Lecture 10:
Open Domain Question Answering, Speech recognition and Distillation

MADE, Moscow
03.06.2020

Radoslav Neychev

Outline

1. Open domain Question Answering
    - DrQA
2. Speech recognition and generation: brief overview
    - ASR
3. Extra: Knowledge Distillation
4. NLP area and perspectives overview
5. Outro

Based on: Series of Medium posts
http://ai.stanford.edu/blog/answering-complex-questions/
Distilling the Knowledge in a Neural Network

Open Domain question answering

Open-domain QA

SQuAD, TREC, WebQuestions, WikiMovies

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

Document Retriever


Article Talk Read Edit View history

Warsaw

From Wikipedia, the free encyclopedia

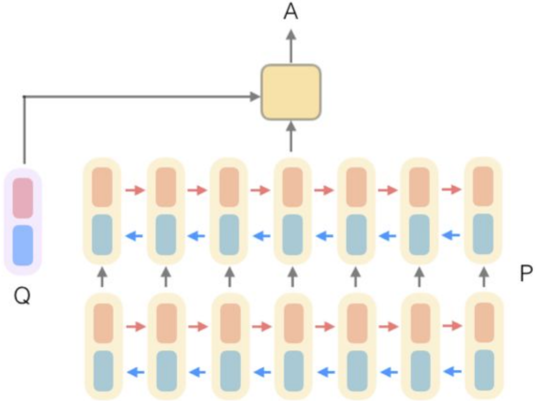
This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F 1934, see Adamowicz brothers.

Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

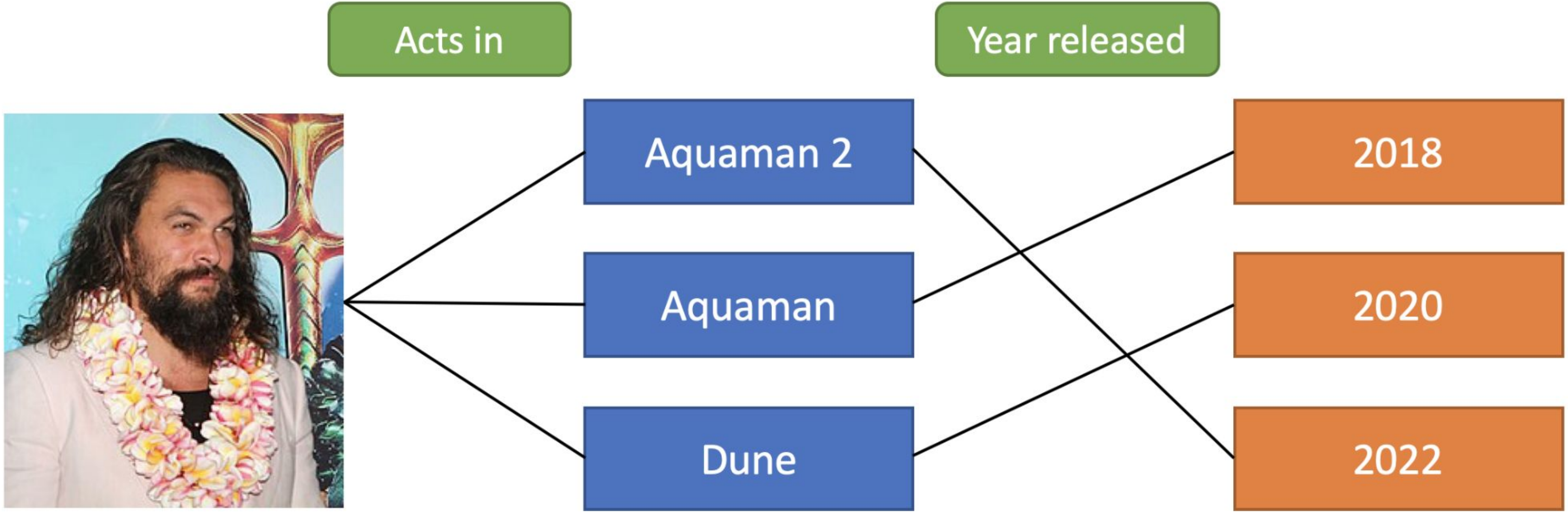
833,500



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

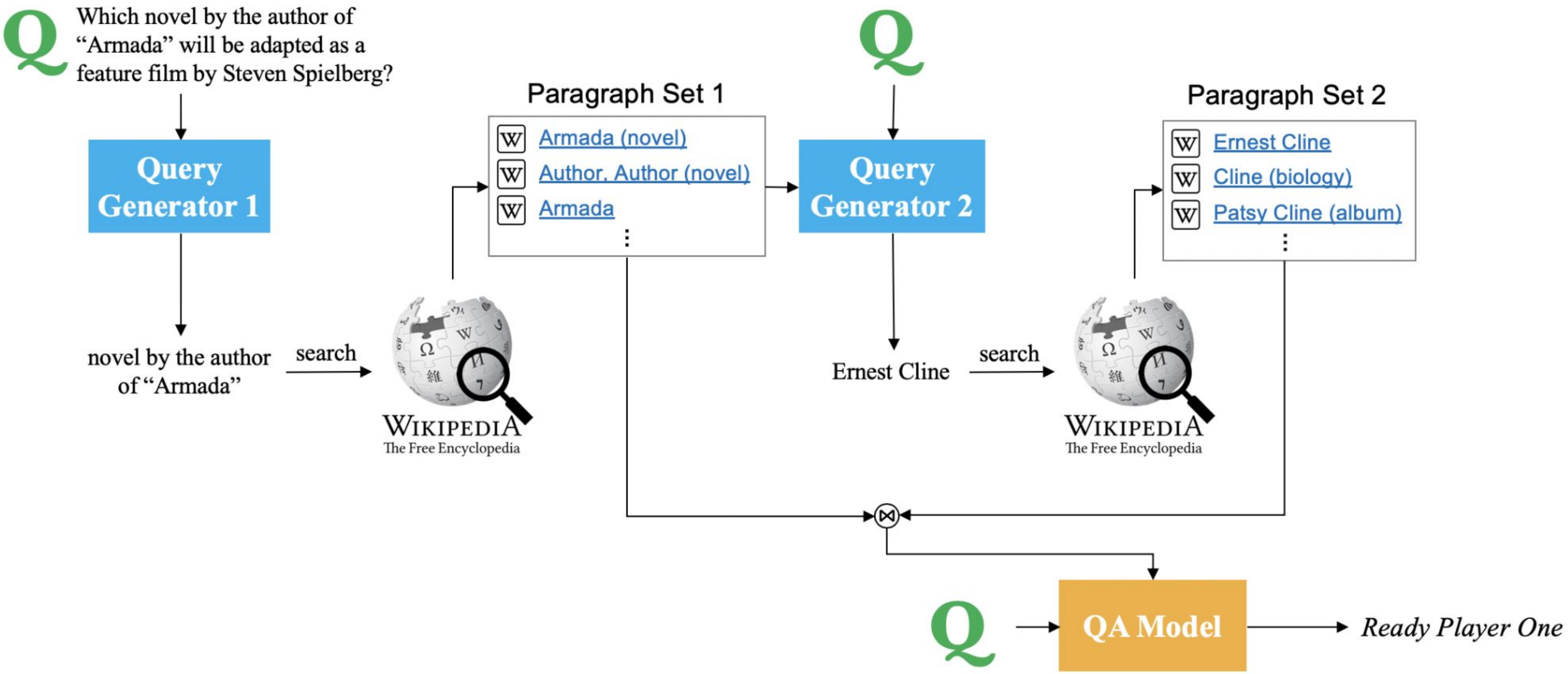
Possible problems

Example question: “What is the Aquaman actor’s next movie?”



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

Possible solutions

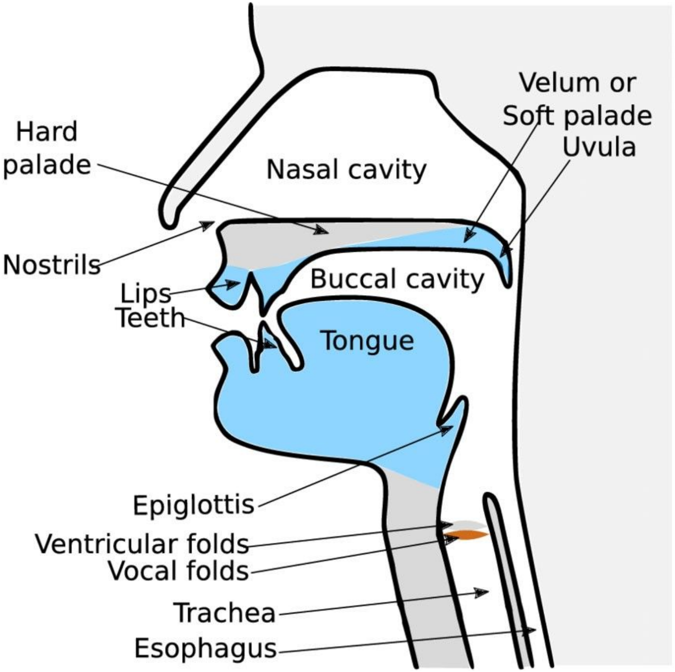


source: http://ai.stanford.edu/blog/answering-complex-questions/

Speech recognition and generation

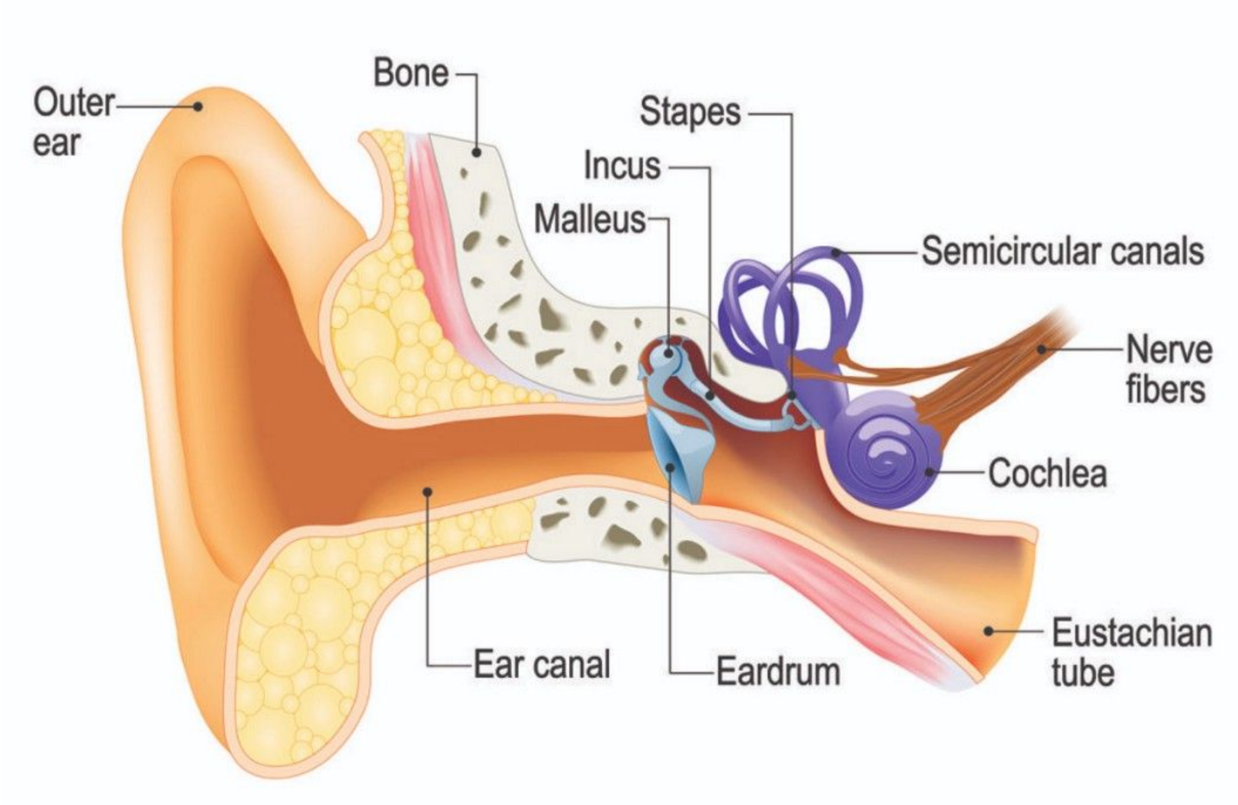
ASR and TTS

Foundations: vocal tract



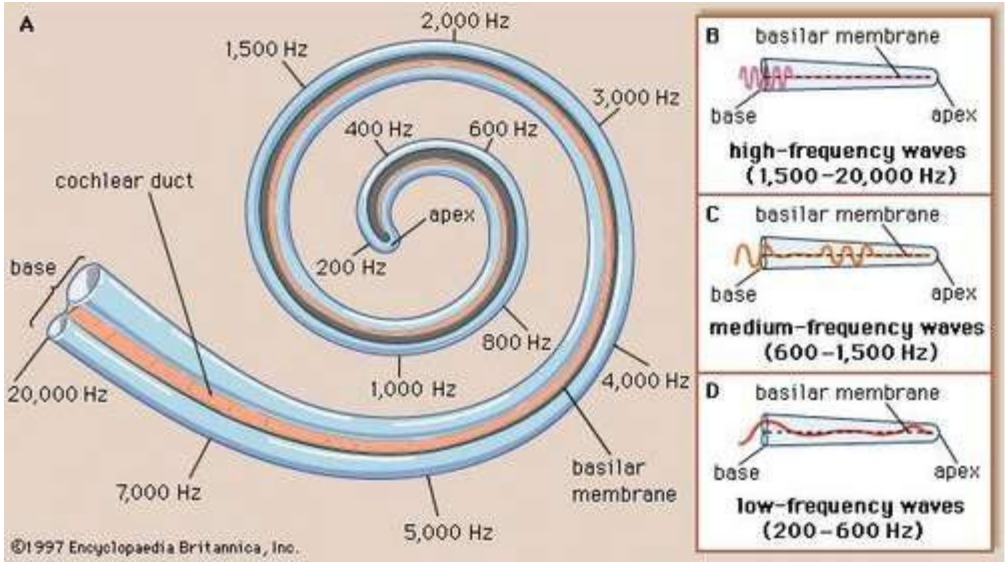
source: [Medium post]

Foundations: ear structure



source: [Medium post]

Foundations: ear structure



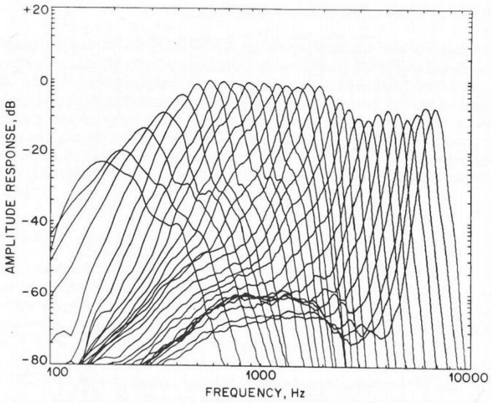


Figure 3.50 Frequency response curves of a cat's basilar membrane (after Ghitza [13]).

source: [Medium post]

Processing the sound

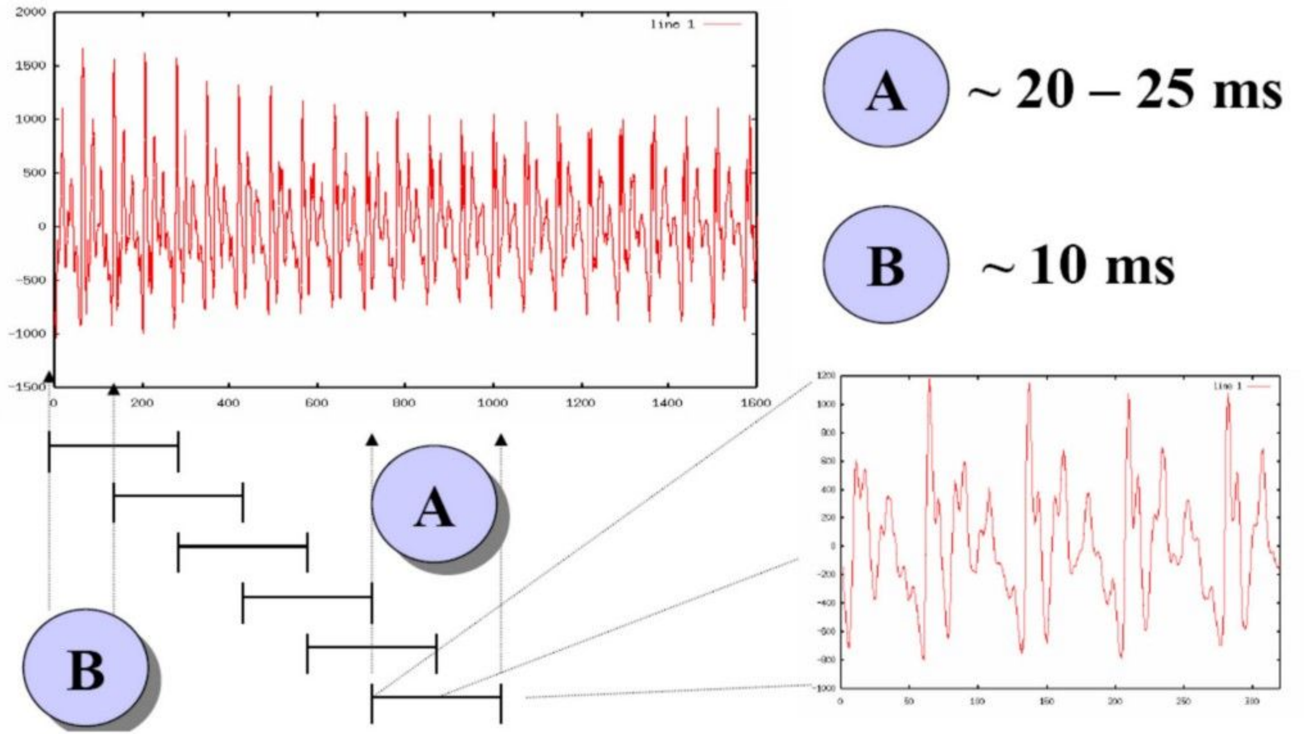


Image from Bryan Pellom

source: [Medium post]()

Fourier Transform

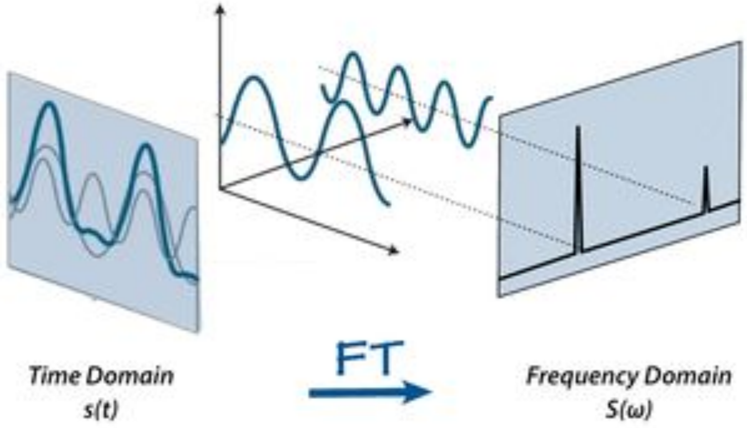
Definition [edit]

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

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

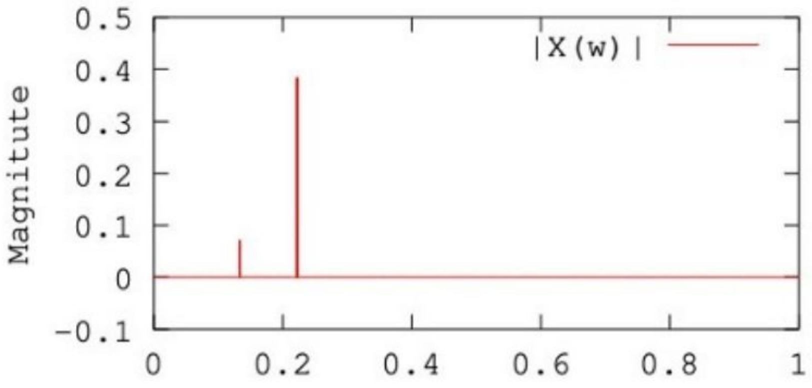
where the last expression follows from the first one by Euler's formula.

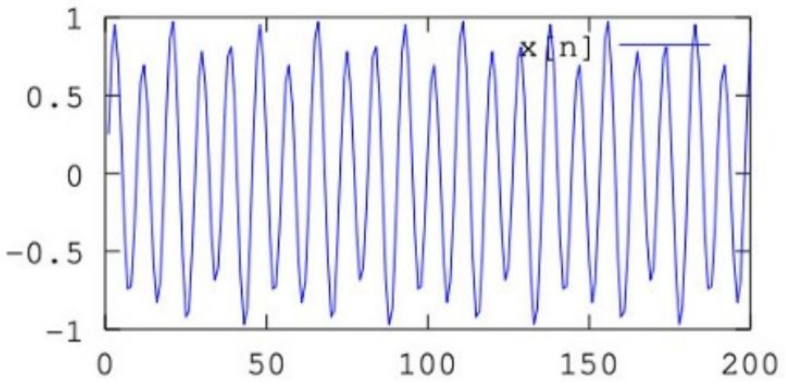
The transform is sometimes denoted by the symbol $\mathcal{F}$, as in $\mathbf{X} = \mathcal{F}\{\mathbf{x}\}$ or $\mathcal{F}(\mathbf{x})$ or $\mathcal{F}\mathbf{x}$.[A]

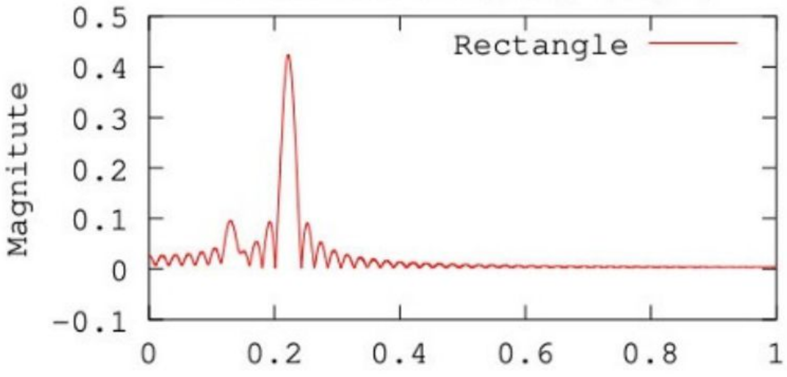


source: Wikipedia and Medium post

Fourier Transform: filtering

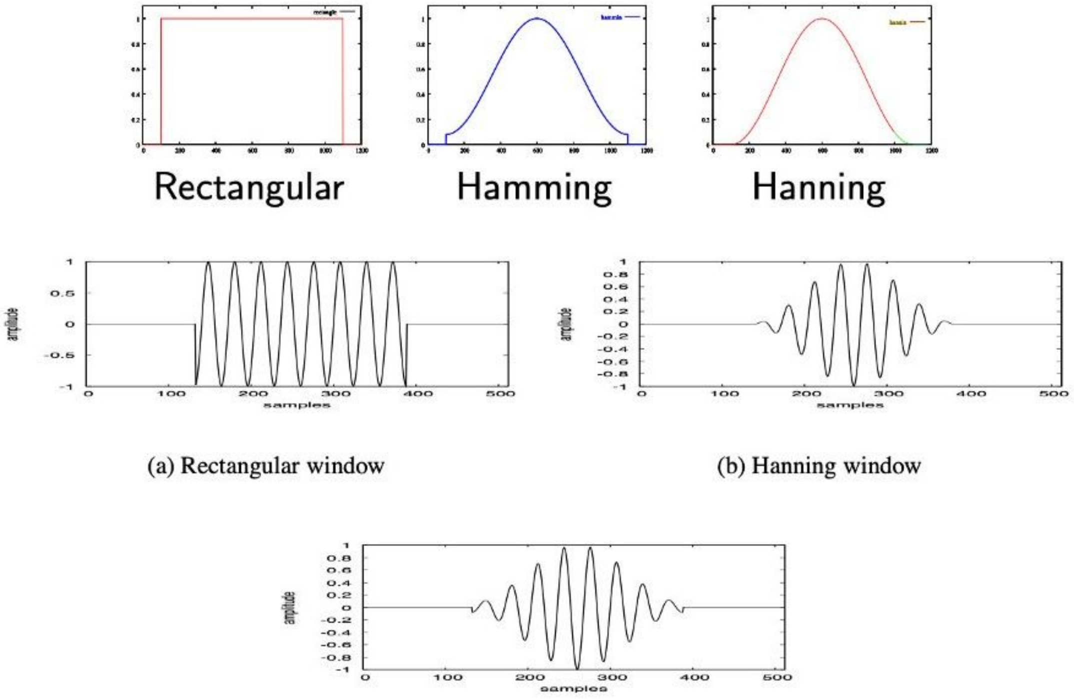






source: [Medium post]

Fourier Transform: filtering



(a) Rectangular window

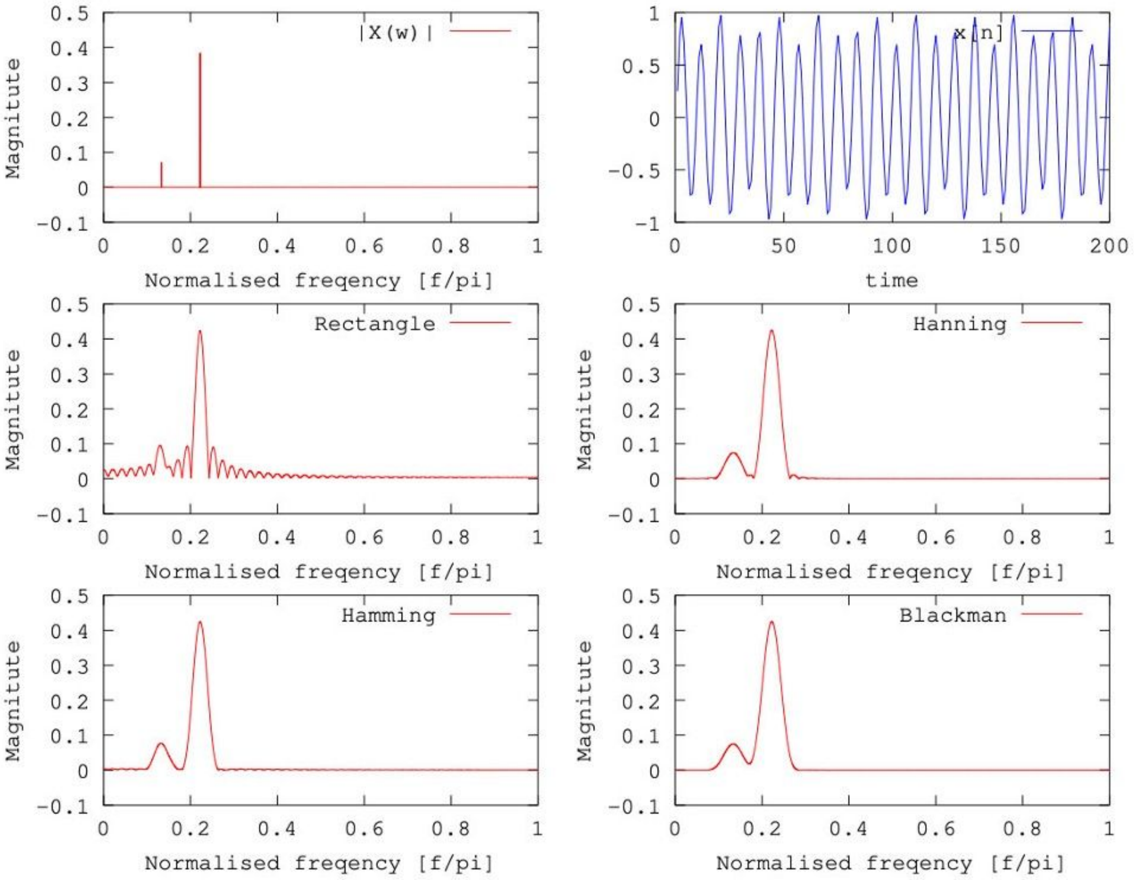
(b) Hanning window

(c) Hamming window

(Taylor, fig 12.1)

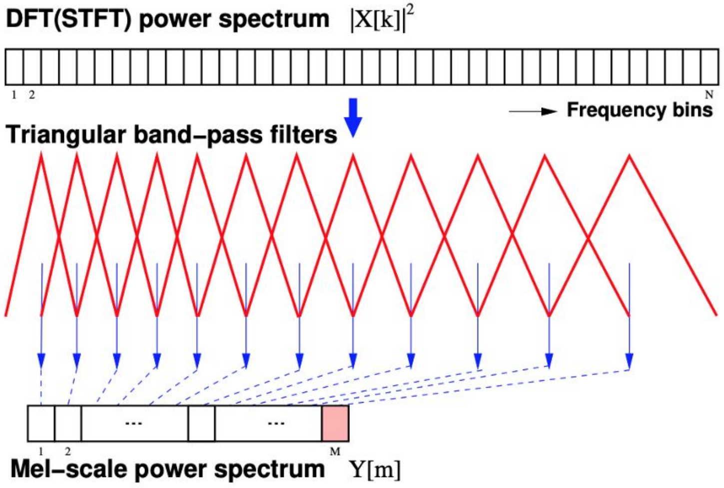
source: [Medium post]

Fourier Transform: filtering



source: [Medium post]

Features generation



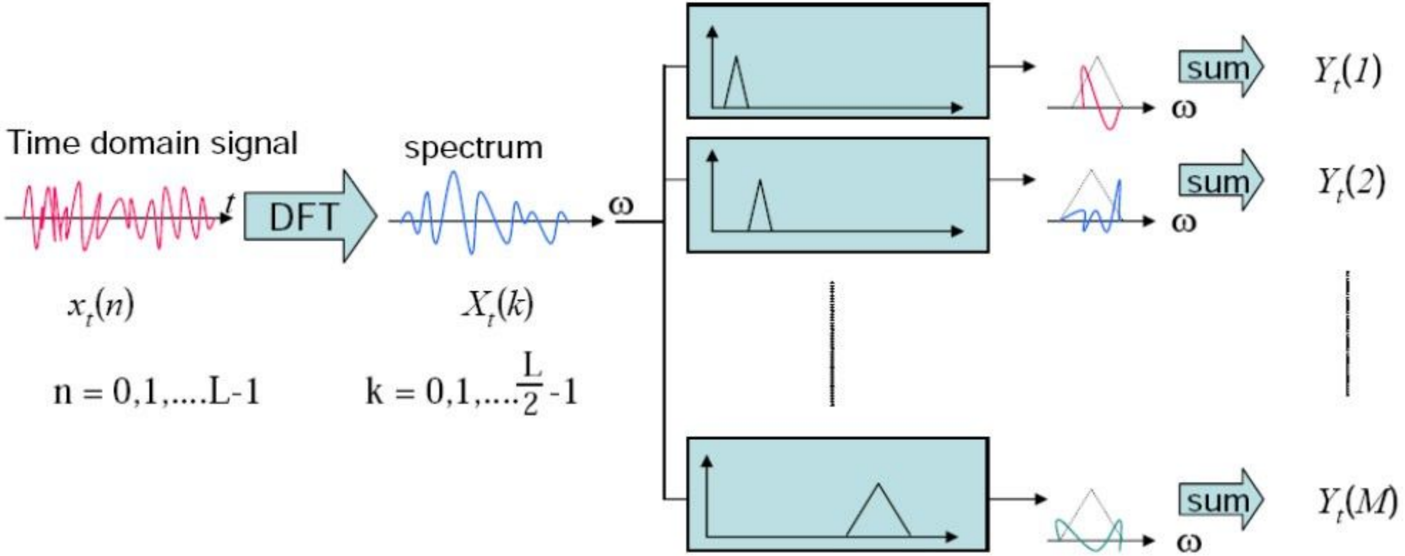
$$Y_t[m] = \sum_{k=1}^{N} W_m[k]\,|X_t[k]|^2$$

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

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

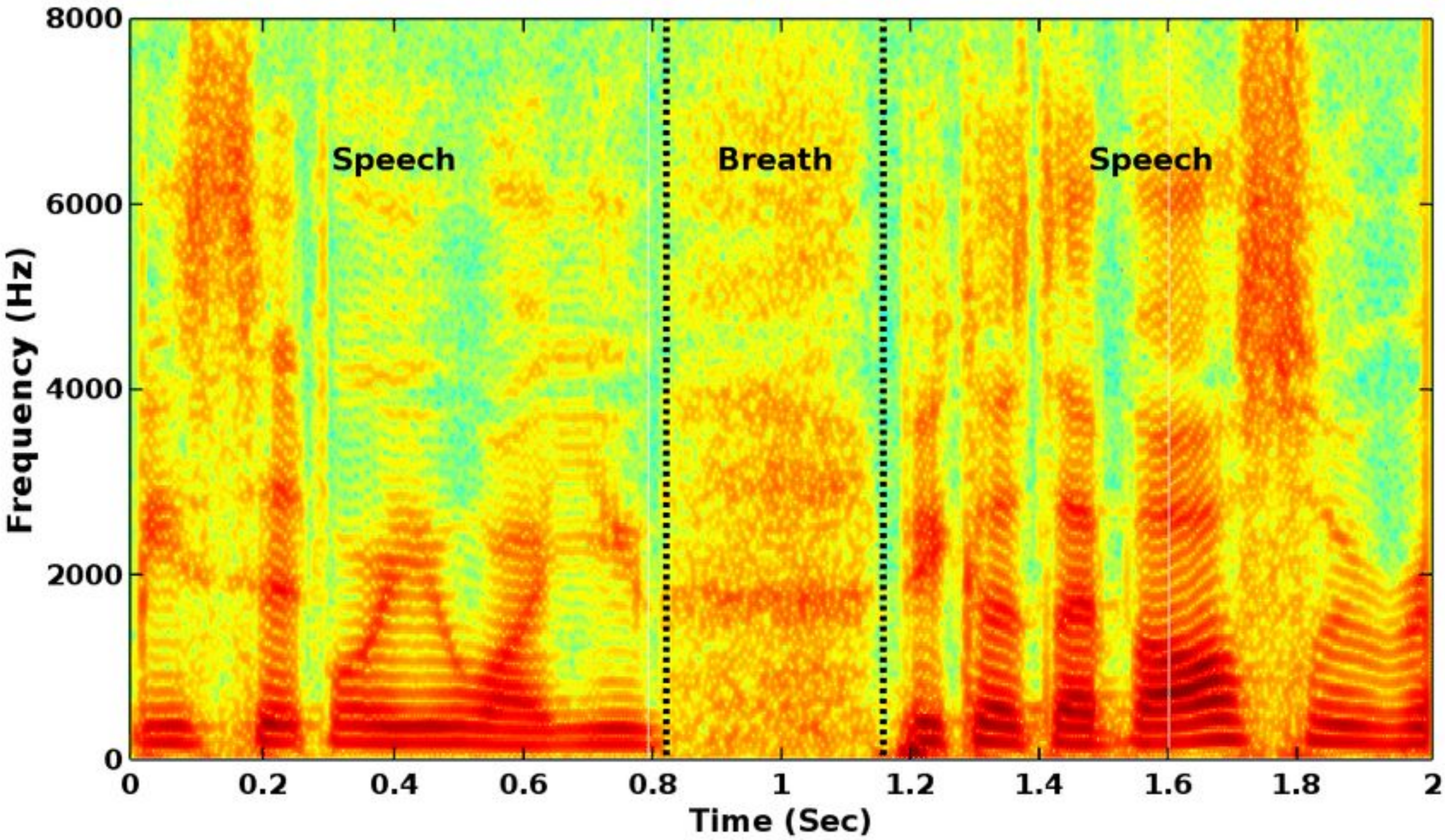
source: [Medium post]

Features generation



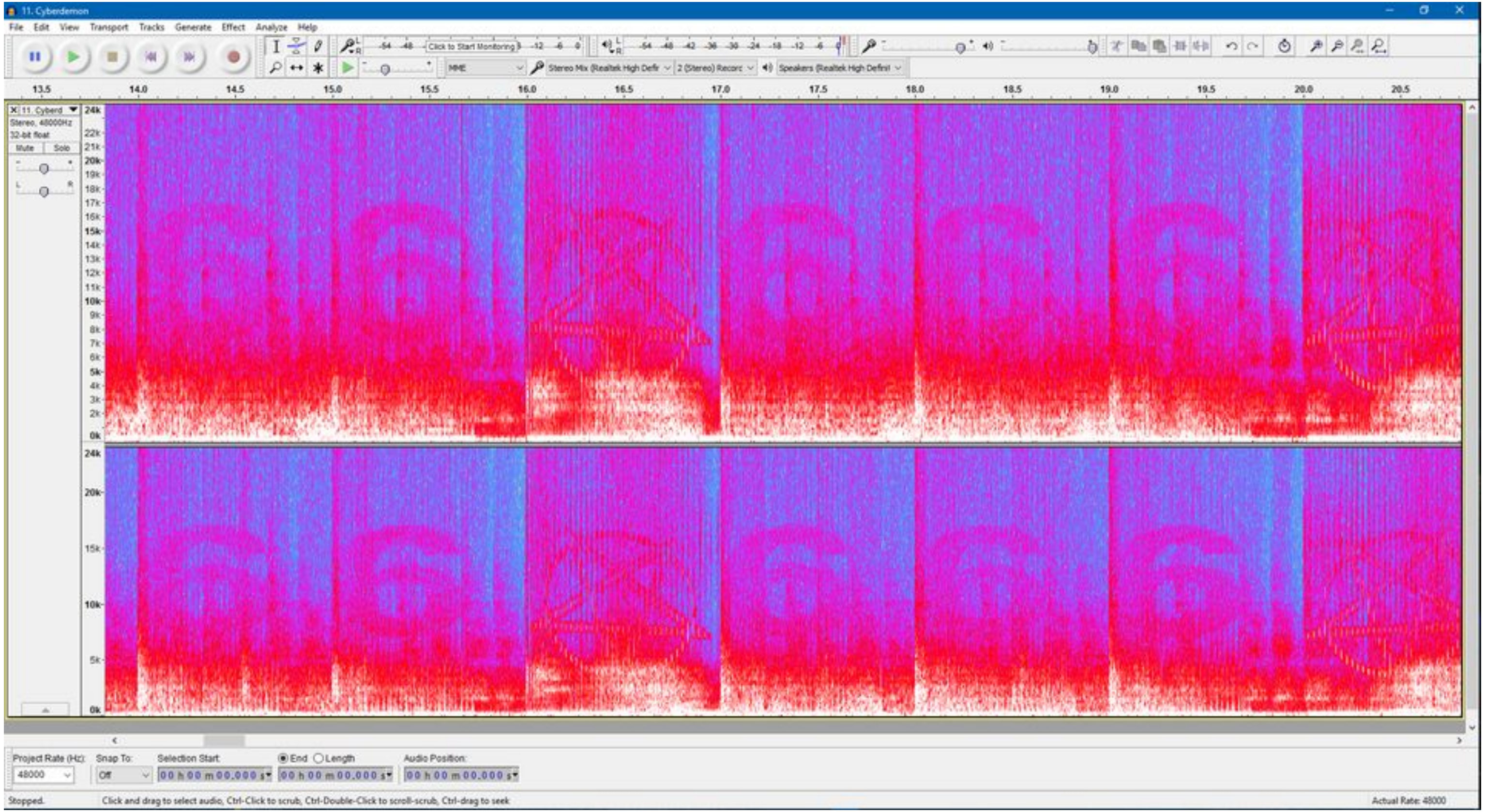
source: [Medium post]

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]

Embeddings

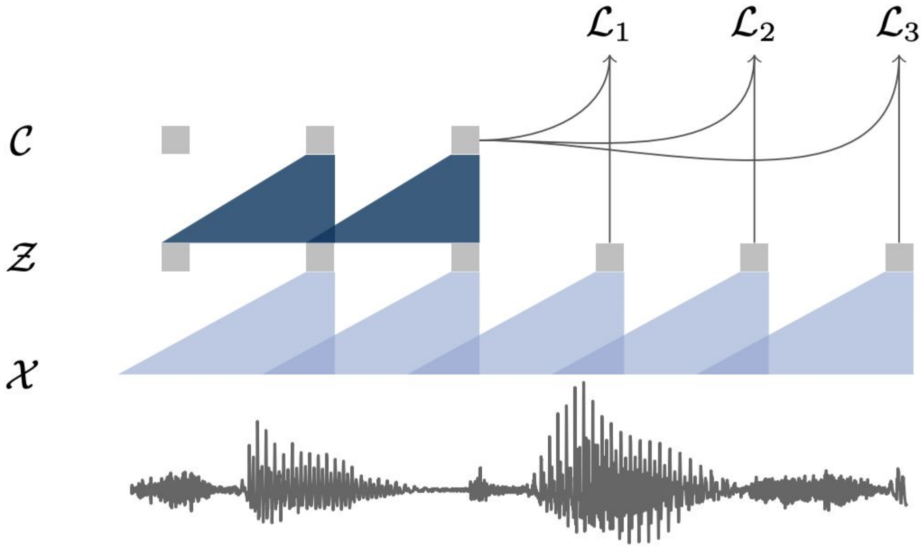


Figure 1: Illustration of pre-training from audio data $\mathcal{X}$ which is encoded with two convolutional neural networks that are stacked on top of each other. The model is optimized to solve a next time step prediction task.

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

Extra: Knowledge Distillation

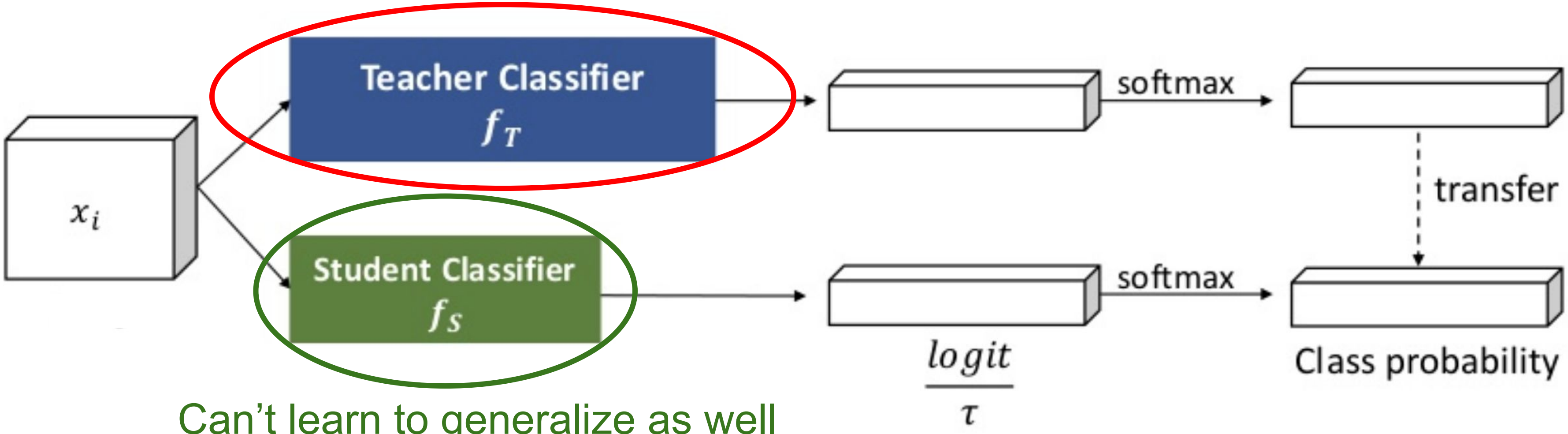
Cerura Vinula in caterpillar and butterfly forms

Do they have the same “life purpose”
and solve the same problems?

Image credits: commons.wikimedia.org (left), Ben Sale/Flickr.com (right), copied from
https://nat-geo.ru/nature/insects/bylo-stalo-kak-vyglyadyat-babochki-i-ikh-gusenitsy/

Knowledge distillation

Learns to generalize
and perform well



Can't learn to generalize as well
as the teacher,
but still can perform well

Source: https://www.slideshare.net/NaverEngineering/relational-knowledge-distillation

Knowledge distillation

Denote **teacher** and **student** models.

Student model has logits z_i and corresponding probabilities q_i , derived with the softmax operation:

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

where *T* stays for the temperature.

Teacher model has logits v_i and corresponding probabilities p_i .

Source: [Distilling the Knowledge in a Neural Network]

Knowledge distillation

Let's derive the cross-entropy gradient on **student** logits using the **teacher** predictions as targets:

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}\left(q_i - p_i\right) = \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right)$$

If the temperature is high, the following equation takes place:

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1 + z_i/T}{N + \sum_j z_j/T} - \frac{1 + v_i/T}{N + \sum_j v_j/T}\right)$$

Source: [Distilling the Knowledge in a Neural Network]

Knowledge distillation

Logits can be centered, so

$$\sum_j z_j = \sum_j v_j = 0$$

Then the gradient takes form:

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1 + z_i/T}{N + \sum_j z_j/T} - \frac{1 + v_i/T}{N + \sum_j v_j/T}\right) \approx \frac{1}{NT^2}(z_i - v_i)$$

Constant

$$\frac{dC}{dz_i} = \frac{1}{NT^2}(z_i - v_i) \sim (z_i - v_i) = M\frac{d(z_i - v_i)^2}{dz_i}$$

Source: [Distilling the Knowledge in a Neural Network]

Recent achievements: DistilBERT

	Nb of parameters (millions)	Inference Time (s)
GLUE BASELINE (ELMo + BiLSTMs)	180	895
BERT base	110	668
DistilBERT	66	410

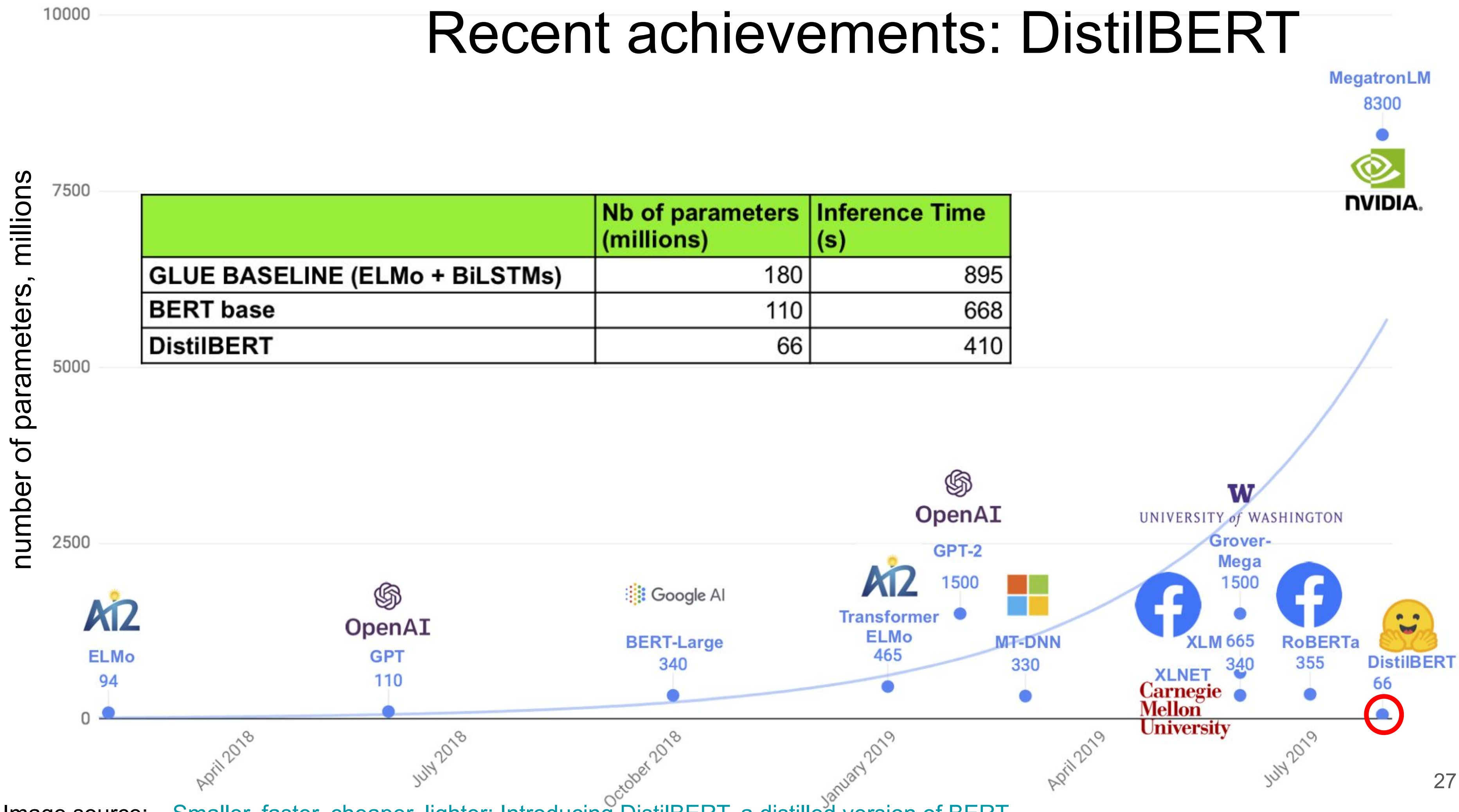


Image source: [Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT]

Main ideas

- DistilBERT is initialized from its teacher, BERT, by taking one layer out of two, leveraging the common hidden size.
  - *Comment: Training a sub-network is not only about the architecture. It is also about finding the right initialization for the sub-network to converge.*

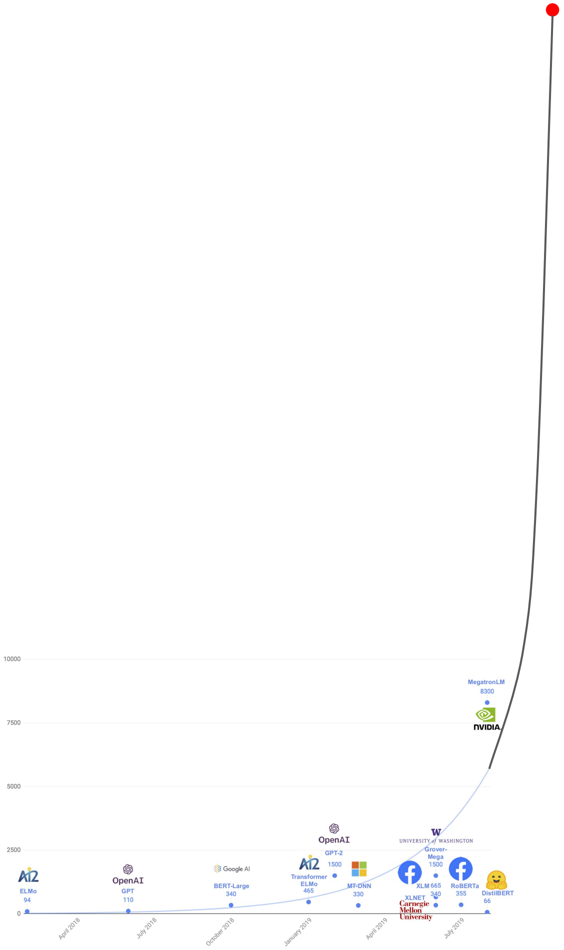
- DistilBERT is trained on very large batches leveraging gradient accumulation (up to 4000 examples per batch), with dynamic masking and removed the next sentence prediction objective.
  - *Comment: the way BERT is trained is crucial for its final performance.*

- DistilBERT was trained on eight 16GB V100 GPUs for approximately three and a half days using the concatenation of Toronto Book Corpus and English Wikipedia (same data as original BERT).

Image source: [Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT]

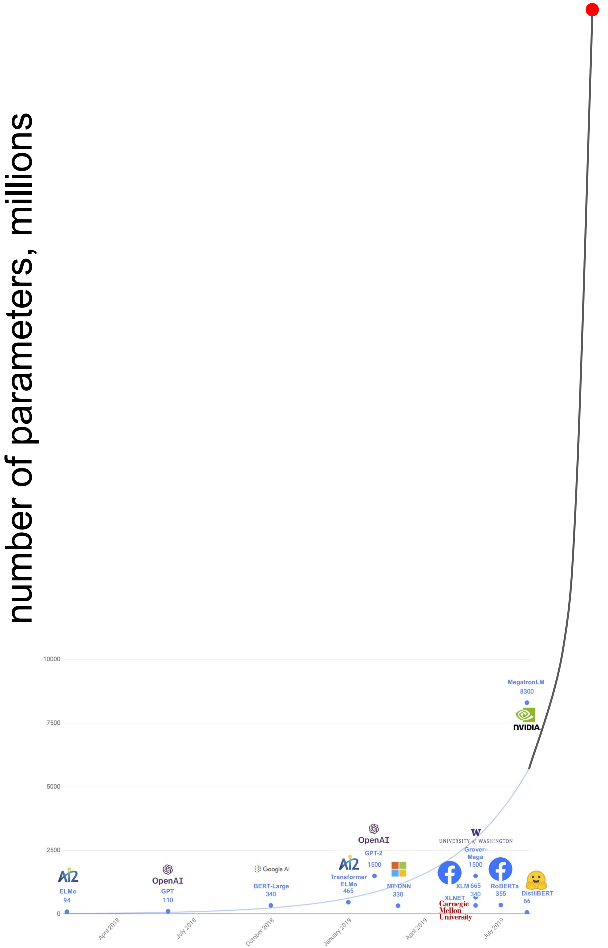
Recent achievements: GPT-3

number of parameters, millions

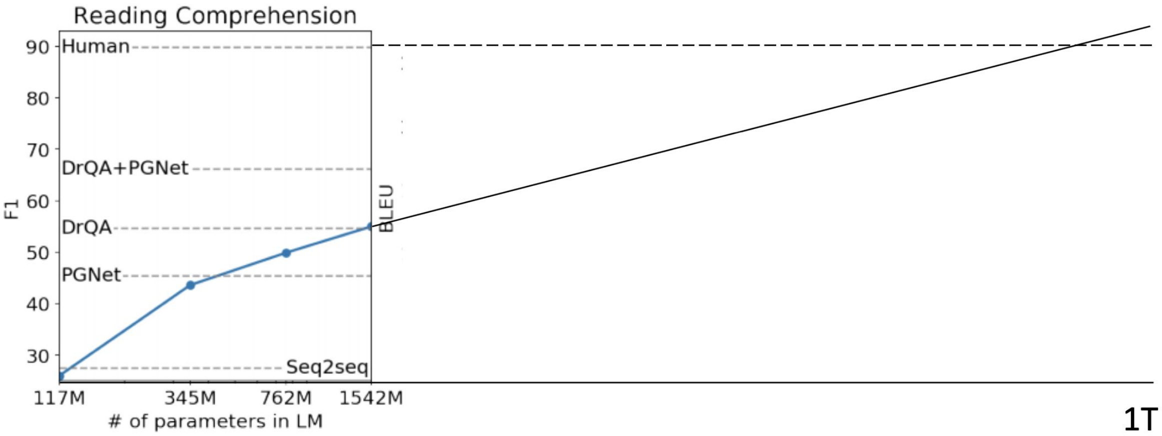


GPT-3, May 2020
175B parameters
(proportions are incorrect for visual sake)

Recent achievements: GPT-3



GPT-3, May 2020
175B parameters
(proportions are incorrect for visual sake)



Hypothesis from Stanford CS224n (2019) lecture 20

Q & A

Bonus: FusionNet

MLP (Additive) form:

$$S_{ij} = s^T \tanh(W_1 c_i + W_2 q_j)$$

Space: O(mnk), W is kxd

Bilinear (Product) form:

$$S_{ij} = c_i^T W q_j$$

$$S_{ij} = c_i^T U^T V q_j$$

$$S_{ij} = c_i^T W^T D W q_j$$

Space: O((m+n)k)

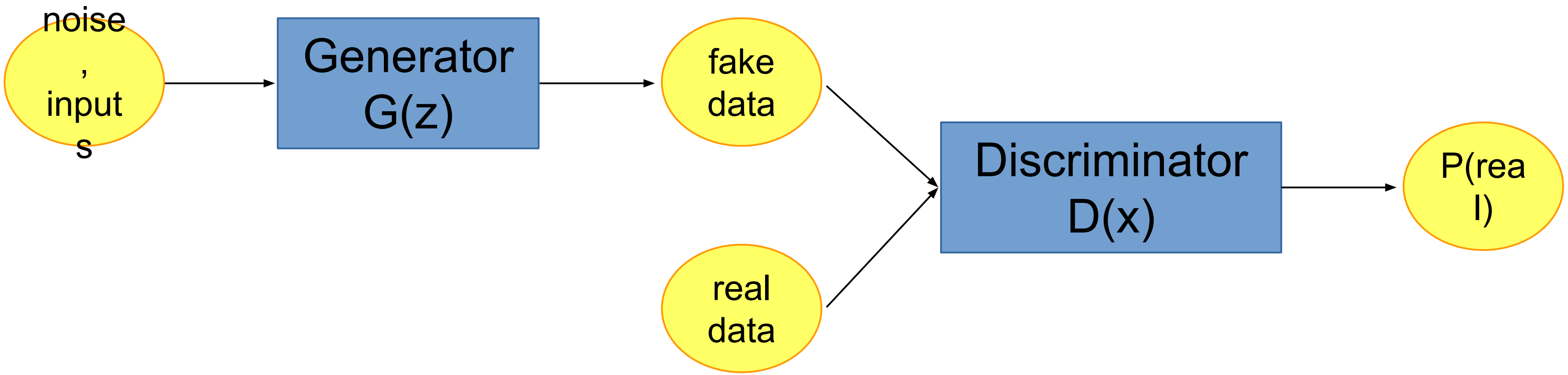
1. Smaller space
2. Non-linearity

$$\boxed{S_{ij} = Relu(c_i^T W^T) D Relu(W q_j)}$$

source: [CS224n Lecture 10]

Bonus: discrete GANs

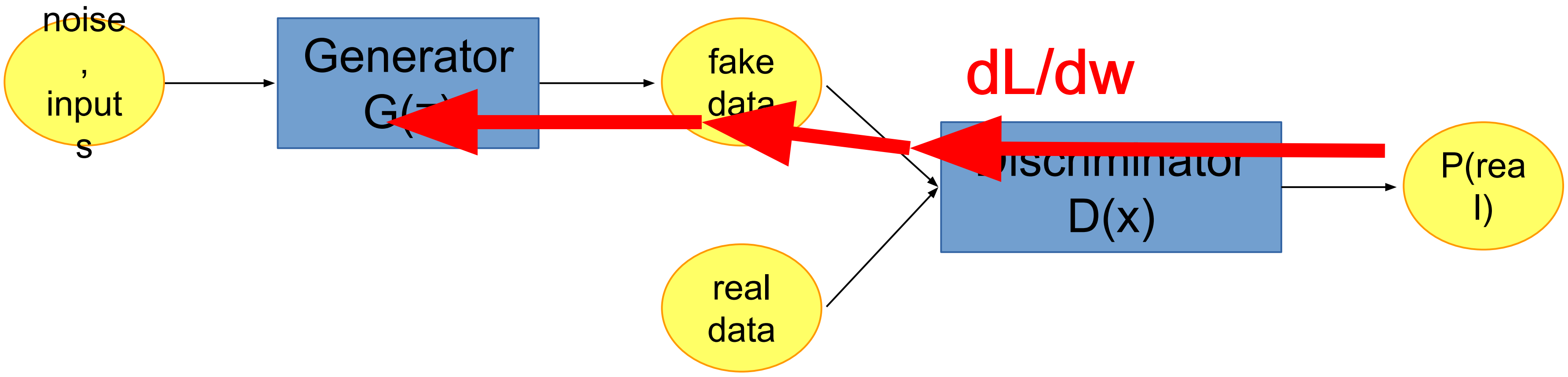
Generalized GAN scheme



source: [Practical RL week07]

Bonus: discrete GANs

Generalized GAN scheme

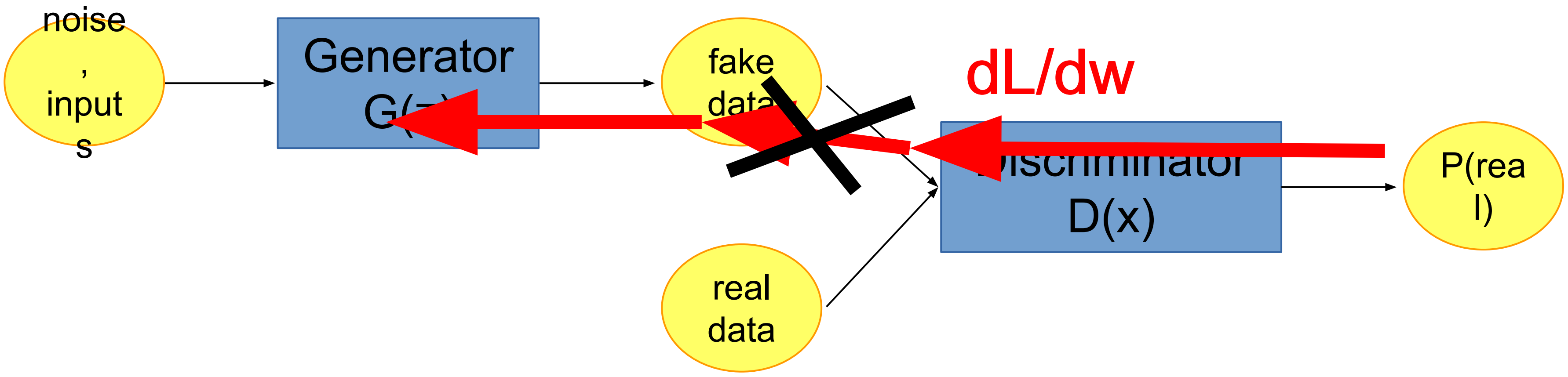


source: [Practical RL week07]

Bonus: discrete GANs

Standard scheme fails if G(z) is discrete

- generating text
- generatng music notes
- generating molecules
- binary image masks



source: [Practical RL week07]()

Bonus: discrete GANs

We can train generator with Reinforcement Learning methods!

$$\nabla J = \underset{\substack{z \sim p(z) \\ x \sim P(x|G_\theta(z))}}{E} \nabla \log P(x|G_\theta(z)) D(x)$$

source: [Practical RL week07]