Lecture 07: Question Answering

Anastasia Ianina

Harbour.Space University
21.06.2022

Outline

1. Question Answering
   - SQuAD, SberQuAD
   - Open-Domain Question Answering
2. More on GPT
   - miniGPT and GPT-3 for Russian language

Based on: http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture13-contextual-representations.pdf
https://jalammar.github.io/illustrated-transformer/
http://jalammar.github.io/illustrated-bert/
https://medium.com/mlreview/understanding-building-blocks-of-ulmfit-818d3775325b

A Brief History of Open-domain Question Answering

- Simmons et al. (1964) did first exploration of answering questions from an expository text based on matching dependency parses of a question and answer
- Murax (Kupiec 1993) aimed to answer questions over an online encyclopedia using IR and shallow linguistic processing
- The NIST TREC QA track begun in 1999 first rigorously investigated answering fact questions over a large collection of documents
- IBM's Jeopardy! System (DeepQA, 2011) brought attention to a version of the problem; it used an ensemble of many methods
- DrQA (Chen et al. 2016) uses IR followed by neural reading comprehension to bring deep learning to Open-domain QA

MCTest Reading Comprehension

Passage (P) + **Question (Q)** ⟶ **Answer (A)**

P Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house…….

Q Why did Alyssa go to Miami?   **A** To visit some friends

Stanford Question Answering Dataset (SQuAD)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

- What causes precipitation to fall?
  - gravity
- What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
  - graupel
- Where do water droplets collide with ice crystals to form precipitation?
  - within a cloud

source: SQuAD: 100,000+ Questions for Machine Comprehension of Text

SQuAD evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1: Take system and each gold answer as bag of words, evaluate Precision, Recall and harmonic mean F1.

    Score is (macro-)average of per-question F1 scores

- F1 measure is seen as more reliable and taken as primary
  - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a, an, the** only)

SQuAD v1.1 leaderboard, end of 2016

		EM	F1
11	Fine-Grained Gating Carnegie Mellon University (Yang et al. '16)	62.5	73.3
12	Dynamic Chunk Reader IBM (Yu & Zhang et al. '16)	62.5	71.0
13	Match-LSTM with Ans-Ptr (Boundary) Singapore Management University (Wang & Jiang '16)	60.5	70.7
14	Match-LSTM with Ans-Ptr (Sequence) Singapore Management University (Wang & Jiang '16)	54.5	67.7
15	Logistic Regression Baseline Stanford University (Rajpurkar et al. '16)	40.4	51.0

Will your model outperform humans on the QA task?

		EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.3	91.2

SQuAD v1.1 leaderboard, (May 2020)

Rank		Model	EM	F1
		Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1	Apr 10, 2020	LUKE (single model) *Studio Ousia & NAIST & RIKEN AIP*	**90.202**	**95.379**
2	May 21, 2019	XLNet (single model) *Google Brain & CMU*	89.898	95.080
3	Dec 11, 2019	XLNET-123++ (single model) *MST/EOI* http://tia.today	89.856	94.903
3	Aug 11, 2019	XLNET-123 (single model) *MST/EOI*	89.646	94.930
4	Sep 25, 2019	BERTSP (single model) *NEUKG* http://www.techkg.cn/	88.912	94.584
4	Jul 21, 2019	SpanBERT (single model) *FAIR & UW*	88.839	94.635
5	Jul 03, 2019	BERT+WWM+MT (single model) *Xiaoi Research*	88.650	94.393

source: SQuAD website

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For `NoAnswer` examples, `NoAnswer` receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
  - Like Natural Language Inference (NLI) or "Answer validation"

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

Prediction: 1234   [from Microsoft nlnet]

SQuAD 2.0 leaderboard (October 2020)

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799
4 Sep 11, 2020	EntitySpanFocus+AT (ensemble) *RICOH_SRCB_DML*	90.454	92.748
4 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
5 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.115	92.580
5 Sep 27, 2020	electra+nlayers (ensemble) *oppo.tensorlab*	90.126	92.535

source: SQuAD website

Now in Russian: SberQuAD

Термин Computer science (Компьютерная наука) появился в 1959 году в научном журнале Communications of the ACM, в котором Луи Фейн (Louis Fein) ратовал за создание Graduate School in Computer Sciences (Высшей школы в области информатики) . . . Усилия Луи Фейна, численного аналитика Джорджа Форсайта и других увенчались успехом: университеты пошли на создание программ, связанных с информатикой, начиная с Университета Пердью в 1962.

- Q11870 Когда впервые был применен термин Computer science (Компьютерная наука)?
- Q28900 Кто впервые использовал этот термин?
- Q30330 Начиная с *каого** учебного заведения стали применяться учебные программы, связанные с информатикой?

*Misspelling is intended

SberQuAD evaluation

Model	SberQuAD		SQuAD	
	EM	F1	EM	F1
simple baseline	0.3	25.0	–	–
ML baseline	3.7	31.5	–	–
BiDAF	51.7	72.2	68.0	77.3
DrQA	54.9	75.0	70.0	79.0
R-Net	58.6	77.8	71.3	79.7
DocQA	59.6	79.5	72.1	81.1
BERT	66.6	84.8	85.1	91.8

Table 7: Model performance on SQuAD and SberQuAD; SQuAD part shows single-model scores on test set taken from respective papers.
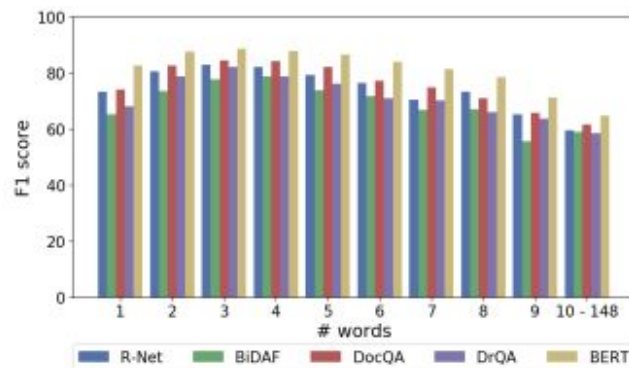
Figure 6: Model performance depending on answer length (# of words).

% test	R-Net	BiDAF	DocQA	DrQA	BERT
w/ typos	5.7	74.1	66.7	77.5	67.5	81.1
correct	94.3	77.1	72.5	79.6	75.4	85.0
Test set		77.8	72.2	79.5	75.0	84.8

Table 8: Answer quality for misspelled questions.

source:

But errors are still present

The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire
③ the Song dynasty

Prediction: Ming dynasty   [BERT (single model) (Google AI)]

S(ber)QuAD limitations

- Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at the passages
  - Not genuine information needs
  - Generally greater lexical and syntactic matching between questions and answer span than you get IRL
  - Barely any multi-fact/sentence inference beyond coreference

- But these datasets are still of a great use

Potential solutions

Passage

$$\alpha_i = \mathrm{softmax}_i \, \mathbf{q}^\mathsf{T} \mathbf{W}_s \tilde{\mathbf{p}}_i$$

$$\mathbf{o} = \sum_i \alpha_i \tilde{\mathbf{p}}_i$$

Two attention heads are used to find the *start* and *end* of the answer.

Attention is computed between encoded question and RNN state corresponding to every position.
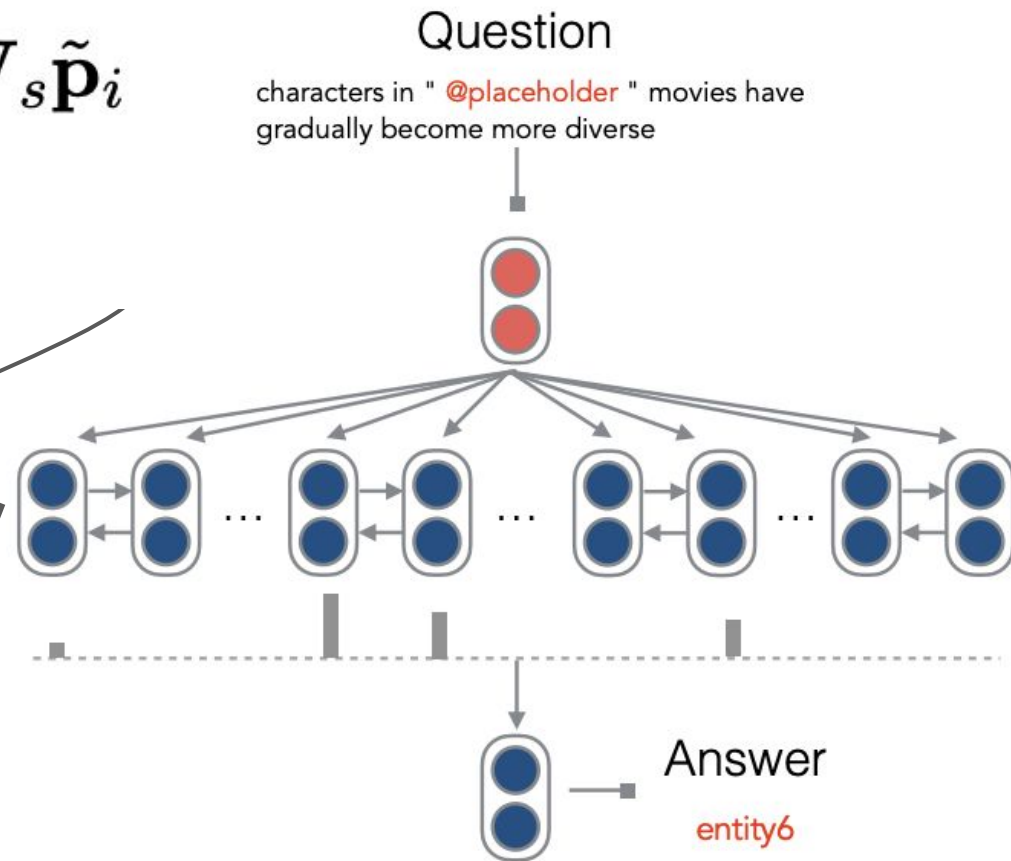
Question

characters in " @placeholder " movies have gradually become more diverse

Answer

entity6

- Use extra information about the text
  - Char embeddings
  - Linguistic features: PoS and NER tags
  - ...

PoS tagging

Pred. Tag	Actual Tag	Correct?	Token
PUNCT	PUNCT	✔	[
DET	DET	✔	this
NOUN	NOUN	✔	killing
ADP	ADP	✔	of
DET	DET	✔	a
ADJ	ADJ	✔	respected
NOUN	NOUN	✔	cleric
AUX	AUX	✔	will
AUX	AUX	✔	be
VERB	VERB	✔	causing
PRON	PRON	✔	us
NOUN	NOUN	✔	trouble
ADP	ADP	✔	for
NOUN	NOUN	✔	years
PART	PART	✔	to
VERB	VERB	✔	come
PUNCT	PUNCT	✔	.
PUNCT	PUNCT	✔	]

- PoS tagging can be performed using
  - Rule-based taggers
  - Dynamic programming
  - Models based on CRF (Conditional Random Field)
  - Neural Networks
  - etc.

- Use extra information about the text
  - Char embeddings
  - Linguistic features: PoS and NER tags
  - ...

- Better use of attention

BiDAF

source: Bidirectional Attention Flow for Machine Comprehension
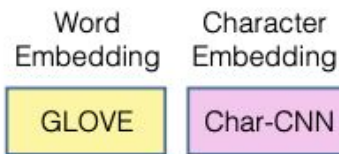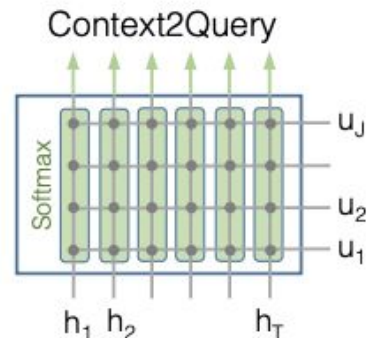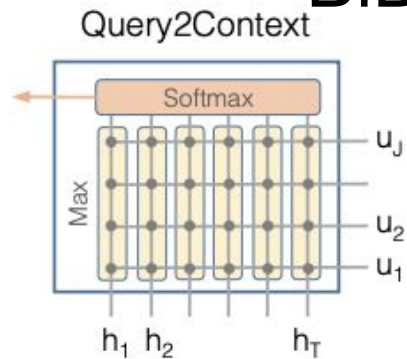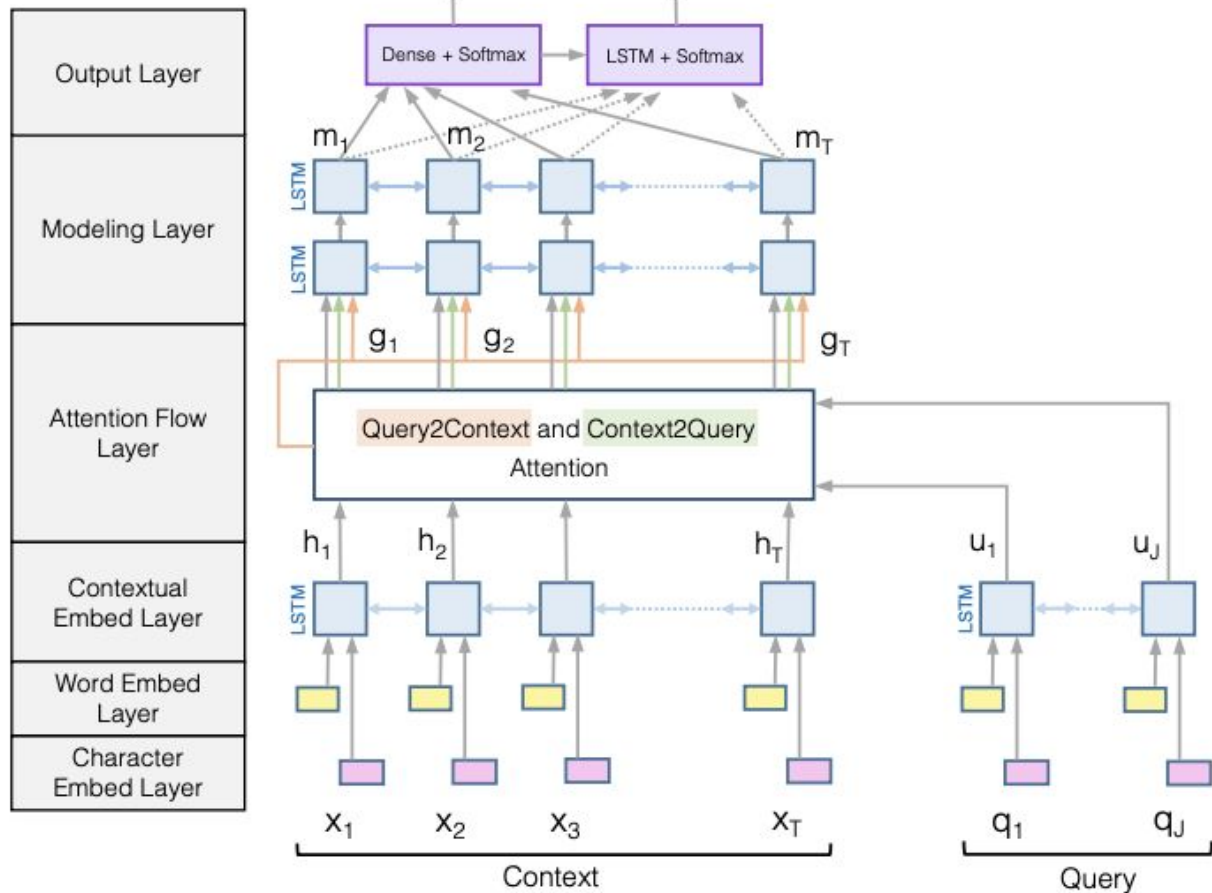
- There are variants of and improvements to the BiDAF architecture, but **the central idea** is the **Attention Flow layer**: attention should flow both ways – from the context to the question and from the question to the context
- Make similarity matrix (with w of dimension 6d):

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^{T}[\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

- Context-to-Question (C2Q) attention (which query words are most relevant to each context word):

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$

BiDAF

- Attention Flow:
  - attention should flow both ways – from the context to the question and from the question to the context
- Question-to-Context (Q2C) attention:
  - the weighted sum of the most important words in the context with respect to the query – slight asymmetry through max

$$m_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \ldots, N\}$$

$$\beta = \mathrm{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^2$$

- For each passage position, output of BiDAF layer is:

$$\boldsymbol{b}_i = [\boldsymbol{c}_i; \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{c}'] \in \mathbb{R}^{8h} \quad \forall i \in \{1, \ldots, N\}$$
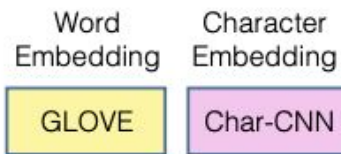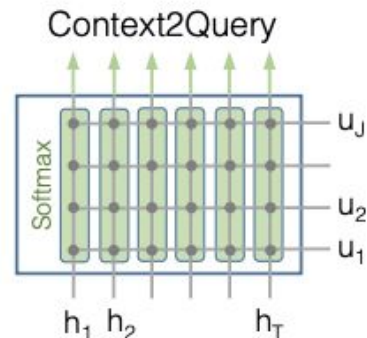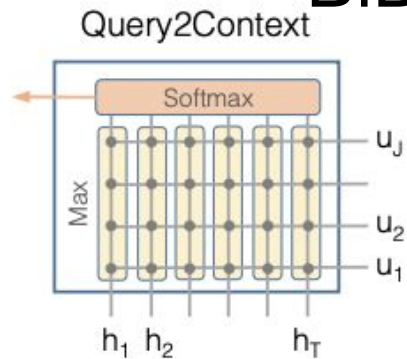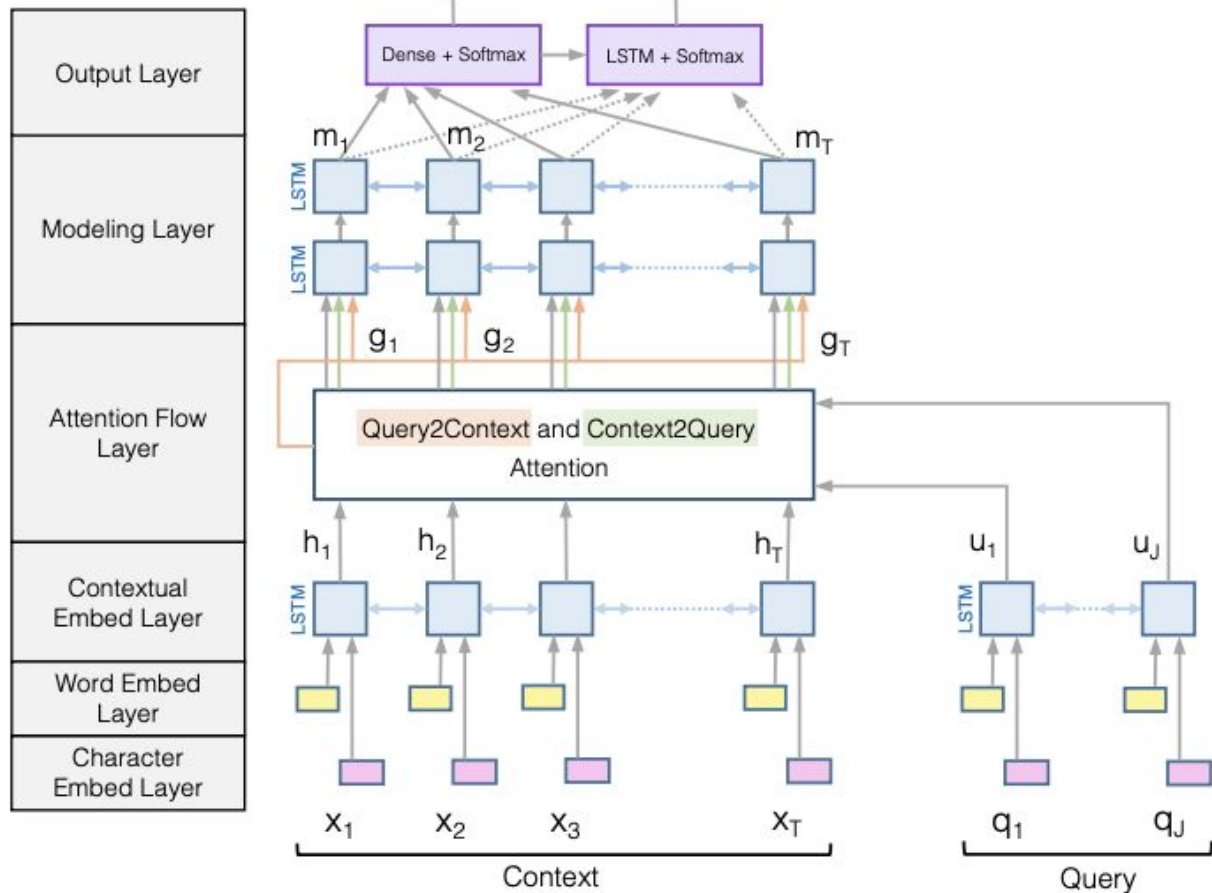
source: CS224n Lecture 10

BiDAF

source: Bidirectional Attention Flow for Machine Comprehension

Open-Domain Question Answering

Q: How many of Warsaw's inhabitants
spoke Polish in 1933?

**Document
Retriever**

**Document
Reader**

833,500

WIKIPEDIA
The Free Encyclopedia

25

source: https://github.com/facebookresearch/DrQA

Possible problems

Example question: "What is the Aquaman actor's next movie?"

source: http://ai.stanford.edu/blog/answering-complex-questions/

Possible solutions

Which novel by the author of "Armada" will be adapted as a feature film by Steven Spielberg?

Query Generator 1 → novel by the author of "Armada" → search → WIKIPEDIA

Paragraph Set 1
- Armada (novel)
- Author, Author (novel)
- Armada
- ⋮

Query Generator 2 → Ernest Cline → search → WIKIPEDIA

Paragraph Set 2
- Ernest Cline
- Cline (biology)
- Patsy Cline (album)
- ⋮

QA Model → *Ready Player One*

source: http://ai.stanford.edu/blog/answering-complex-questions/

More on GPT

GPT-3, May 2020

Proportions are not preserved for visual sake

Number of trainable parameters, millions
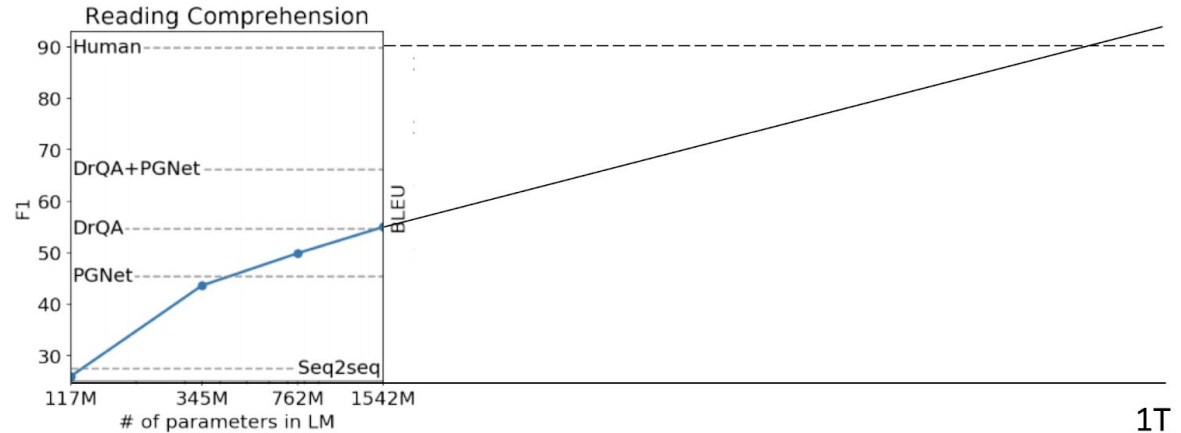
Hypothesis from Stanford CS224N Lecture 20 (2019)

Reaction: GPT-3

Sam Altman ✓ @sama · Jul 19
The GPT-3 hype is way too much. It's impressive (thanks for the nice compliments!) but it still has serious weaknesses and sometimes makes very silly mistakes. AI is going to change the world, but GPT-3 is just a very early glimpse. We have a lot still to figure out.

💬 148   🔁 963   ♡ 7K   ⬆

Paul Graham ✓ @paulg · Jul 19
Hackers are fascinated by GPT-3. To everyone else it seems a toy.

Pattern seem familiar to anyone?

💬 153   🔁 604   ♡ 5K   ⬆

Source: Twitter, more links available at https://lambdalabs.com/blog/gpt-3/

Reaction: GPT-3

Andriy Burkov · Following
ML at Gartner, author of The Hundred-Page Machine Learning Book
2d · 🌐

GPT-3 is the closest thing to artificial general intelligence (AGI) that I ever saw.
It's so strong that it makes me nervous.

Andrej Karpathy ✔ @karpathy · Jul 19
The transformer architecture of GPT upper bounds its ability at memorization. It cannot learn many algorithms due to the functional form of its forward pass, and spends a fixed compute per token - i.e. it can't "think for a while". Progress here critical, likely but non-trivial.
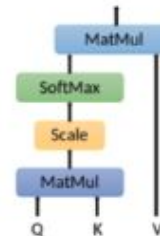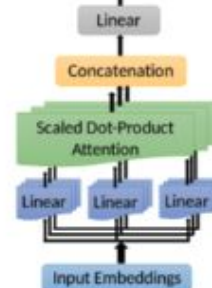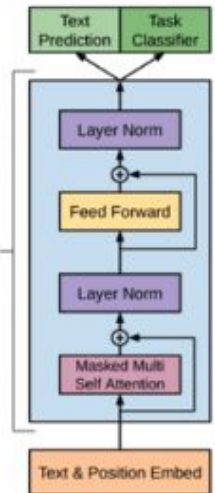
PRE-SOFTWARE:
SPECIAL-PURPOSE
COMPUTER
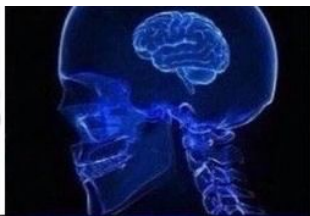
SOFTWARE 1.0:
DESIGN
THE ALGORITHM

SOFTWARE 2.0:
DESIGN
THE DATASET

SOFTWARE 3.0:
DESIGN
THE PROMPT

Andrej Karpathy ✔
@karpathy

Replying to @ch402

Love the idea for Software 3.0 😂. Programming moving from curating datasets to curating prompts to make the meta learner "get" the task it's supposed to be doing. LOL 🤣👌

source: Andrej Karpathy twitter

- Great way to understand GPT
  - ~300 lines of code and examples

available GPT implementations

minGPT

source: miniGPT by Andrej Karpathy

Speech recognition and generation

ASR and TTS

source: Medium post
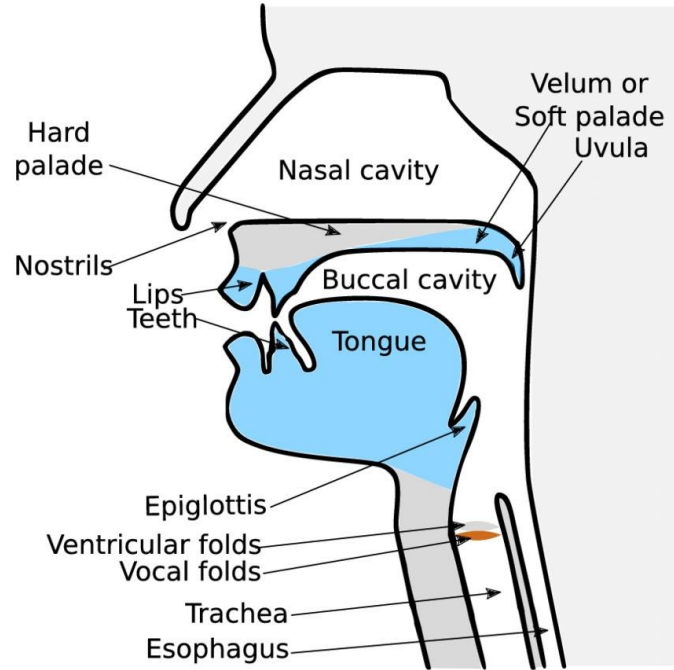
Foundations: ear structure

source: [Medium post](#)
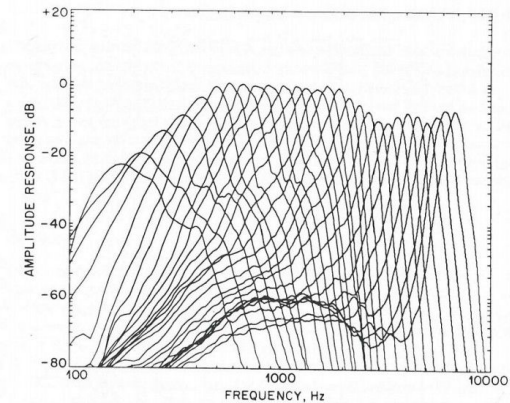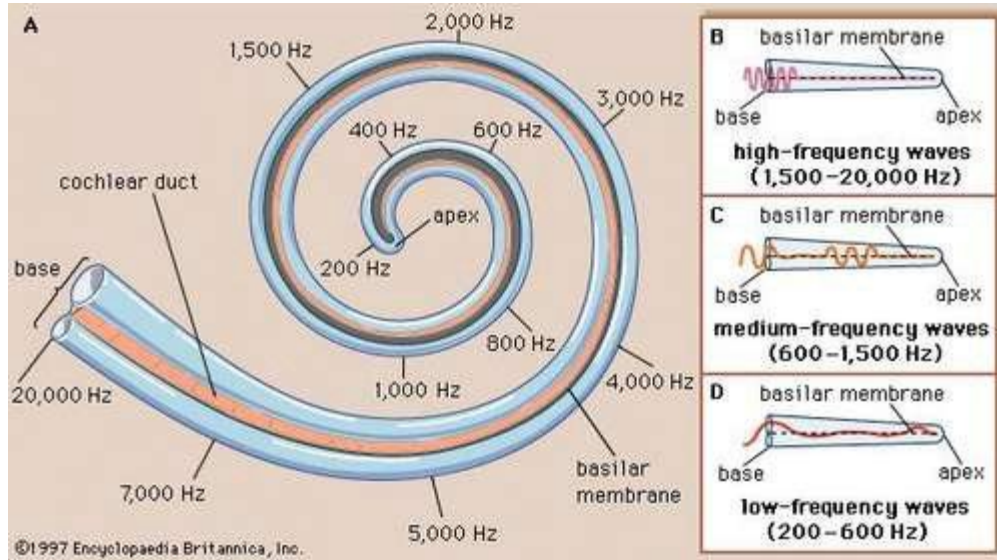
Figure 3.50   Frequency response curves of a cat's basilar membrane (after Ghitza [13]).

source: Medium post

$$A \sim 20 - 25 \text{ ms}$$

$$B \sim 10 \text{ ms}$$
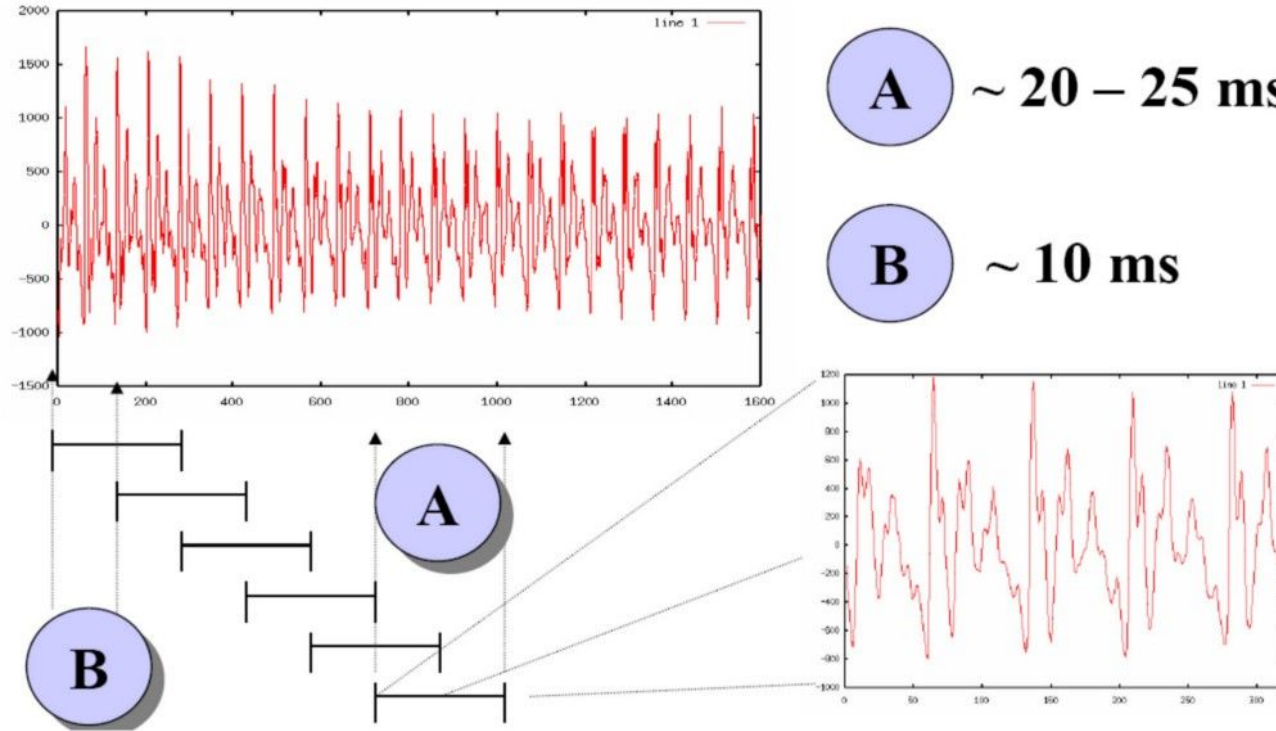
Image from Bryan Pellom

source: Medium post

Fourier Transform

Definition [edit]

The *discrete Fourier transform* transforms a sequence of N complex numbers $\{\mathbf{x_n}\} := x_0, x_1, \ldots, x_{N-1}$ into another sequence of complex numbers, $\{\mathbf{X_k}\} := X_0, X_1, \ldots, X_{N-1}$, which is defined by

$$
\begin{aligned}
X_k &= \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N}kn} \\
&= \sum_{n=0}^{N-1} x_n \cdot \left[\cos\left(\frac{2\pi}{N}kn\right) - i \cdot \sin\left(\frac{2\pi}{N}kn\right) \right],
\end{aligned}
\quad \textbf{(Eq.1)}
$$

where the last expression follows from the first one by Euler's formula.

The transform is sometimes denoted by the symbol $\mathcal{F}$, as in $\mathbf{X} = \mathcal{F}\{\mathbf{x}\}$ or $\mathcal{F}(\mathbf{x})$ or $\mathcal{F}\mathbf{x}$.[A]
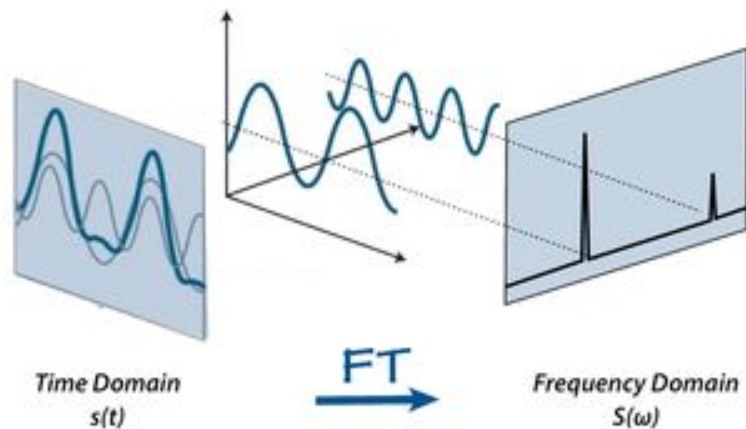
Time Domain
s(t)

FT

Frequency Domain
S(ω)

source: Wikipedia and Medium post

Fourier Transform

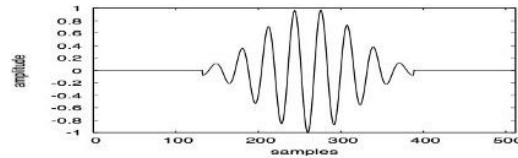source: Wikipedia and Medium post

Fourier Transform: filtering

source: [Medium post](#)
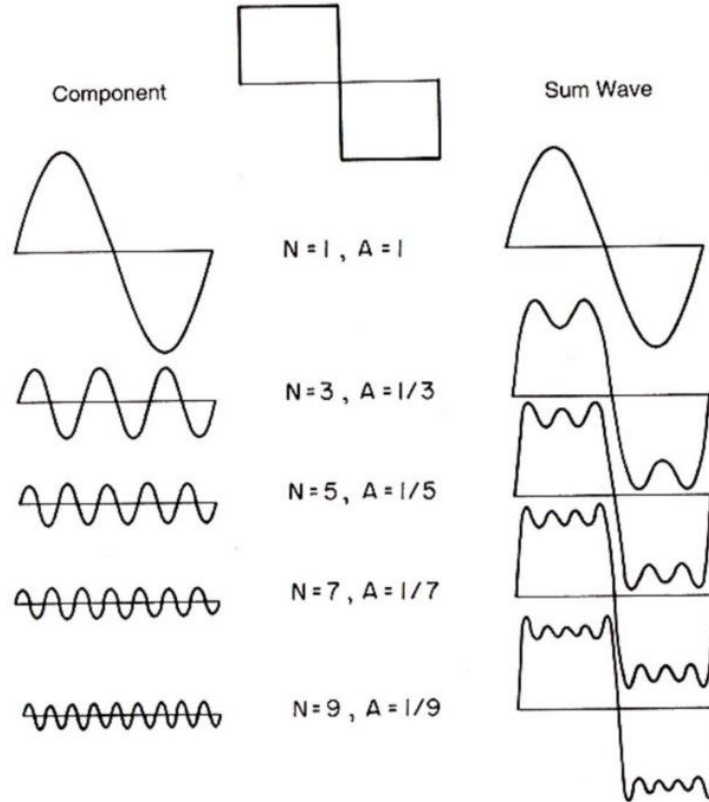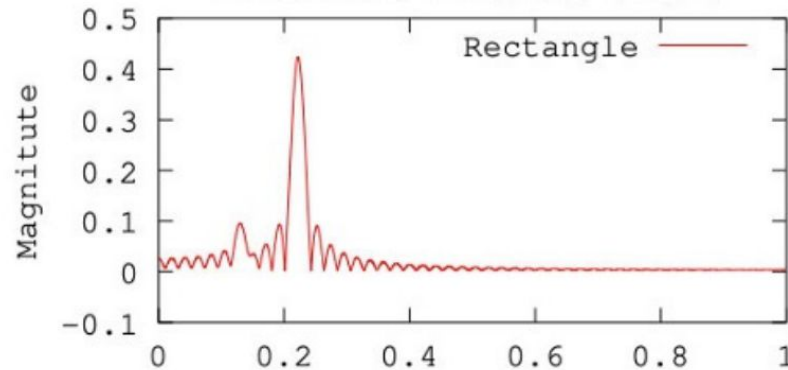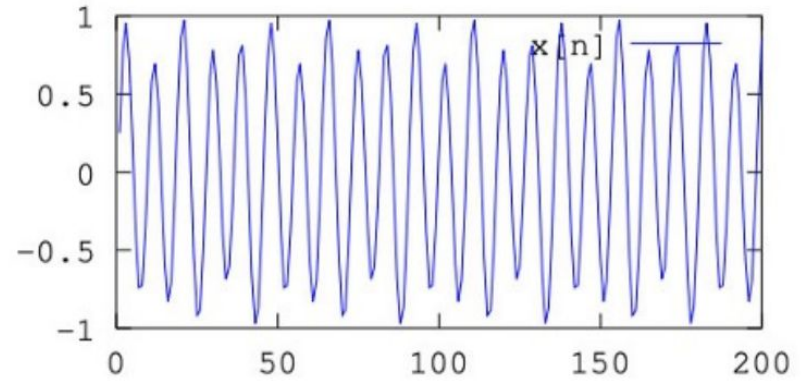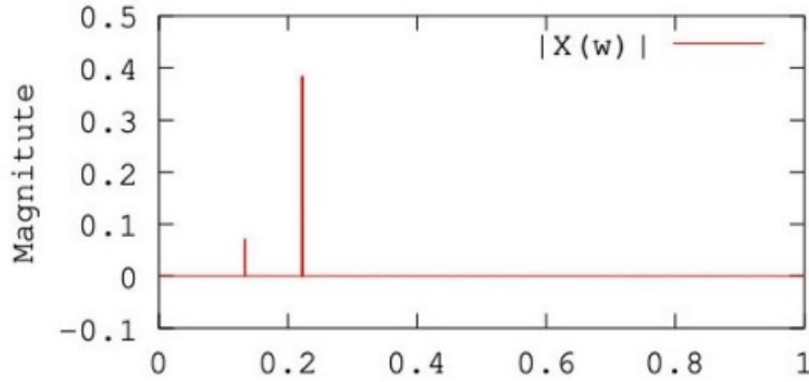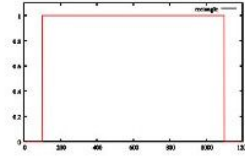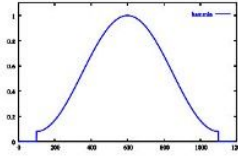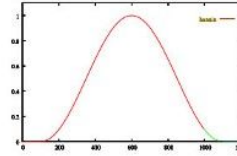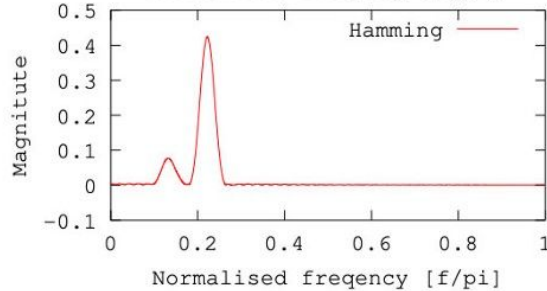
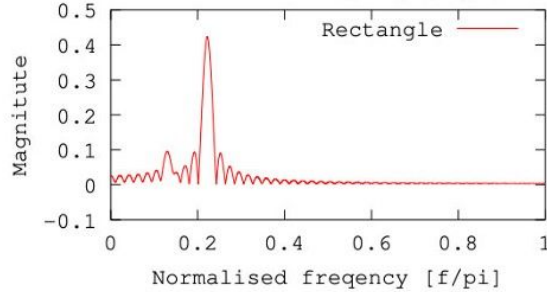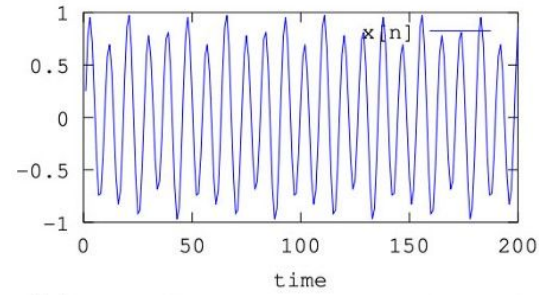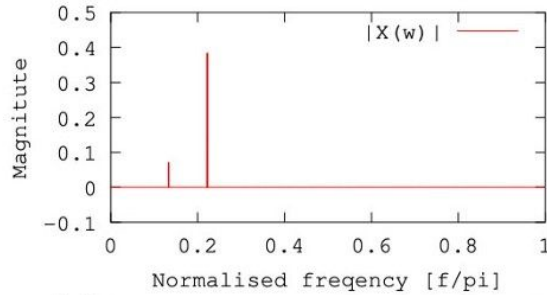Rectangular   Hamming   Hanning
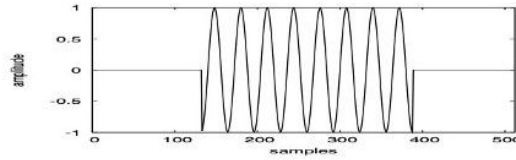
(a) Rectangular window
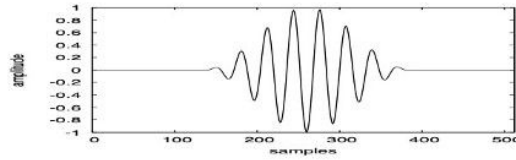
(b) Hanning window

(c) Hamming window
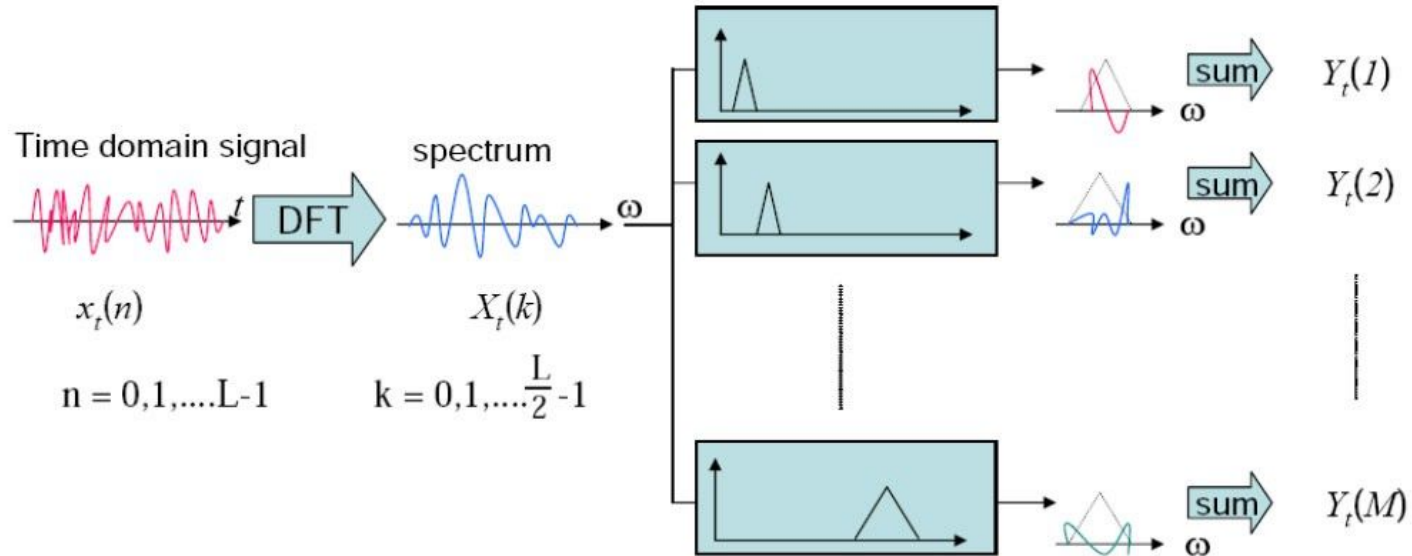
(Taylor, fig 12.1)

42

Fourier Transform: filtering

source: Medium post

DFT(STFT) power spectrum $|X[k]|^2$

Triangular band–pass filters
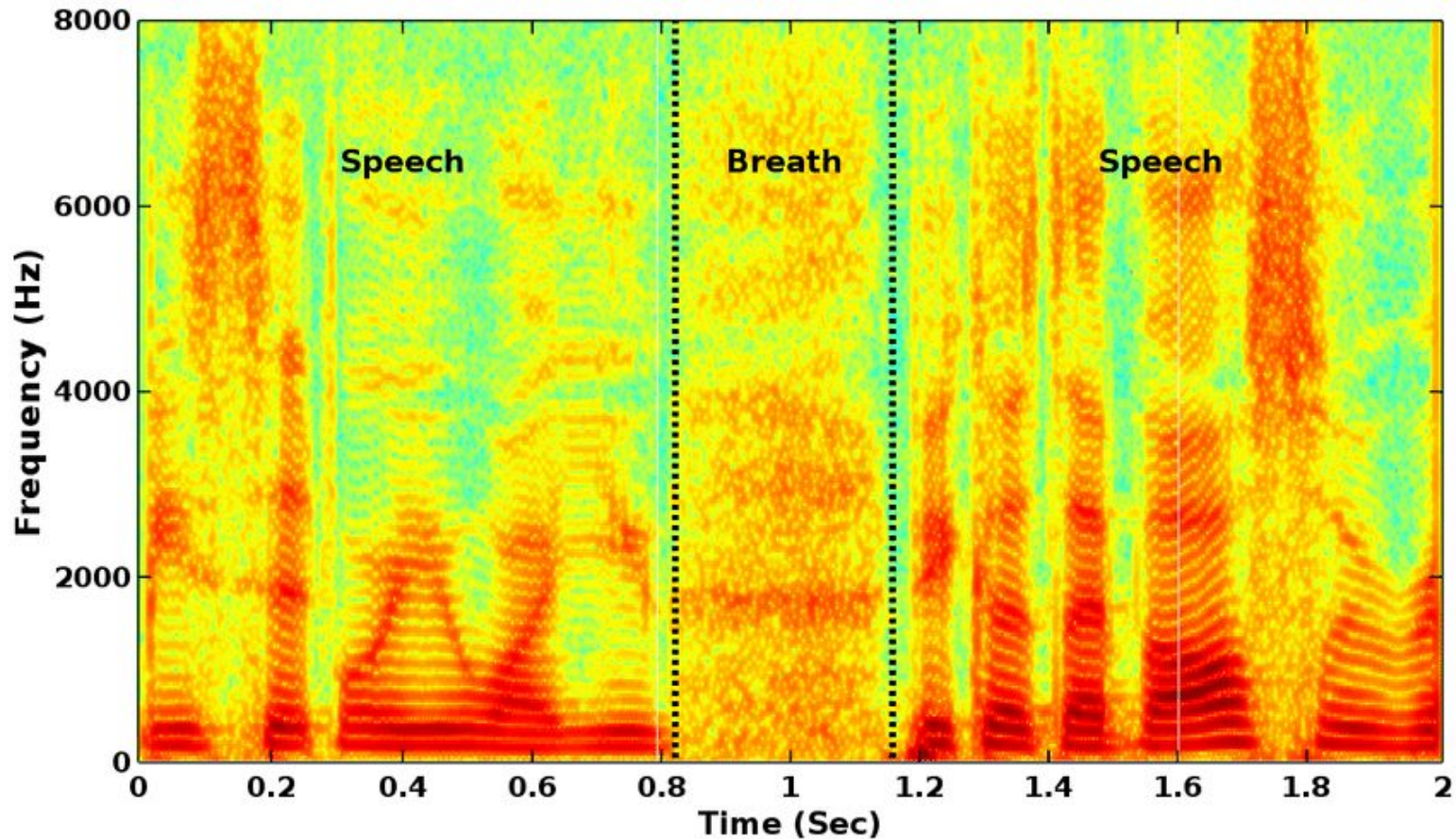
→ **Frequency bins**

Mel–scale power spectrum $Y[m]$
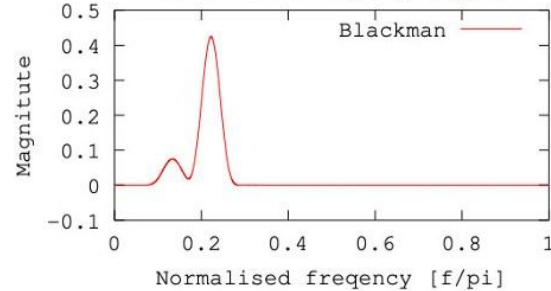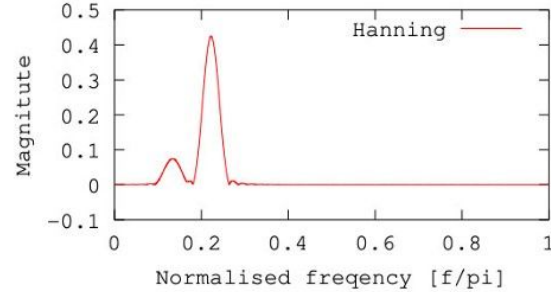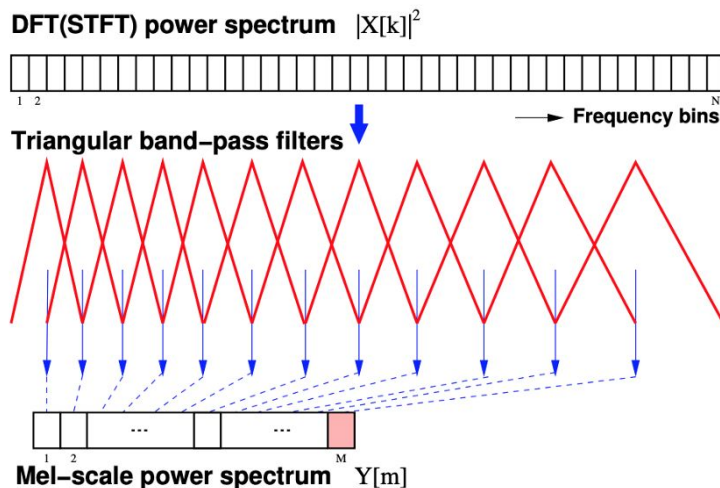
$$Y_t[m] = \sum_{k=1}^{N} W_m[k] \, |X_t[k]|^2$$

where

k : DFT bin number $(1, \ldots, N)$

m : mel-filter bank number $(1, \ldots, M)$

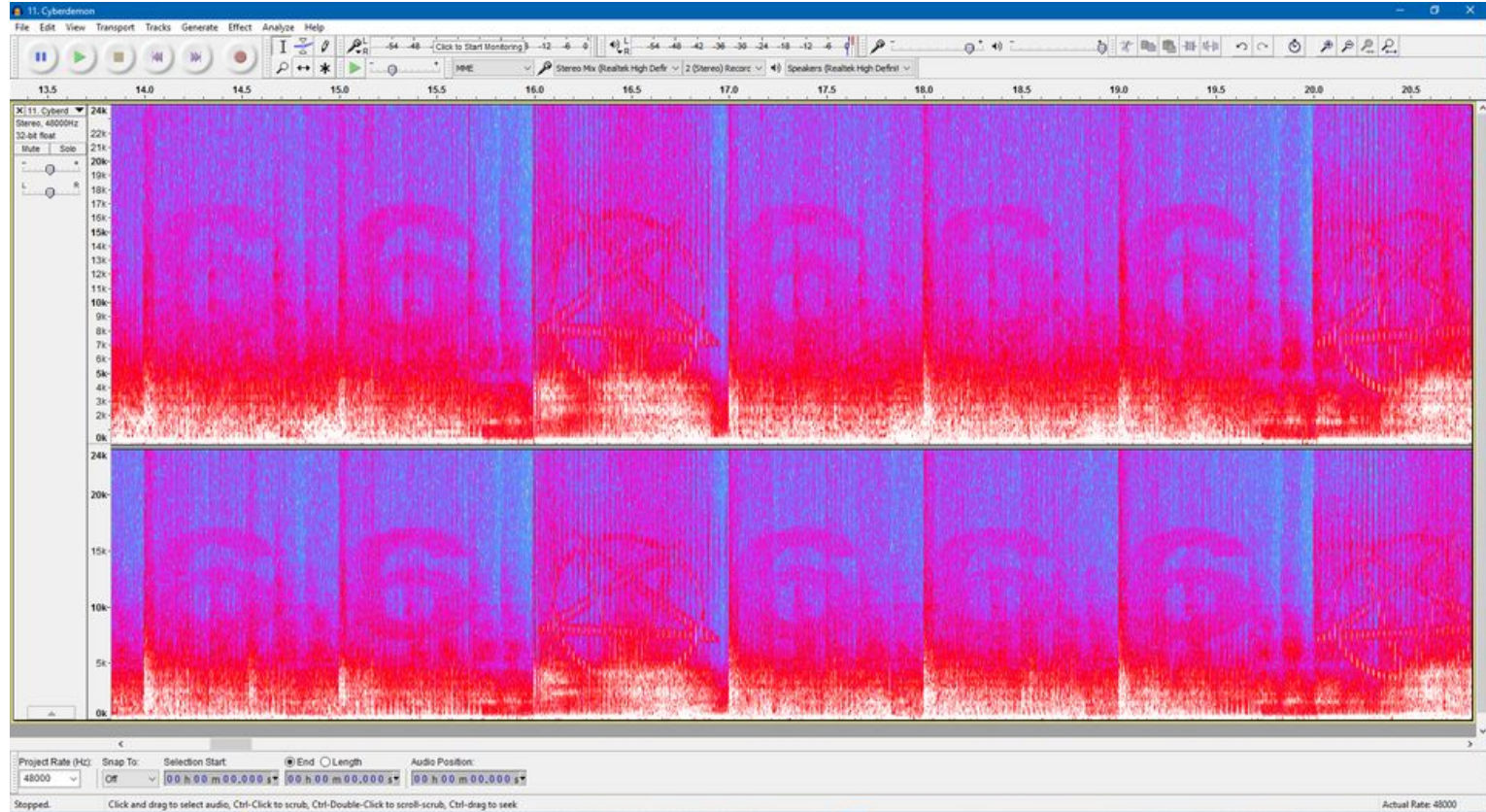source: Medium post

Spectrogram

Source: An Algorithm for Detection of Breath Sounds in Spontaneous Speech with Application to Speaker Recognition

Spectrograms as Easter eggs. DOOM 2016 soundtrack
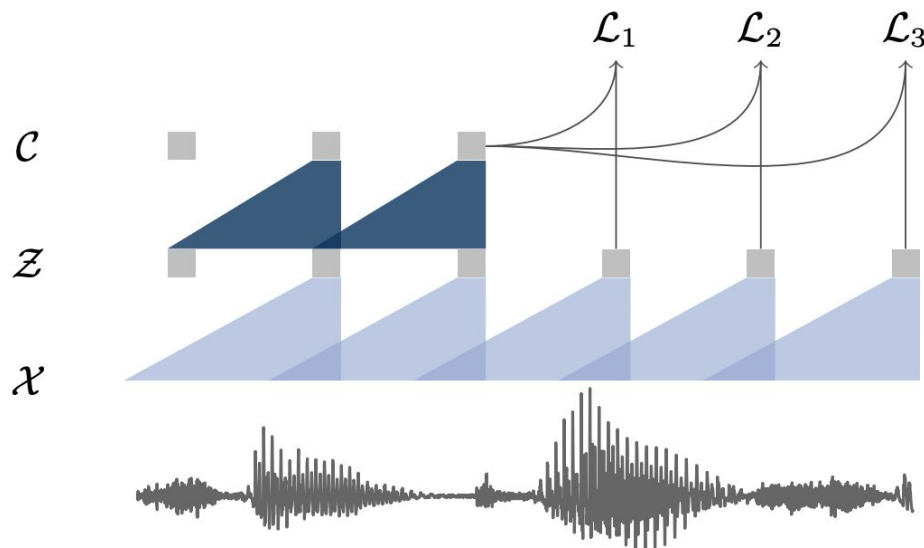
Source: reddit thread

Figure 1: Illustration of pre-training from audio data $\mathcal{X}$ which is encoded with two convolutional neural networks that are stacked on top of each other. The model is optimized to solve a next time step prediction task.

Source: [wav2vec: Unsupervised Pre-Training for Speech Recognition](#)

Fixed window to back-propagate will always have the same words contributing for the update with same weight of gradients flowing from last word to the first

Let's randomize windows selected at each step!

More about BPTT:

https://machinelearningmastery.com/gentle-introduction-backpropagation-time/