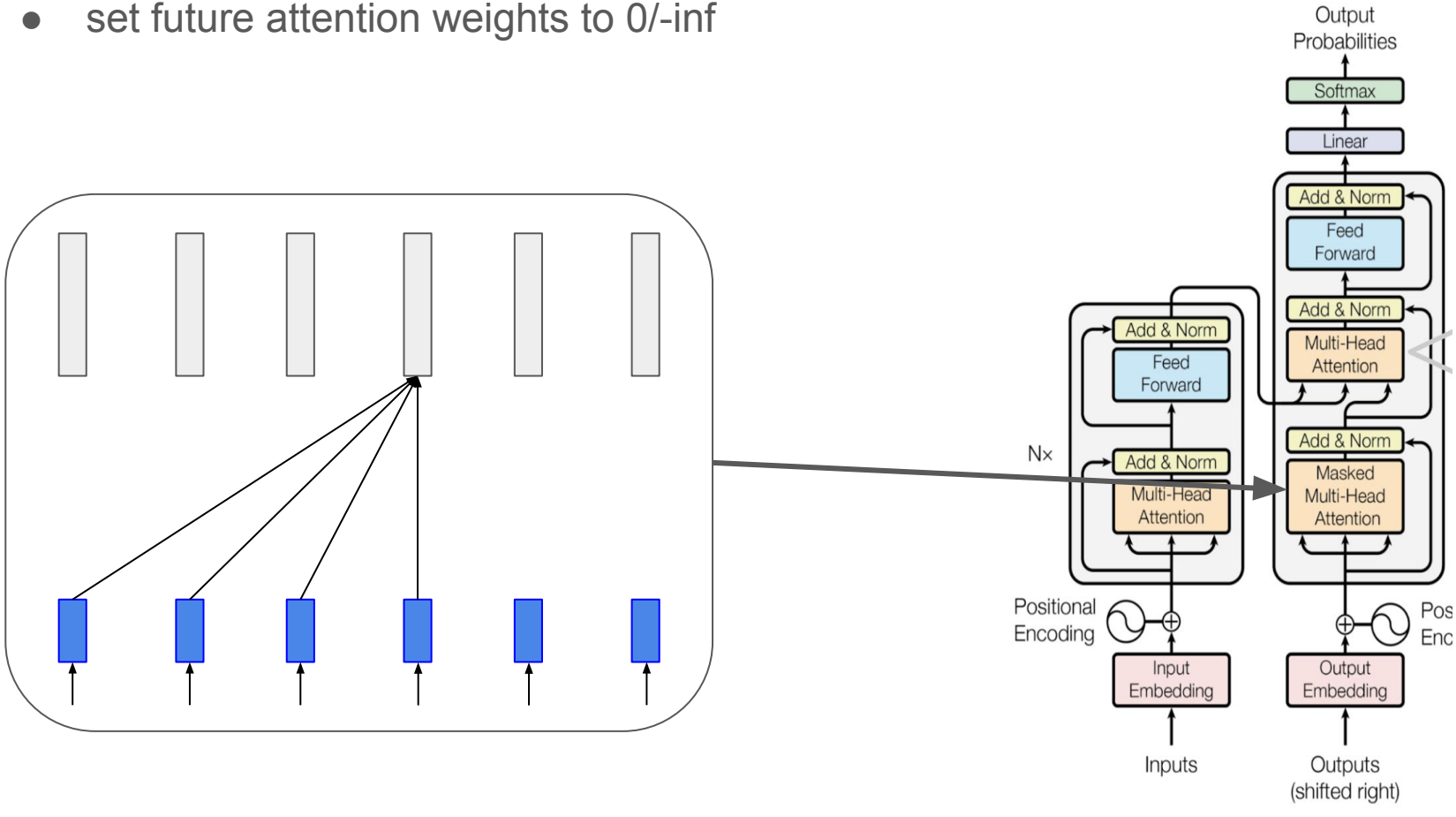
Attention Is All You Need (Vaswani et al, NIPS 2017)

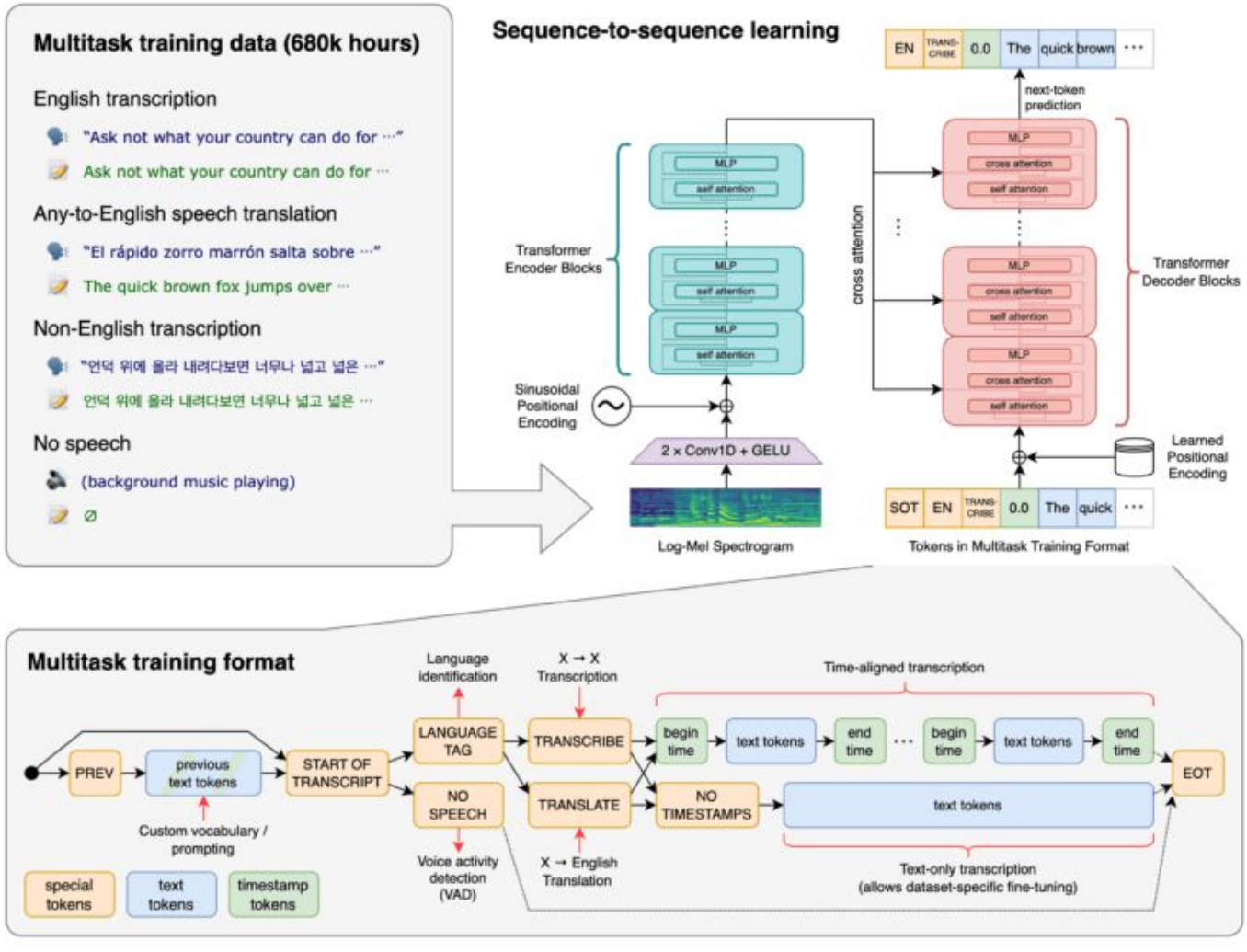
Tool: [Tensor2tensor]



Causal attention in autoregressive models

- set future attention weights to 0/-inf




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)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
⋯
next-token prediction
MLP
self attention
cross attention
Transformer Encoder Blocks
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Learned Positional Encoding
Log-Mel Spectrogram
SOT
EN
TRANS-CRIBE
0.0
The
quick
⋯
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
Custom vocabulary / prompting
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
Voice activity detection (VAD)
TRANSCRIBE
TRANSLATE
X → English Translation
begin time
text tokens
end time
NO TIMESTAMPS
EOT
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Encoder-decoder vs. decoder only

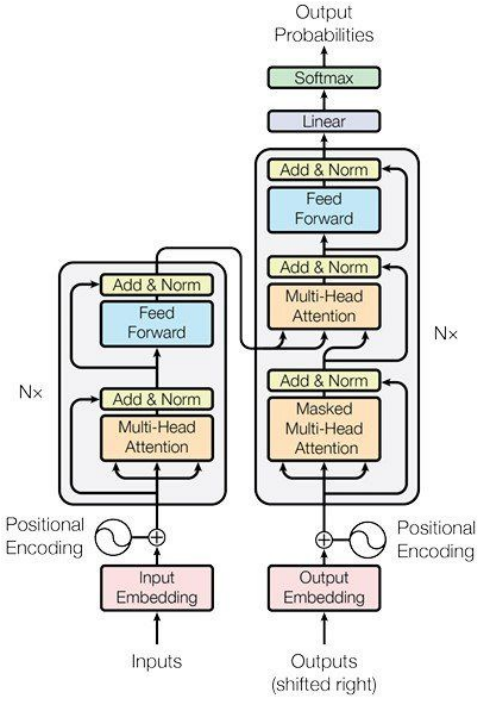
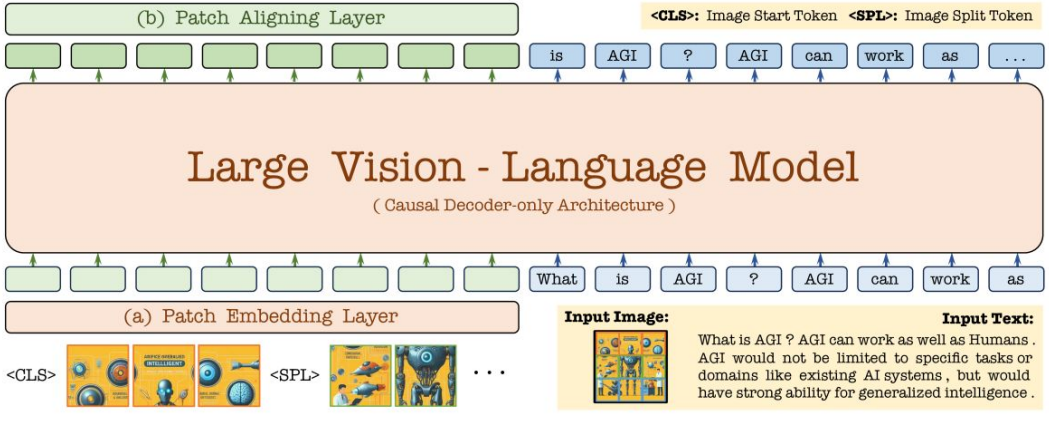
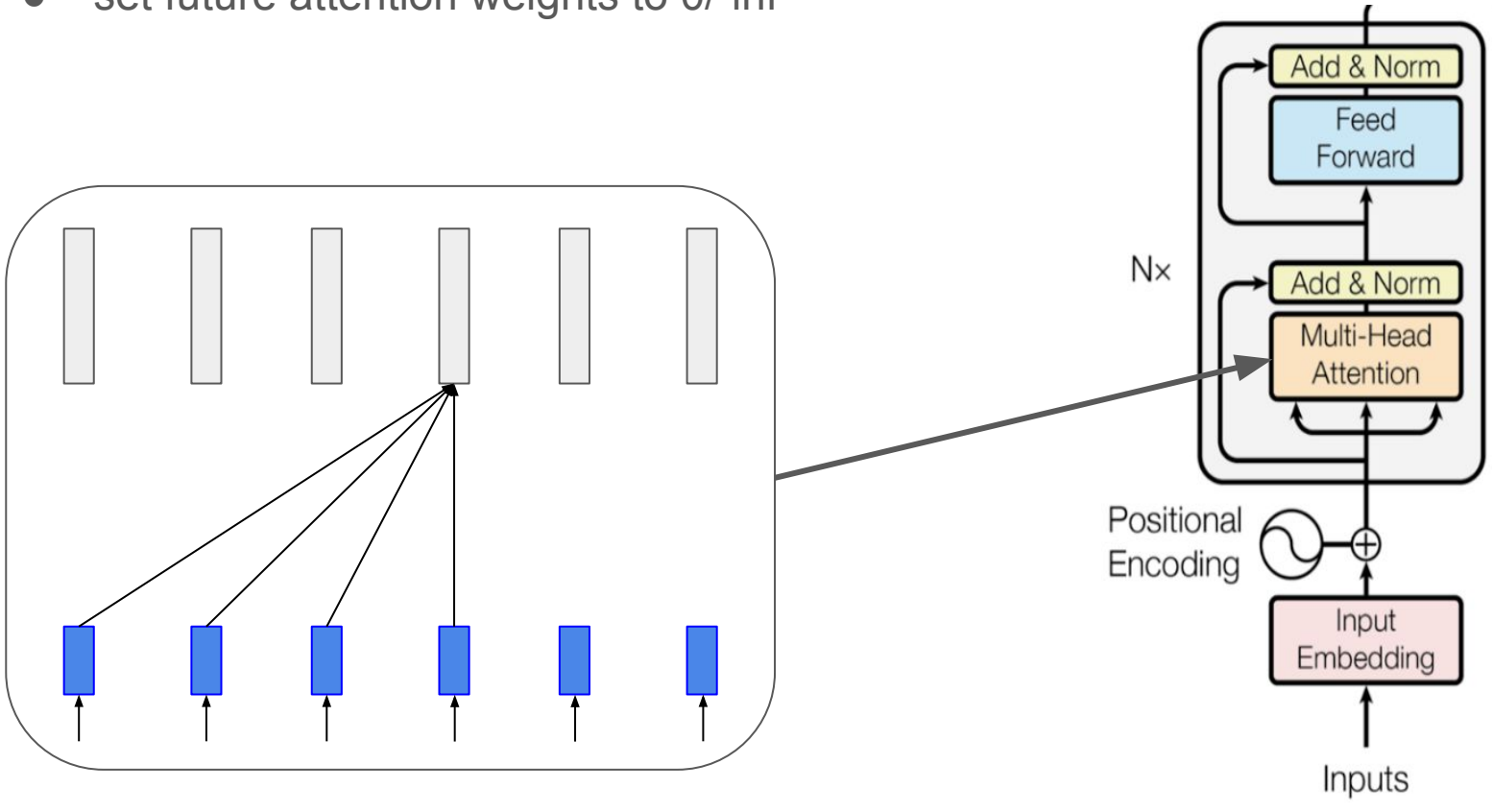


Figure 1: The Transformer - model architecture.



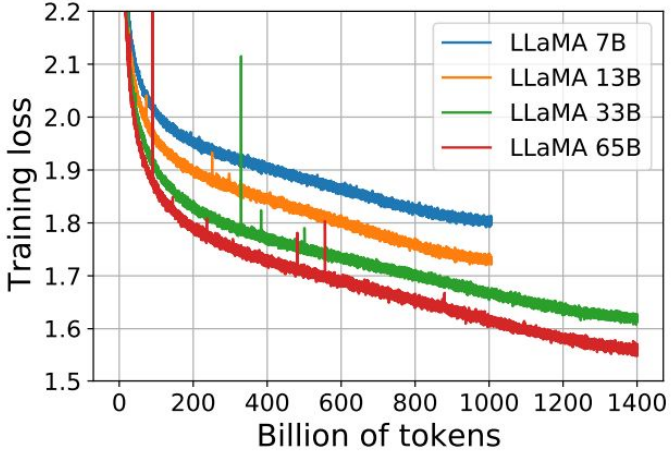
Causal attention in autoregressive Language Models

- set future attention weights to 0/-inf

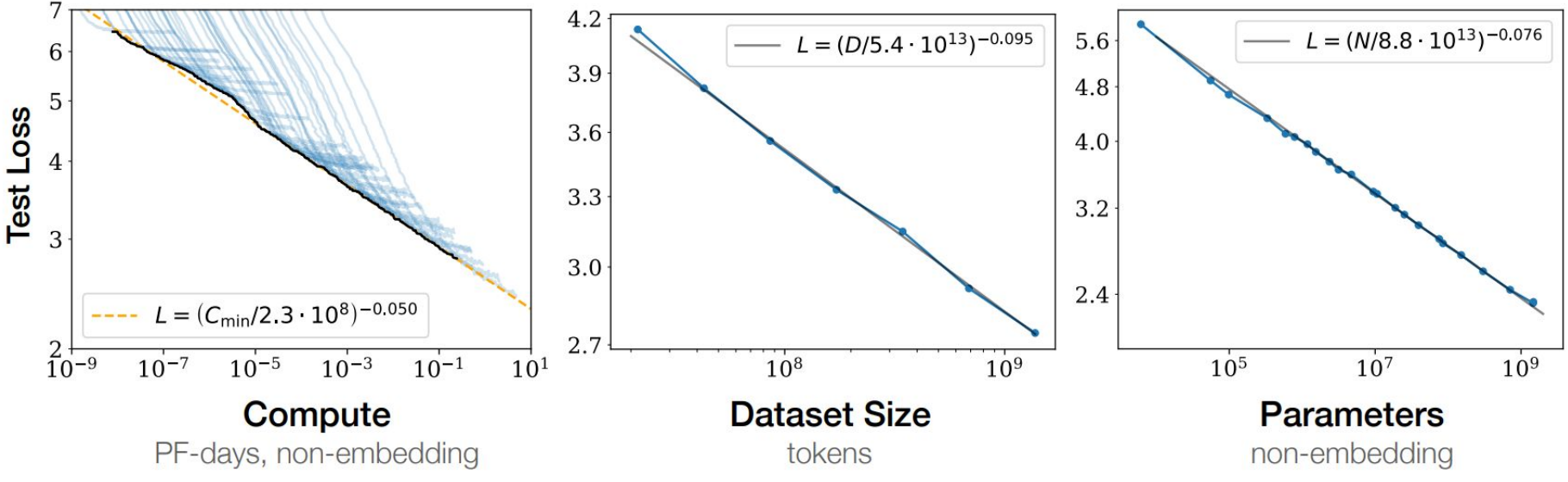


Technical challenge - training / inference

- Model does not fit on a single GPU or even single multi-GPU node
- Model parallel and data parallel training
- LLaMA:
  - 32.5B, dim. 6656, hads 52, layers 60
  - batch size 4M tokens
  - training 1.4T tokens = 350000 iterations
  - 380 tokens/sec/A100 GPU with 80GB
  - Training on 2048 A100 GPUs
  - Training time 21 days

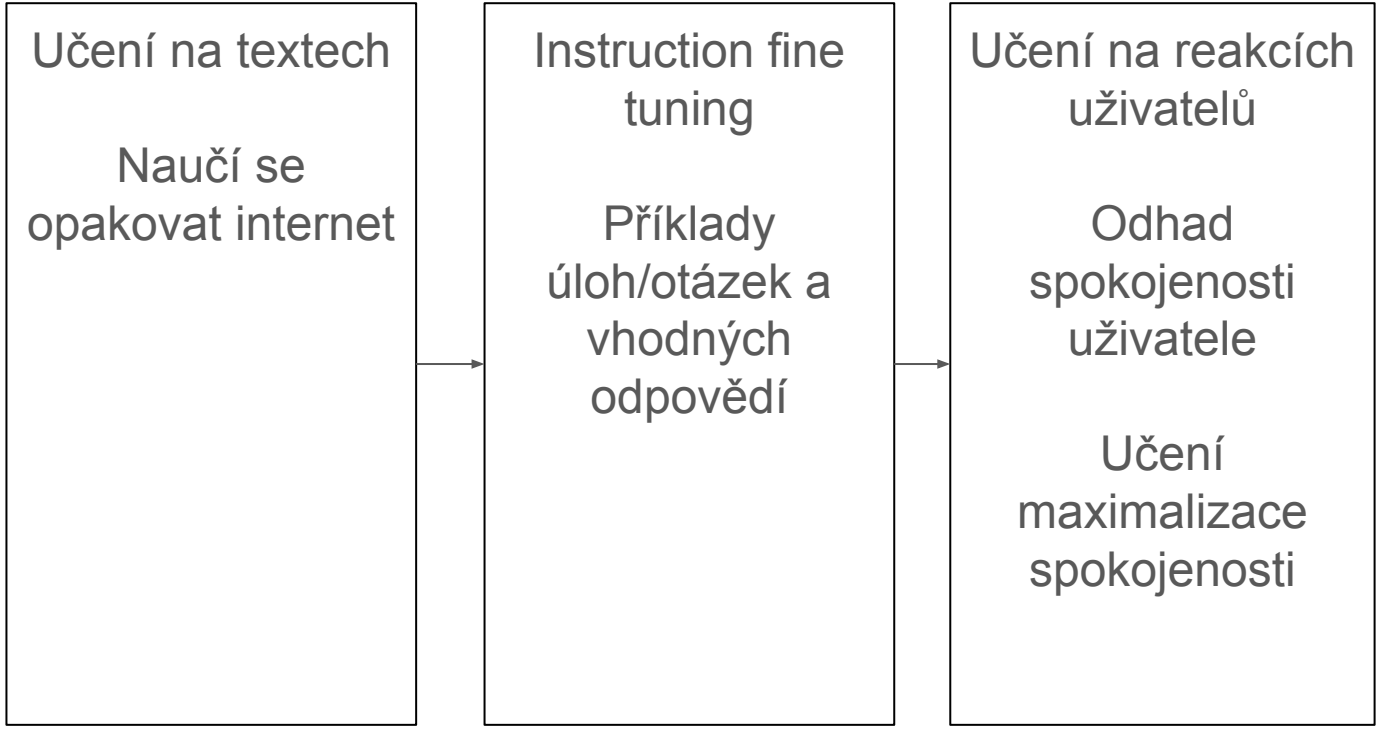


Model scaling laws

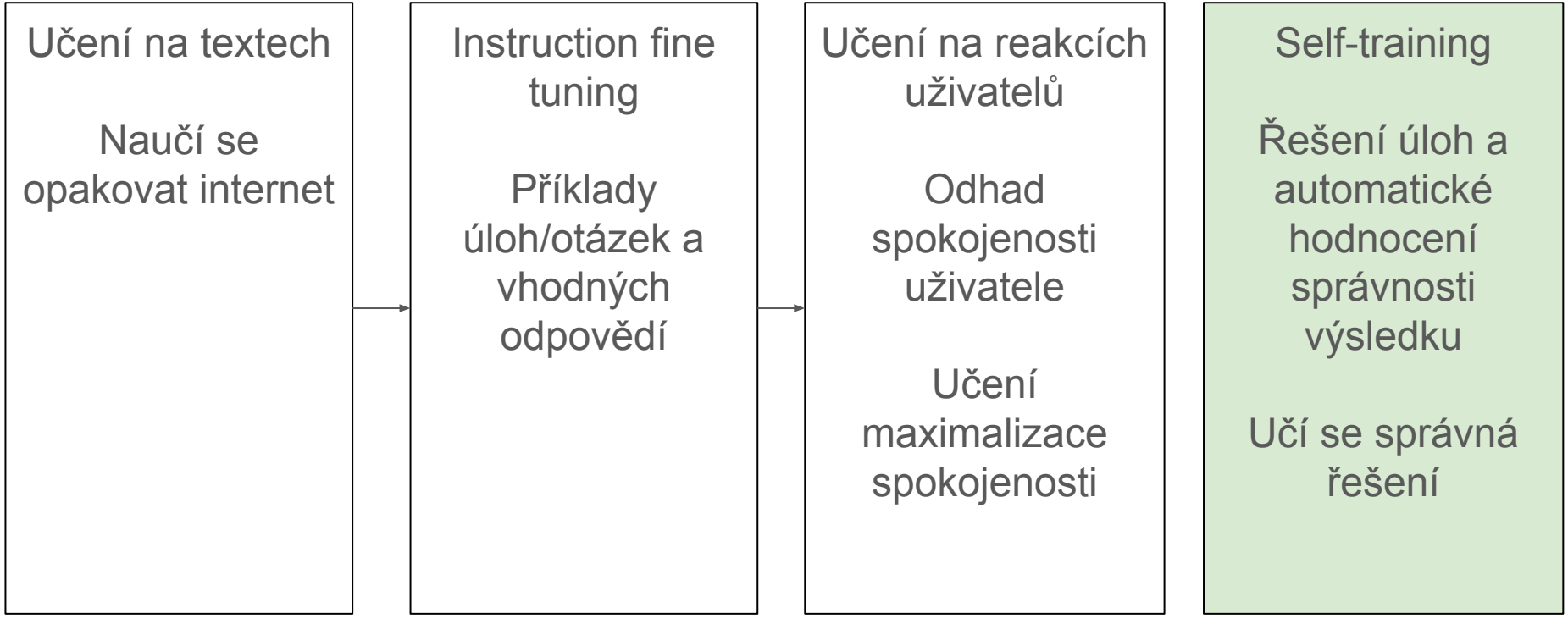


Jared Kaplan et al.: Scaling Laws for Neural Language Models, 2021.

Jazykové modely - učení



Jazykové modely - učení



Model Alignment

- We (sometimes) want the models to useful to humans
  - (and NOT to provide bad press for the companies)
  - (and NOT to cause lawsuits for the the companies)
- Some people may be:
  - offended by the models for many reasons
  - offended that the model talks to someone else about some topics
  - offended that models are biased
  - angry that models “harm children”
  - angry that models provide some “harmful” information
- If we get AGI? How will we assure that it acts in the “best interest” of humanity?

How to work with LLM directly - Ollama

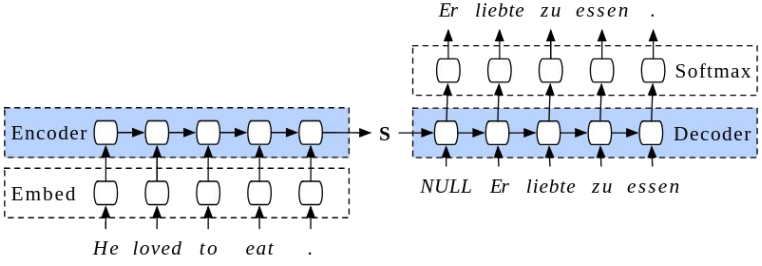
- **Ollama -** Simple installation - inference server
  - Install Ollama
  - Start ollama
  - Download model
  - Load model into memory
  - Send requests to the model
  - Receive responses of the model - process it

OpenAI api

- Create account, get credit
- Send requests from python, javascript, curl, …

NOTEBOOK - https://colab.research.google.com/drive/1B7xYG3HBEUwWAp012TA0qN4208a5oE3E?usp=sharing

Machine translation



The machine learning summer school in Brno will surely be great!

Google Translate

Letní škola strojového učení v Brně jistě bude skvělá!

브르노 (Brno)의 여름 학교를 학습하는 기계가 반드시 좋을 것입니다!

Летняя школа машинного обучения в Брно, безусловно, будет отличной!

L’école d’été d’apprentissage automatique à Brno sera sûrement géniale!

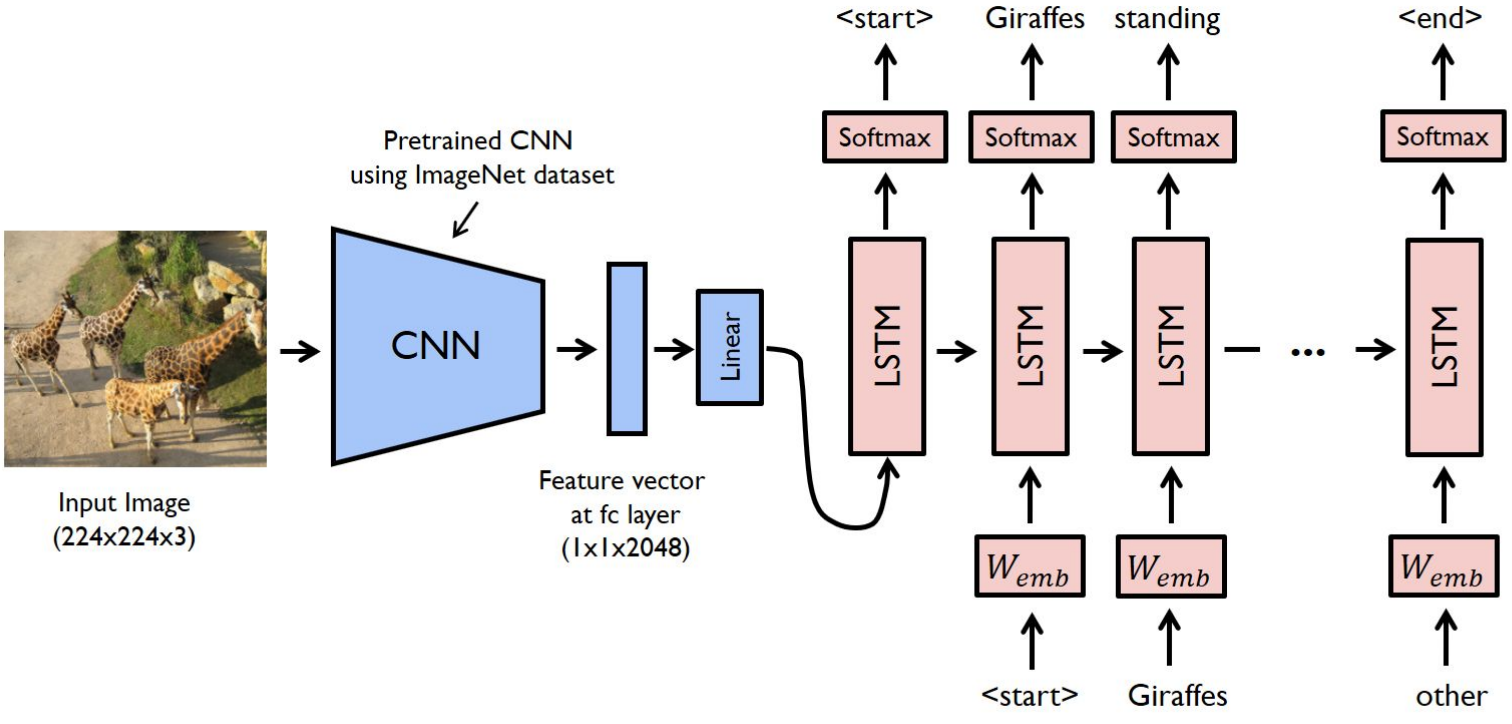
Language Tasks

- Parsing and tagging
- Language modelling
- Translation and unsupervised translation
- Coreference
- Reading comprehension
- Named entity
- What is the next sentence?

General Language Understanding Evaluation (GLUE) benchmark

- Is a sentence grammatically correct?
- Sentiment (positive/negative)
- Paraphrase: Are sentences semanticaly equivalent?
- Paraphrase: Are questions from Quora semanticaly equivalent?
- Paraphrase: news headlines, captions
- Inference: Does one sentece support a hypothesis, contradict it or is it neutral?
- Question aswering: question-paragraph pairs
- What does a pronoun in sentence correspond to?
- ...

Image2Seq



https://www.captionbot.ai/

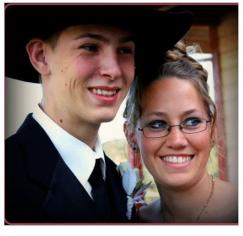
Visual question answering

VQA Challenge 2019

Who is wearing glasses?

man

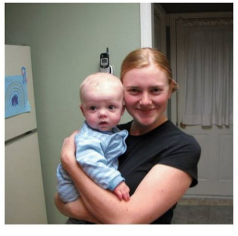
woman



Where is the child sitting?

fridge

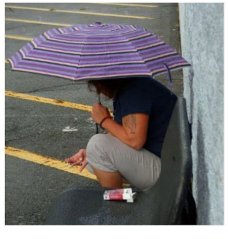
arms



Is the umbrella upside down?

yes

no



How many children are in the bed?

2

1