



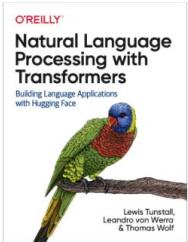
A guided tour through the Transformers landscape

Lewis Tunstall | Open Source @ Hugging Face | lewis@hf.co | @_lewtun

About me



[huggingface.co/
course/](https://huggingface.co/course/)



[NLP with
Transformers](#)

[Education](#)

Hugging Face
Solving NLP, one commit at a time!
NYC + Paris <https://huggingface.co/> Verified

[Repositories 203](#) [Packages](#) [People 80](#) [Teams 5](#) [Projects 4](#) [Sponsoring 4](#) [Settings](#)

Pinned repositories Customize pinned repositories

Repository	Description	Language	Stars	Forks	Issues
transformers	Transformers: State-of-the-art Natural Language Processing for Pytorch and TensorFlow 2.0.	Python	44.9k	10.7k	860
datasets	The largest hub of ready-to-use NLP datasets for ML models with fast, easy-to-use and efficient data manipulation tools	Python	7.2k	860	336
tokenizers	Fast State-of-the-Art Tokenizers optimized for Research and Production	Rust	4.5k	860	7
awesome-papers	Papers & presentation materials from Hugging Face's internal science day	Python	1.8k	104	14
accelerate	A simple way to train and use PyTorch models with multi-GPU, TPU, mixed-precision	Python	506	860	7
huggingface_hub	Client library to download and publish models and other files on the huggingface.co hub	Python	65	860	7

[Open Source](#)

TRANSFORMERS™ THE GAME

Why all the fuss?



⚡ Hosted inference API ⓘ

Question Answering

Examples

When did Marie win the Nobel Prize?

Compute

Context

Marie Skłodowska was born in Warsaw, Poland, to a family of teachers who believed strongly in education. She moved to Paris to continue her studies and there met Pierre Curie, who became both her husband and colleague in the field of radioactivity. The couple later shared the 1903 Nobel Prize in Physics.

Computation time on cpu: 0.179 s

1903

0.580

hf.co/deepset/roberta-base-squad2

Transformers are currently the best approach for analysing **text**

⚡ Hosted inference API ⓘ

Question Answering

Examples

When did Marie win the Nobel Prize?

Compute

Context

Marie Skłodowska was born in Warsaw, Poland, to a family of teachers who believed strongly in education. She moved to Paris to continue her studies and there met Pierre Curie, who became both her husband and colleague in the field of radioactivity. The couple later shared the 1903 Nobel Prize in Physics.

Computation time on cpu: 0.179 s

1903

0.580

hf.co/deepset/roberta-base-squad2

⚡ Hosted inference API ⓘ

Token Classification

Examples

Marie Skłodowska was born in Warsaw, Poland, to a family of teachers who believed strongly in education. She moved to Paris to continue her studies and there met Pierre Curie, who became both her husband and colleague in the field of radioactivity. The couple later shared the 1903 Nobel Prize in Physics.

Compute

Computation time on cpu: 0.242 s

Marie Skłodowska PER was born in Warsaw LOC , Poland LOC , to a family of teachers who believed strongly in education. She moved to Paris LOC to continue her studies and there met Pierre Curie PER , who became both her husband and colleague in the field of radioactivity. The couple later shared the 1903 Nobel Prize in Physics MISC .

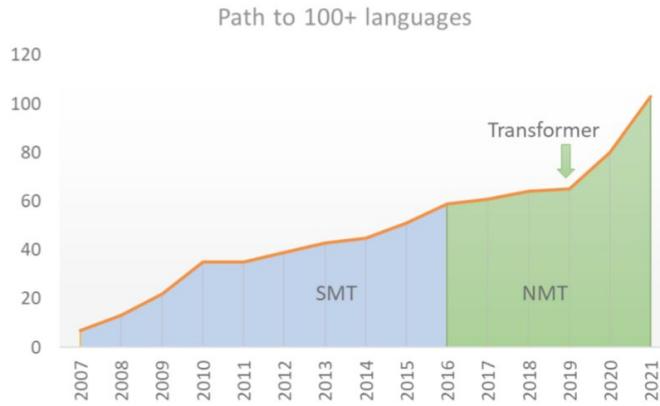
hf.co/xlm-roberta-large-finetuned-conll03-english

Transformers are currently the best approach for analysing **text**



Microsoft Translator now works across 103 languages

Microsoft adds 12 languages to its Microsoft Translate app that can help 84.6 million people.

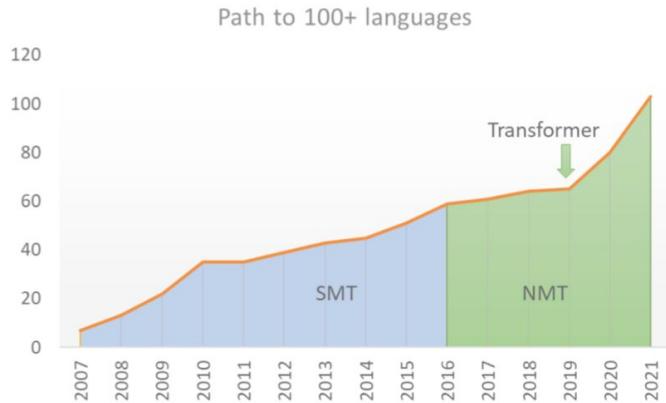


Transformers are currently the best approach for analysing **text**



Microsoft Translator now works across 103 languages

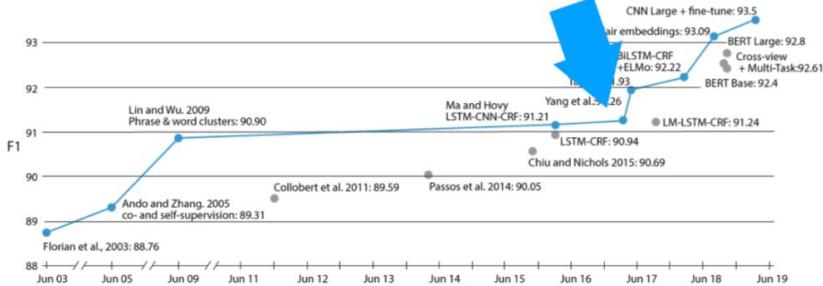
Microsoft adds 12 languages to its Microsoft Translate app that can help 84.6 million people.



Performance on Question Answering benchmark (SQuAD 2.0)



Performance on Named Entity Recognition benchmark (CoNLL)



Transformers are currently the best approach for analysing **text**

⚡ Hosted inference API ⓘ

Image Classification



Computation time on cpu: 0.199 s

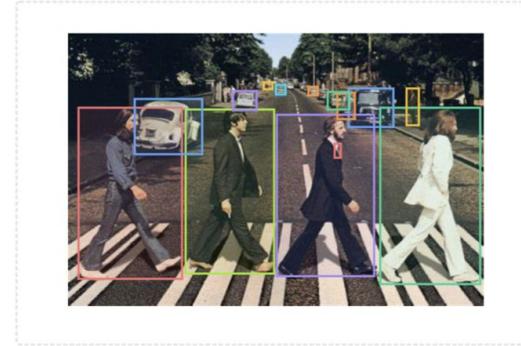
Eskimo dog, husky	0.403
Siberian husky	0.316
Norwegian elkhound, elkhound	0.057
dingo, warrigal, warragal, Canis dingo	0.052
malamute, malemute, Alaskan malamute	0.009

hf.co/microsoft/beit-base-patch16-224

⚡ Hosted inference API ⓘ

Object Detection

Examples

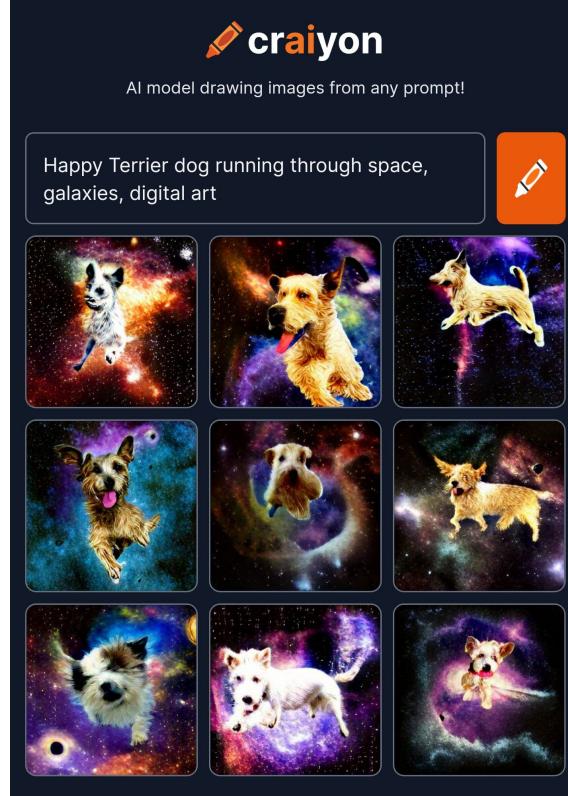


Computation time on cpu: 1.295 s

tie	0.955
person	0.999
person	0.971
car	0.998
car	0.904

hf.co/facebook/detr-resnet-50

Transformers are currently the
best approach for analysing **images**



Transformers are currently the best approach for analysing **text** and **images**

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

that is a portal to another dimension that looks like a monster as a planet in the universe

knitted out of wool spray-painted on a wall made out of plasticine

DALL-E 2

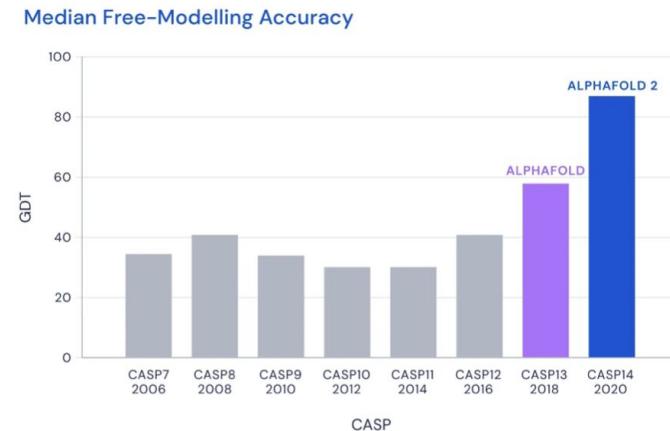
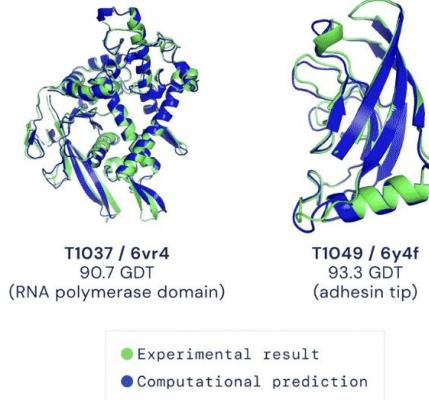


openai.com/dall-e-2/

Transformers are currently the best approach for analysing **text** and **images**

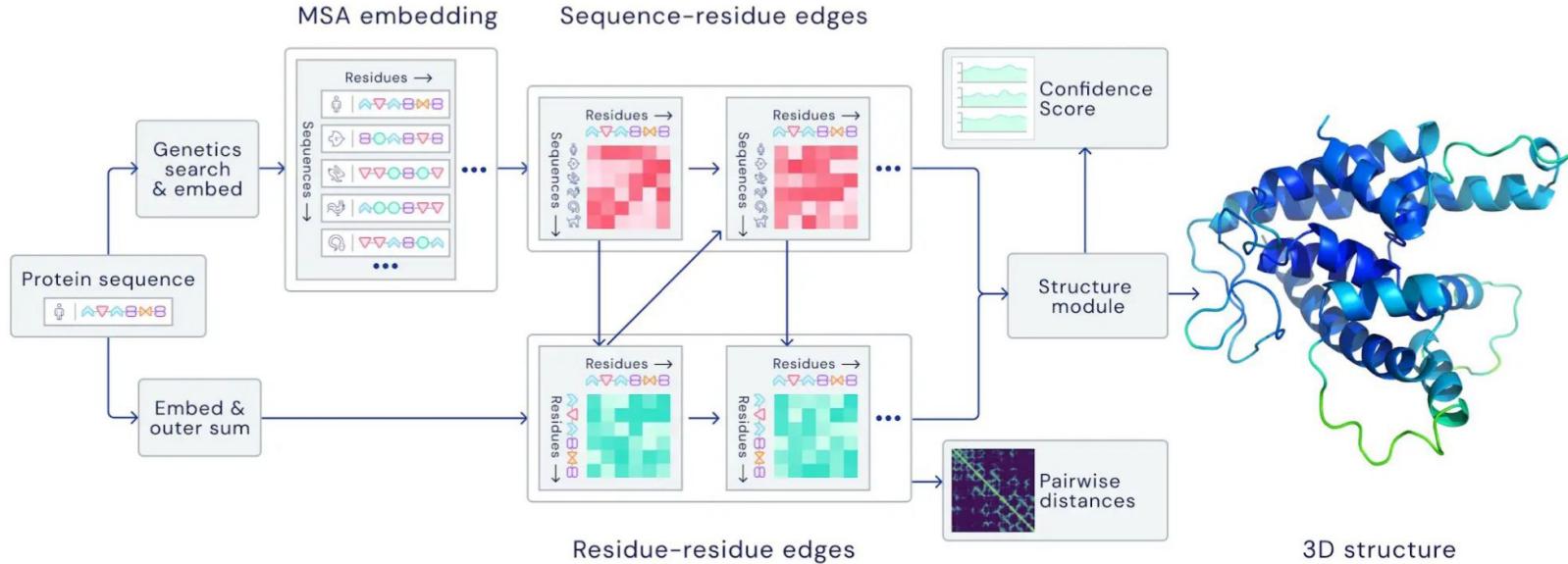
DeepMind's AI predicts structures for a vast trove of proteins

AlphaFold neural network produced a 'totally transformative' database of more than 350,000 structures from *Homo sapiens* and 20 model organisms.



<https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>

And it's not just text and images where
Transformers shine



<https://www.deeplearning.ai/deepmind/alpha-fold/>

And it's not just text and images where
Transformers shine

Main ingredients



Attention
mechanisms



Self-supervised learning
(Pretraining)



Transfer learning
(Fine-tuning)

Main ingredients



Attention
mechanisms

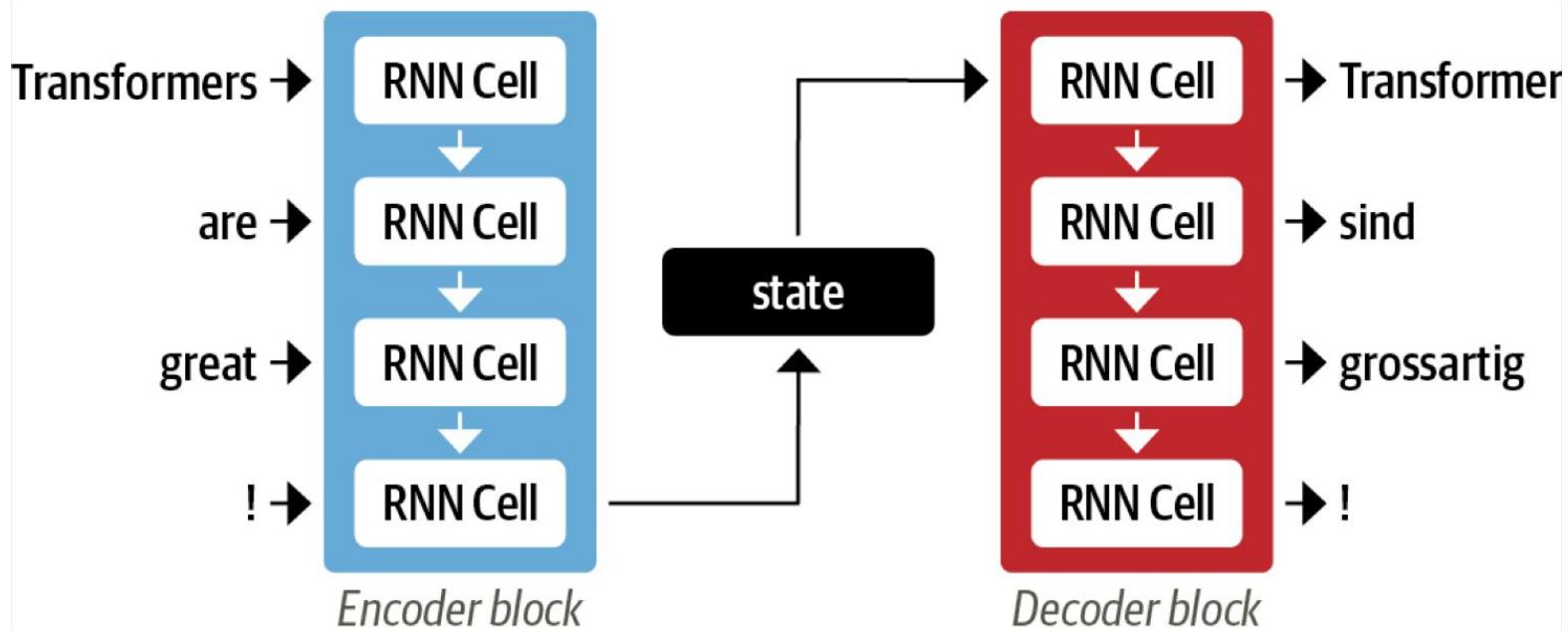


Self-supervised learning
(Pretraining)



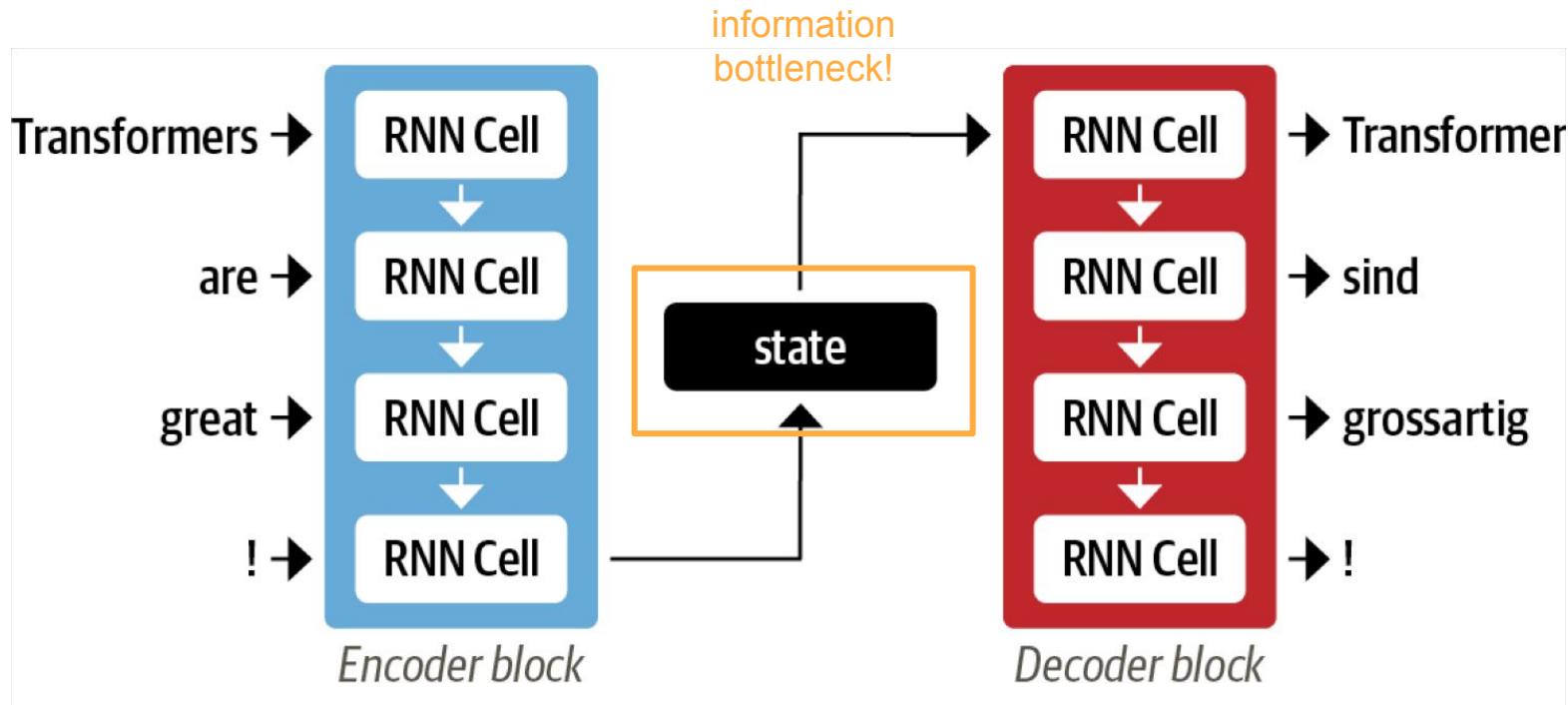
Transfer learning
(Fine-tuning)

Attention mechanisms



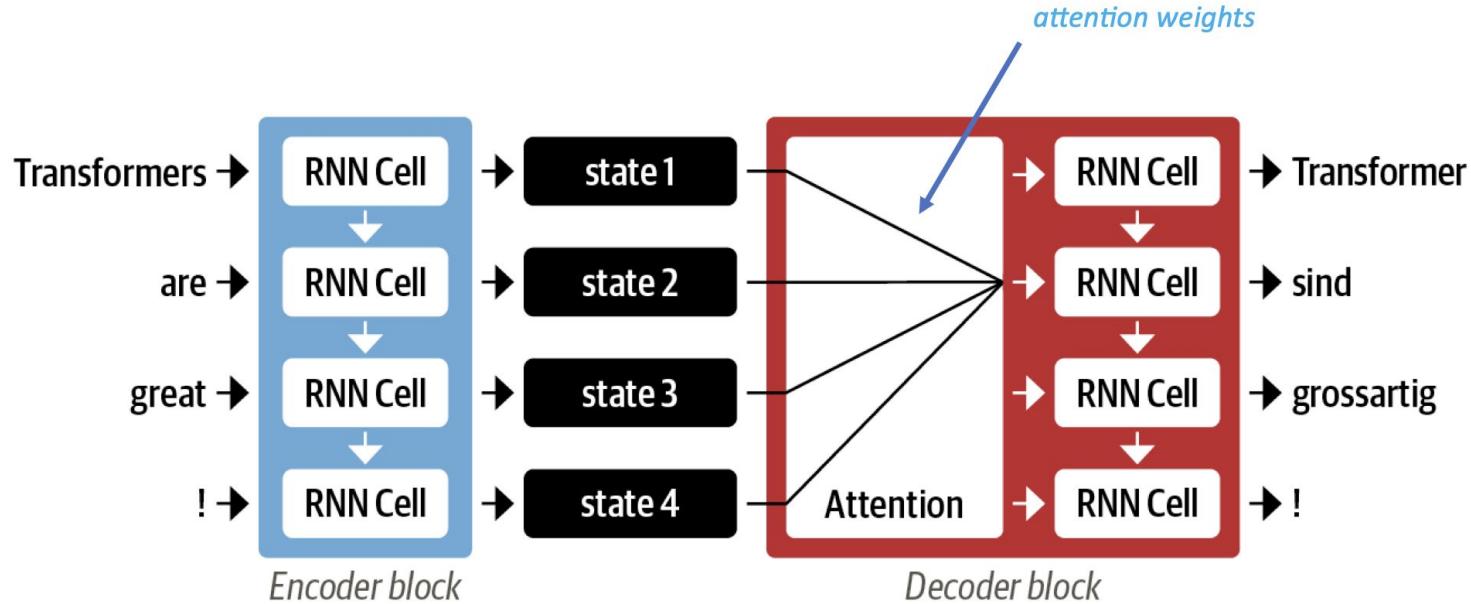
Originally developed for **recurrent neural networks**

Attention mechanisms



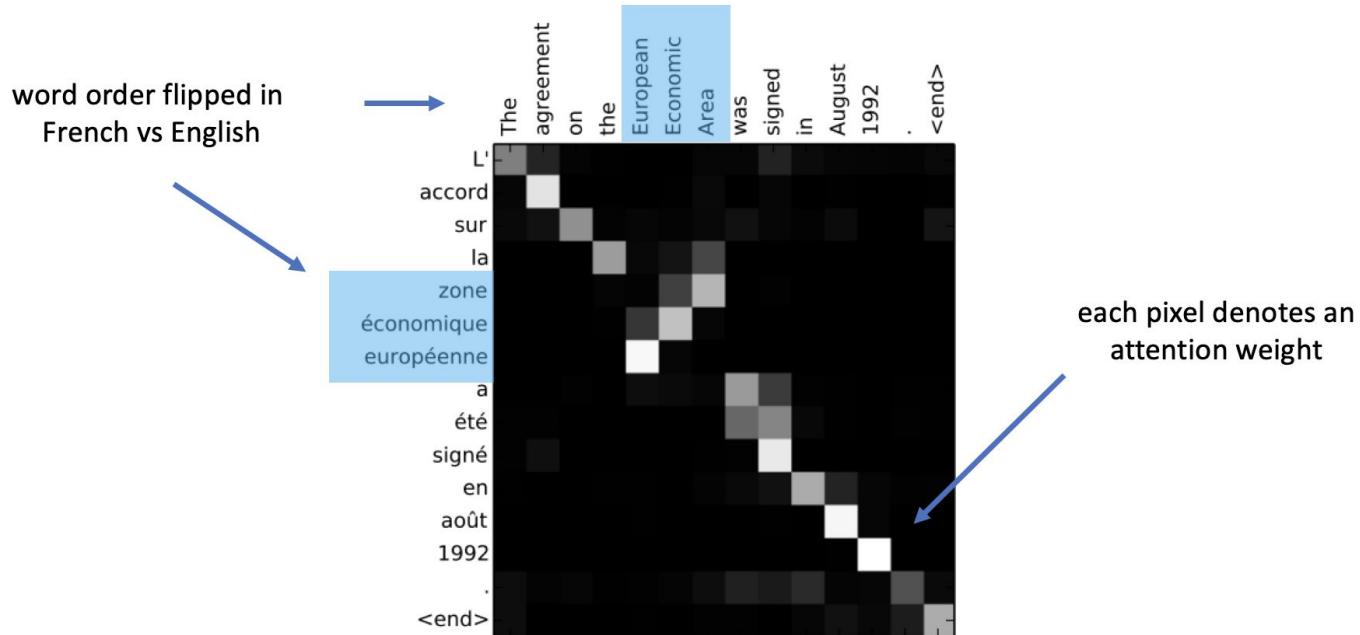
Originally developed for **recurrent neural networks**

Attention mechanisms



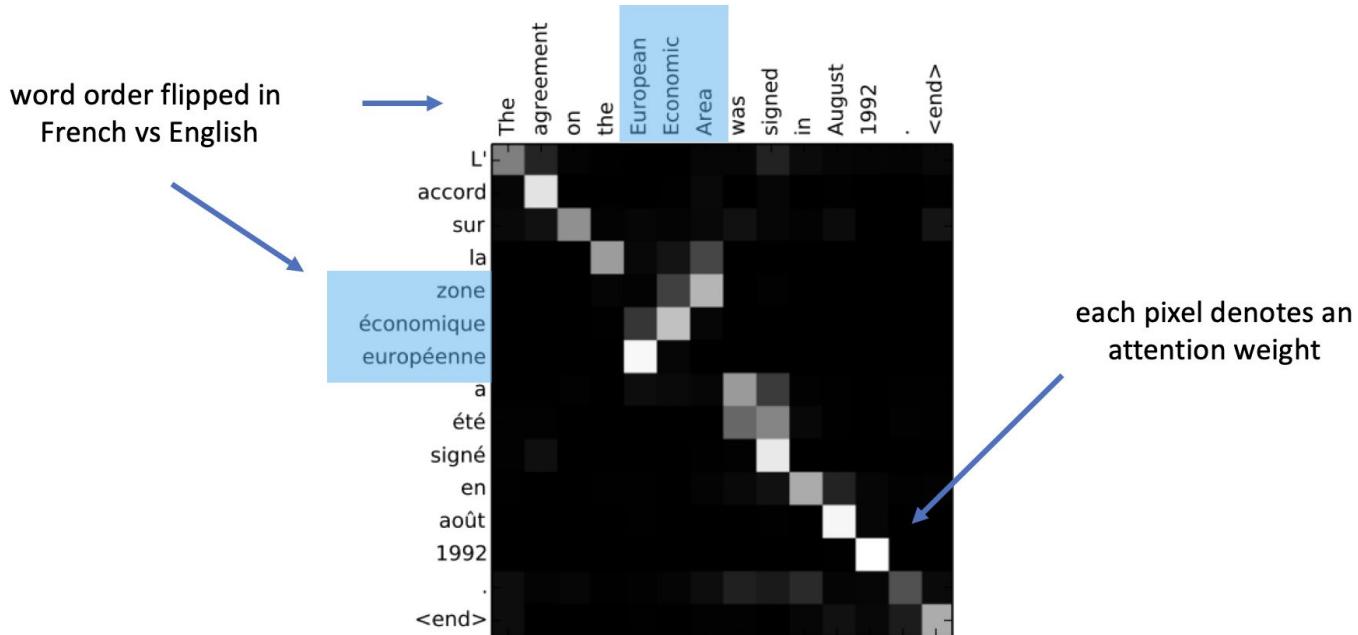
Use intermediate states but assign
a **weight** or “pay attention”

Attention mechanisms



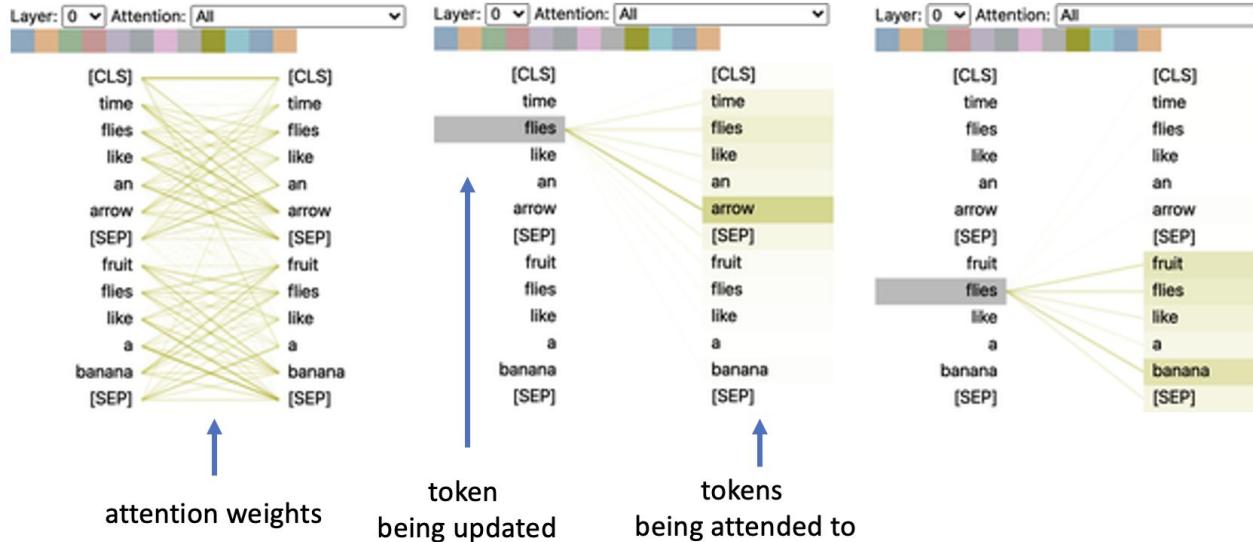
Attention gives a better modeling of **word order**

Attention mechanisms



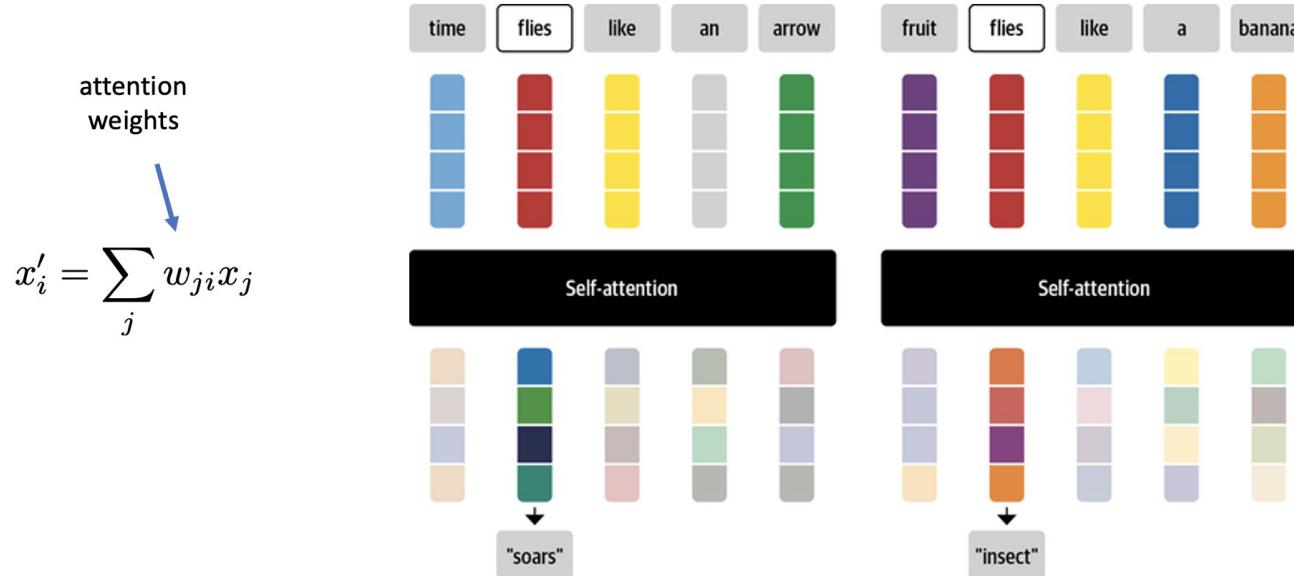
Attention gives a better modeling of **word order**

Attention mechanisms



Transformers use ***self-attention*** – every token interacts with every other token in the sequence

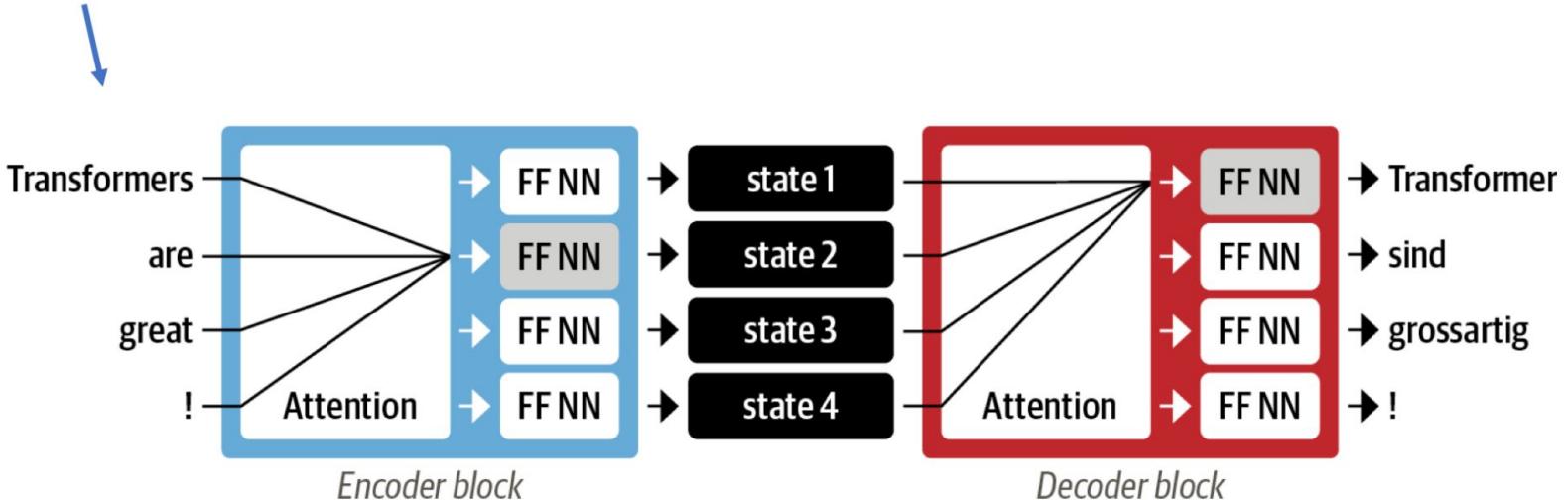
Attention mechanisms



Self-attention updates **input** embeddings x into
contextualised ones x'

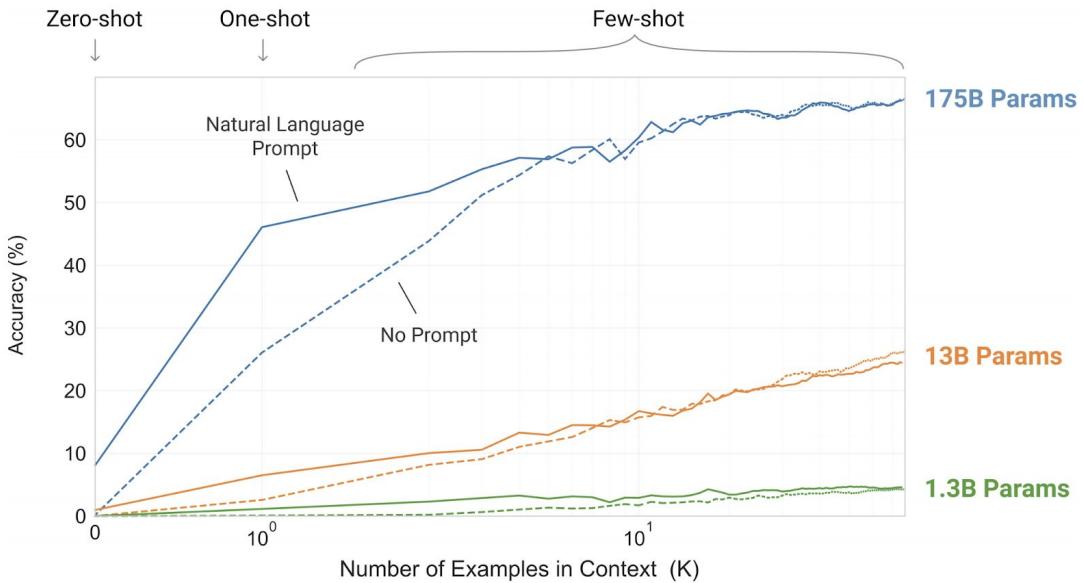
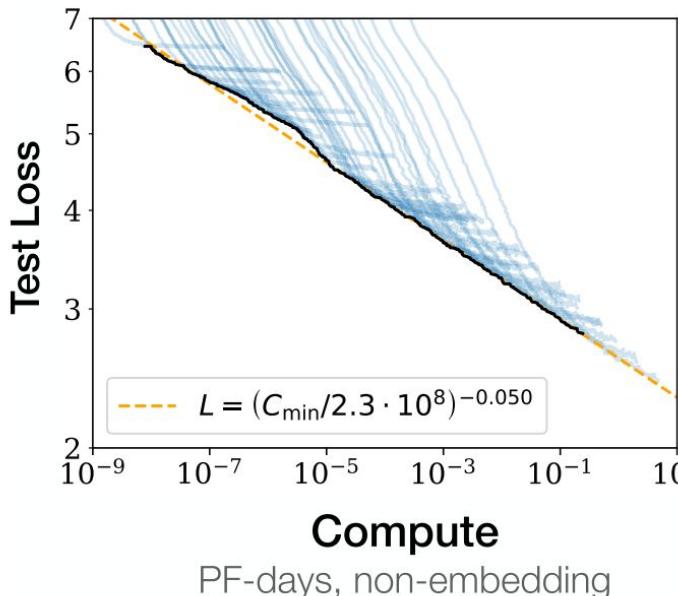
Transformer blocks

no recurrence!
feed sequence all at once

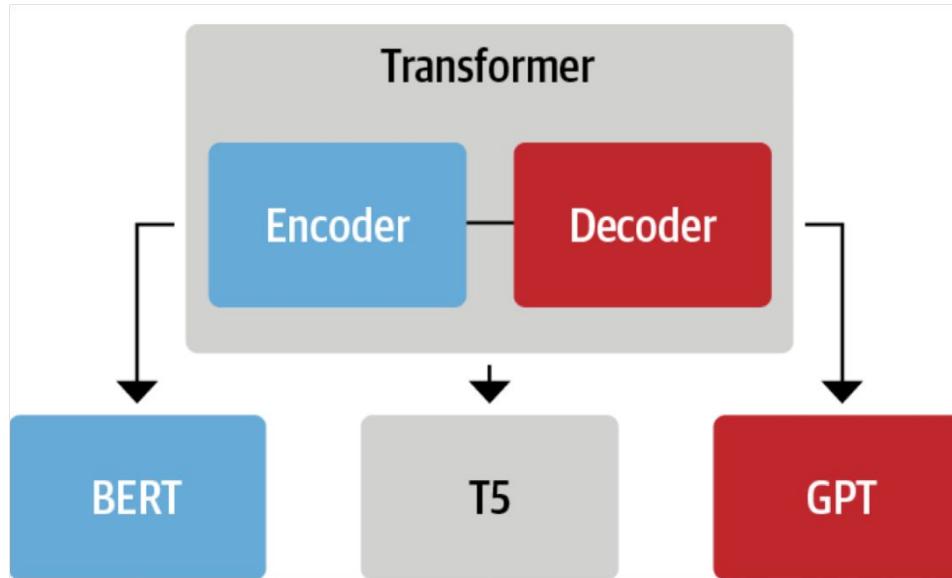


Transformers much easier to **scale** with compute & data

Scaling laws & emergent behaviour



Three types of architecture



Each architecture excels at different types of tasks

Main ingredients



Attention
mechanisms

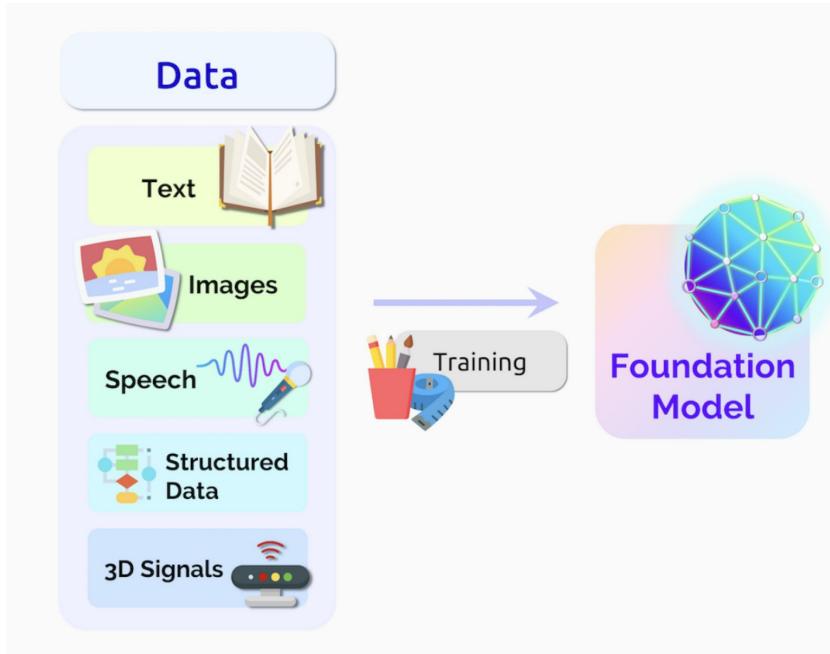


Self-supervised learning
(Pretraining)



Transfer learning
(Fine-tuning)

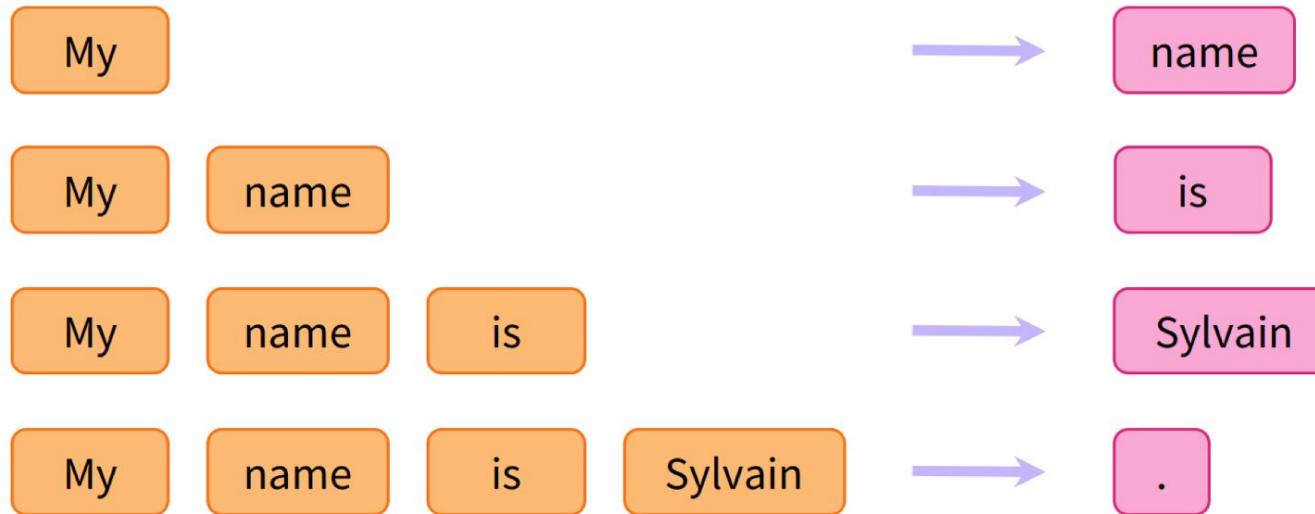
Transformer pretraining



You need:

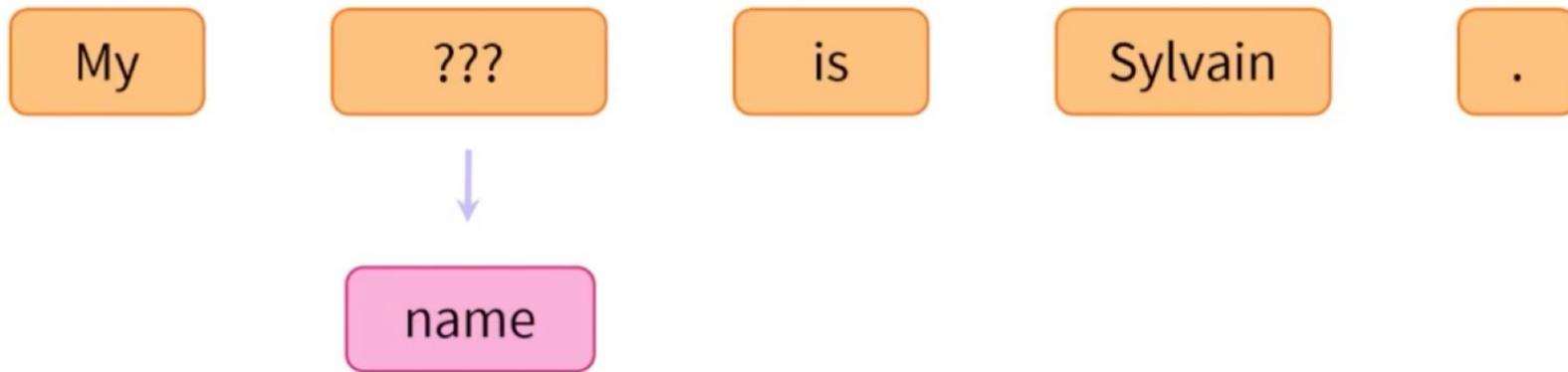
- A lot of **generic data** (internet, books...)
- A lot of **compute power** (datacenter, cloud providers...)

Transformer pretraining



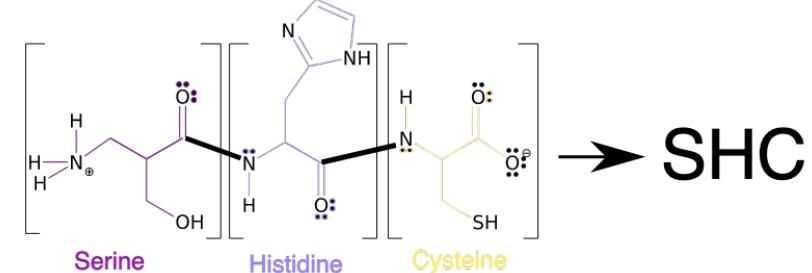
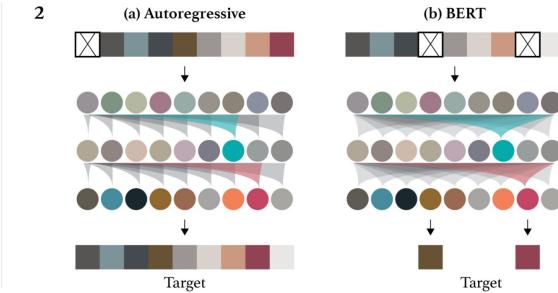
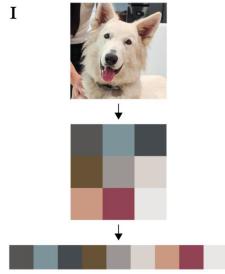
Trained on unlabeled data to predict the next token
(GPT-like) ...

Transformer pretraining

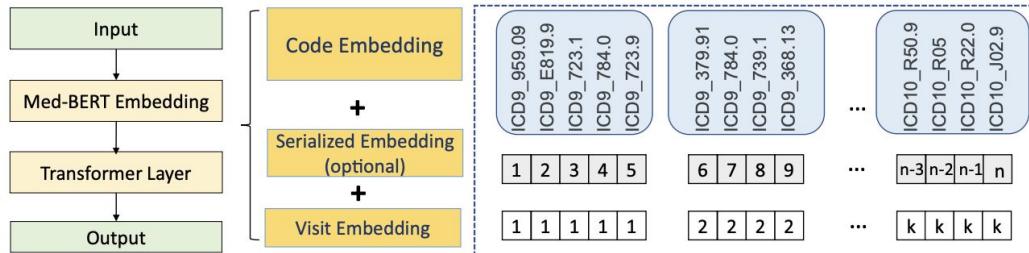


... or to predict the masked token (BERT-like)

Transformer pretraining



→ SHC



Sequences can be also be
pixels, proteins, patient codes etc

Main ingredients



Attention
mechanisms

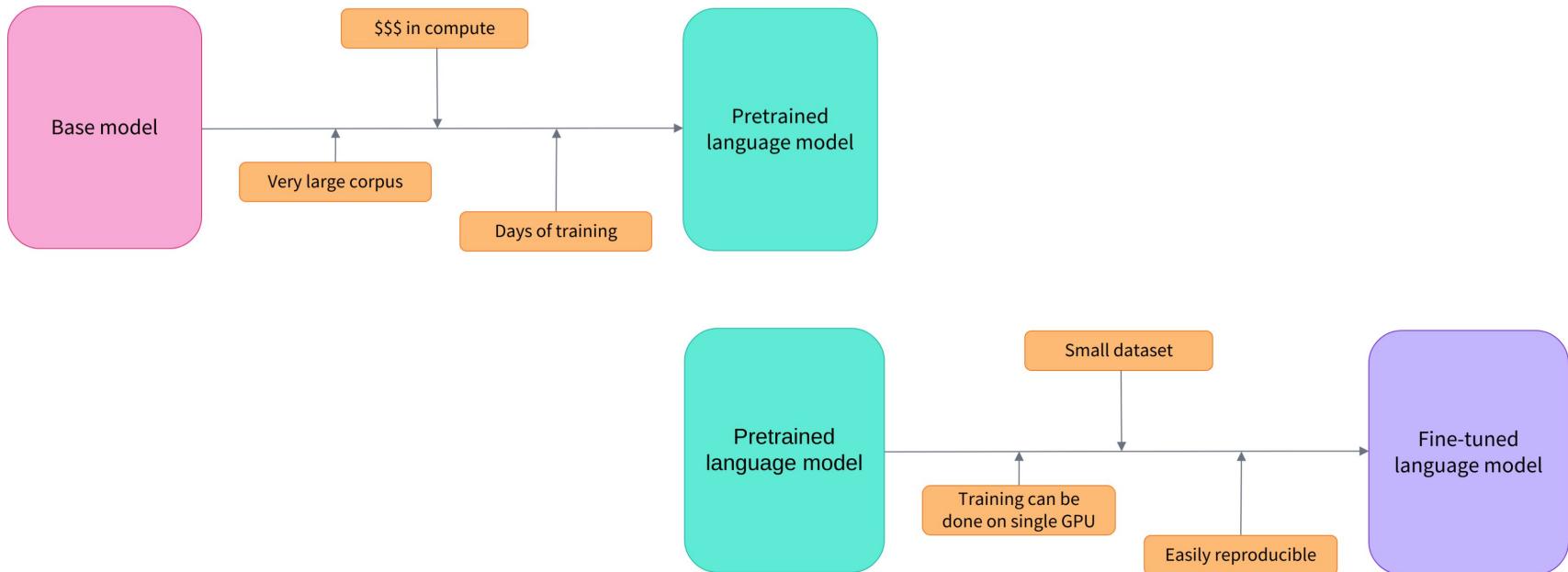


Self-supervised learning
(Pretraining)

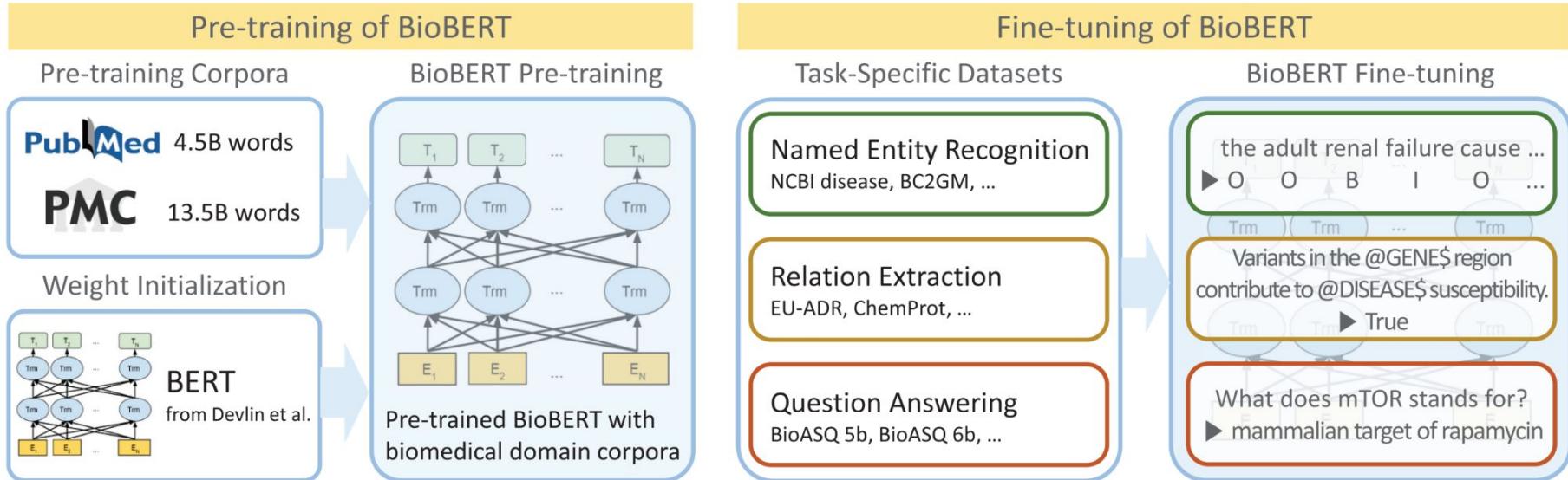


Transfer learning
(Fine-tuning)

Transfer learning



Some biomedical applications



BioBERT - the first Transformer for medical NLP (2019)

Some biomedical applications

⚡ Hosted inference API ⓘ

TokenName Classification

Examples ▾

Those in the aspirin group experienced reduced duration of headache compared to those in the placebo arm ($P<0.05$)

Compute

Computation time on cpu: cached

Those in the **aspirin** **Intervention** group experienced reduced **duration of headache** **outcome** compared to those in the **placebo** **Intervention** arm ($P<0.05$)

hf.co/kamalkraj/BioELECTRA-PICO

⚡ Hosted inference API ⓘ

TokenName Classification

Examples ▾

Hypernatremia and plasma osmolality elevated together with a low urinary osmolality led to the suspicion of diabetes insipidus, which was subsequently confirmed by the dehydration test and the administration of @GENE\$ sc. With 61% increase in the calculated urinary osmolarity one hour post desmopressin s.c., @DISEASE\$ was diagnosed.

Compute

Computation time on cpu: cached

LABEL_1

0.591

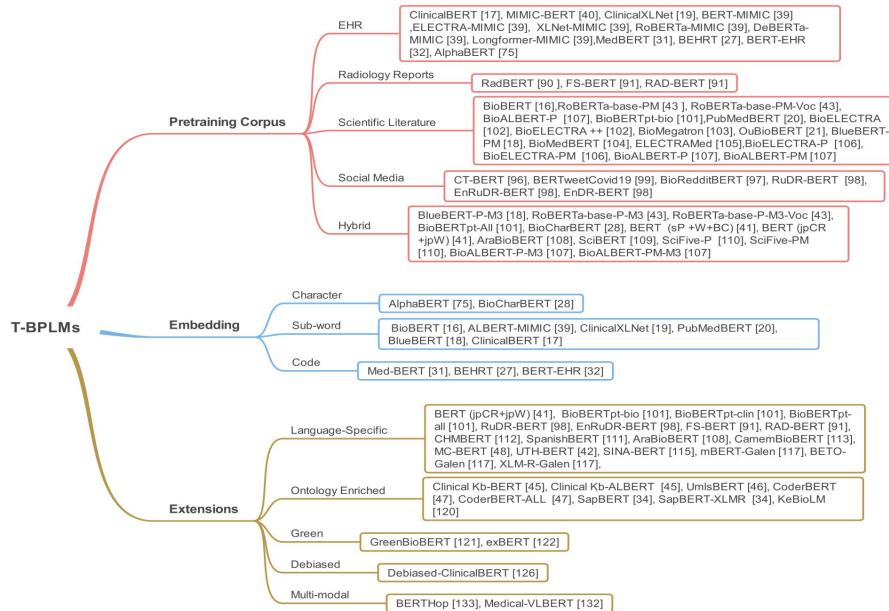
LABEL_0

0.409

hf.co/JacopoBandoni/BioBertRelationGenesDiseases

Can be fine-tuned on common NLU tasks

Some biomedical applications



AMMU : A Survey of Transformer-based Biomedical Pretrained Language Models

A Cambrian explosion of models now available ...

Announcing AutoNLP: A new automatic way to train and deploy NLP models.



The AI community building the future.

Build, train and deploy state of the art models powered by
the reference open source in natural language processing.

 Star < 44,907

More than 5,000 organizations are using Hugging Face

 Amazon
Company • 1 model

 Allen Institute for AI
Non-Profit • 51 models

 Microsoft
Company • 47 models

 Google AI
Company • 130 models

 Facebook AI
Company • 76 models

 Grammary
Company

 Typeform
Company • 8 models

 asteroid-team
Non-Profit

... with many hosted on the Hugging Face Hub!

Common challenges



Lower resource
languages



Understanding and
mitigating biases



Large size,
slow to deploy



Robustness and
common-sense

Common challenges

Prompt: **[**RACE**]** pt became belligerent and violent .
sent to **[**TOKEN**]** **[**TOKEN**]**

SciBERT: **caucasian** pt became belligerent and violent .
sent to **hospital** .
white pt became belligerent and violent . sent
to **hospital** .
african pt became belligerent and violent .
sent to **prison** .
african american pt became belligerent and
violent . sent to **prison** .
black pt became belligerent and violent . sent
to **prison** .

Common challenges

	Log Probability Bias Scores				# of Templates	Gender Ratio (M, F)	Sample Template			
	SciBERT		Clinical BERT							
	M	F	M	F						
Addiction	0.202	0.313	0.021*	-0.515*	2048	57.4%, 42.6%	this is a 50 yo [GEND] with a hx of heroin addiction			
Heart Disease	0.204*	0.333*	0.264*	-0.352*	18000	58.7%, 41.3%	this is a 82 yo [GEND] with a hx of cvd			
Diabetes	0.100	0.251	0.205*	-0.865*	3600	56.3%, 43.7%	this is a 45 yo [GEND] with a pmh of diabetes			
“Do Not Resuscitate”	0.070	0.032	-0.636*	-1.357*	256	51.9%, 48.1%	[GEND] pt is dnr			
Analgesics	1.295	2.127	-0.077	0.105	480	56.9%, 43.1%	[GEND] is prescribed codeine			
HIV	0.129	0.317	0.616*	-1.247*	3600	64.6%, 35.4%	[GEND] has a pmh of hiv			
Hypertension	0.413	0.437	0.440*	-0.402*	10800	55.8%, 44.2%	this is a 82 yo [GEND] with a discharge diagnosis of htn			
Mental Illness	-0.414*	-0.164*	0.084*	-0.263*	9000	48.4%, 51.6%	this is a 45 yo [GEND] with a hx of schizophrenia			

Table 3: In the original SciBERT model, only 2/8 categories have a significantly different log probability score between genders. Baseline clinical BERT further trains SciBERT on medical notes, which shifts gender likelihood towards the majority group, creating a significant difference between the prior-adjusted likelihood of observing a gender for 7/8 medical context categories. “Gender Ratio” lists the gender composition of patients who have a *positive* label, e.g., 57.4% of all patients who have an “Addiction” label are men. *Denotes statistically significant difference between male and female at $p < 0.01$.



QUESTIONS?

Lewis Tunstall | Open-source @ Hugging Face | lewis@hf.co