AIM 2: Artificial Intelligence in Medicine II

Embeddings and Transformers



Carlos Morato, Phd.

Outline for today's class

- Embeddings and their role in NLP
- Transformers and RNNs
- Stack-encoder and Stack-decoder architectures
- BERT and GPT
- Hugging Face library for NLP applications
- Clinical BERT and BioBERT
- LLM-based medical question-answering.

Text Representations

- Co-occurrence statistics
 - Brown Clusters
 - Count vectors, TF-IDF vectors, co-occurrence matrix decomposition
- Predictive
 - word2vec, GloVe, CBOW, Skip-Gram, etc
- Contextualized language models
 - Representation of word *changes* based on context
 - CoVE, ELMo, GPT, BERT, etc

Word Embeddings

Representing words as vectors in a canonical space:

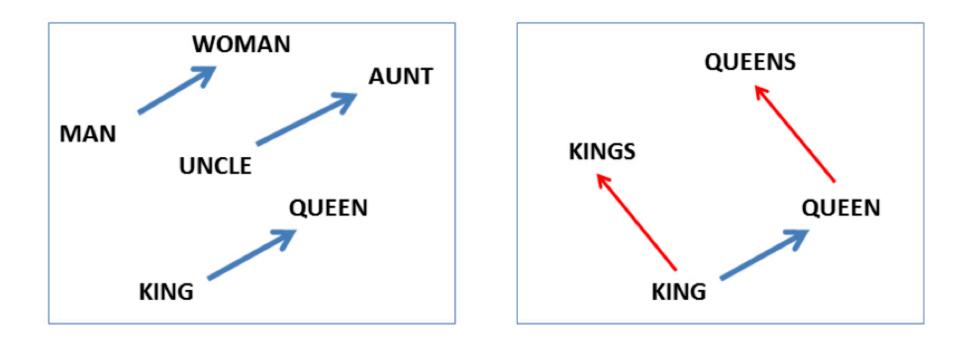
- Two distinct models
 - CBoW
 - Skip-Gram

(SG)

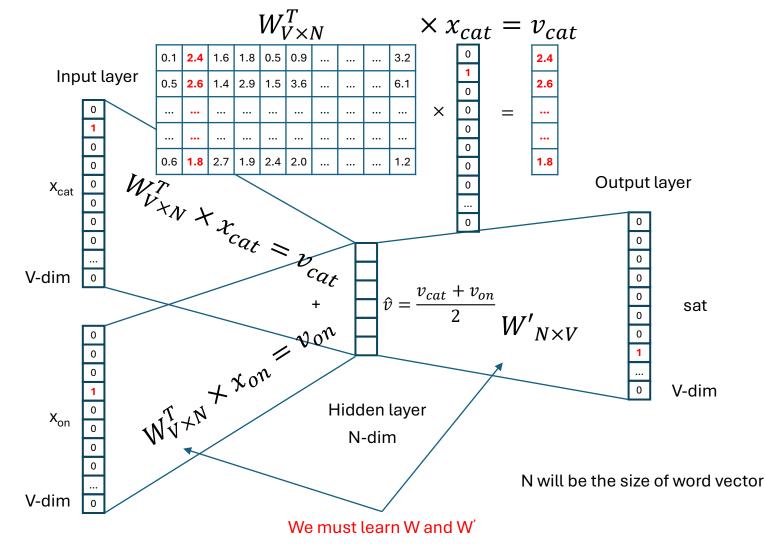
- Various training methods
 - Negative Sampling (NS)
 - Hierarchical Softmax
- A rich preprocessing pipeline
 - Dynamic Context Windows
 - Subsampling
 - Deleting Rare Words

Embeddings capture relational meaning

vector('king') - vector('man') + vector('woman') \approx vector('queen') vector('Paris') - vector('France') + vector('Italy') \approx vector('Rome')



Computing Embeddings

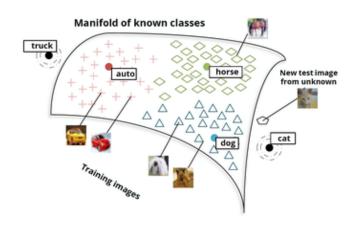


Word embedding applications

- The use of word representations... has become a key "secret sauce" for the success of many NLP systems in recent years, across tasks including named entity recognition, part-of-speech tagging, parsing, and semantic role labeling. (Luong et al. (2013))
- Learning a good representation on a task A and then using it on a task B is one of the major tricks in the Deep Learning toolbox.
 - Pretraining, transfer learning, and multi-task learning.
 - Can allow the representation to learn from more than one kind of data.
- Can learn to map multiple kinds of data into a single representation.

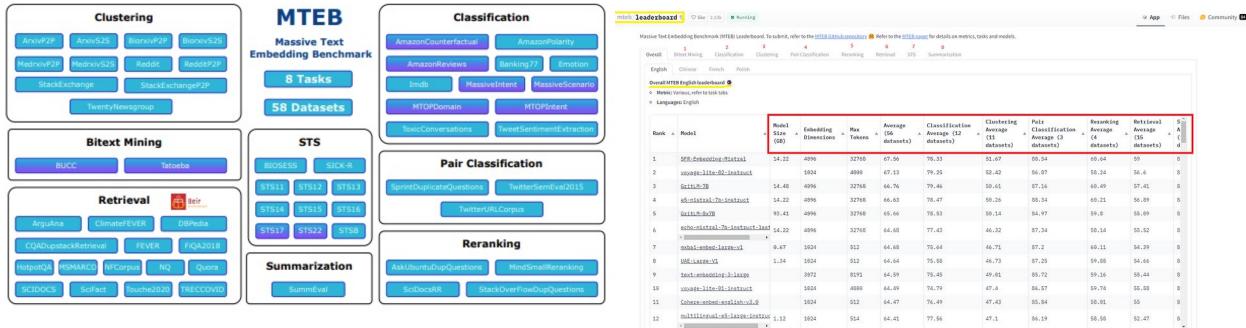
Word embedding applications

- Can apply to get a joint embedding of words and images or other multi-modal data sets.
- New classes map near similar existing classes: e.g., if 'cat' is unknown, cat images map near dog.



(Socher et al. (2013b)) http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

Massive text Embedding Benchmark



1.000

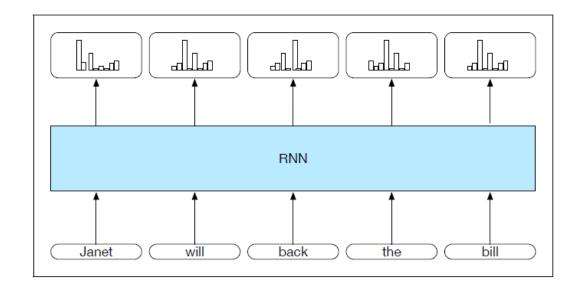
.

An overview of tasks and datasets in MTEB. Multilingual datasets are marked with a purple shade.

https://huggingface.co/spaces/mteb/leaderboard

Encoder-Decoder

• **RNN:** input sequence is transformed into output sequence in a one-to-one fashion.



• **Goal:** Develop an architecture capable of generating contextually appropriate, arbitrary length, output sequences

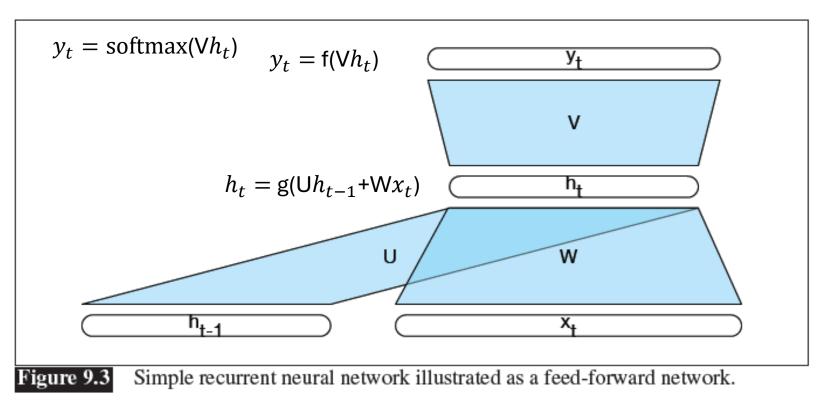
• Applications:

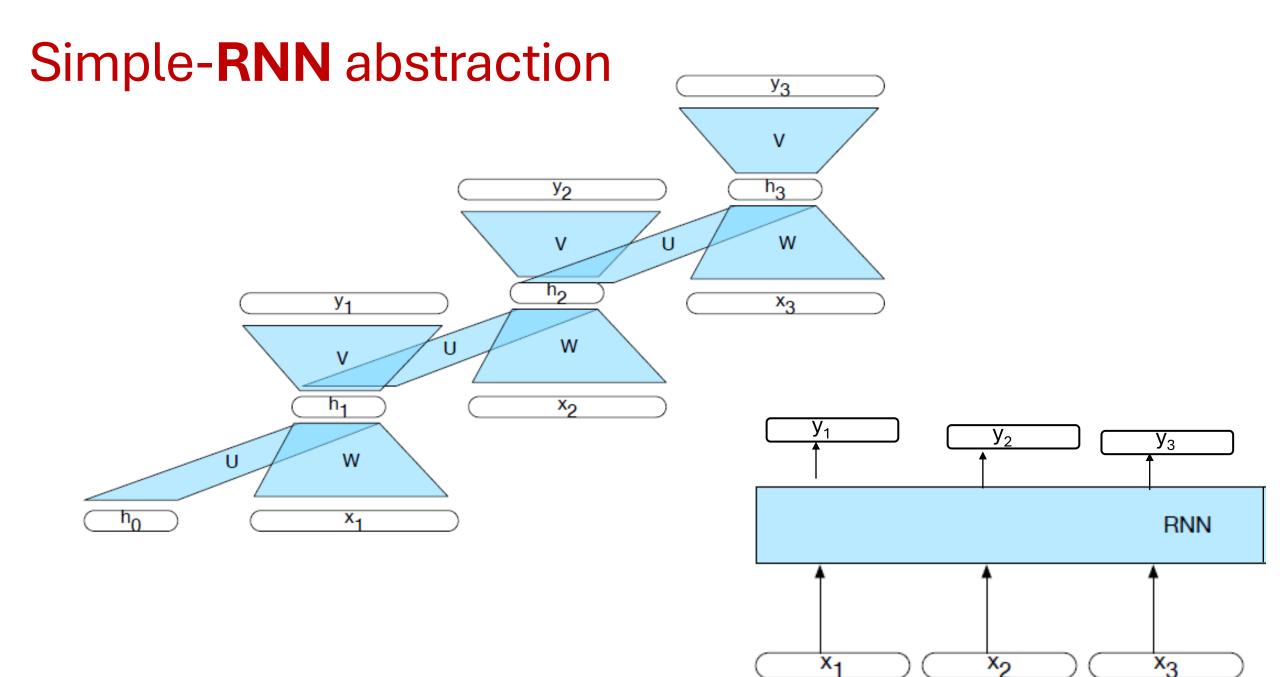
- Machine translation,
- Summarization,
- Question answering,
- Dialogue modeling.

Simple recurrent neural network illustrated as a feed-forward network

Most significant change: new set of weights, U

- connect the hidden layer from the previous time step to the current hidden layer.
- determine how the network should make use of past context in calculating the output for the current input.



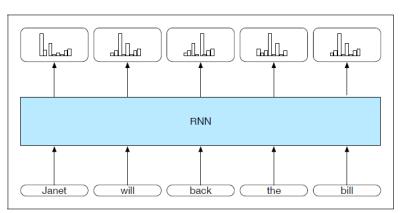


P(W I "The") P(W | "...quick") P(W | "...brown") P(W | "...fox") **RNN** Applications Softmax Softmax Softmax Softmax ملله ملالم ممالا Gallo RNN RNN **RNN** RNN Language Modeling "fox" 'The' brown RNN Softmax Xa Xo Sequence Classification (Sentiment, Topic, intent, RNN

Xo

• Sequence to Sequence

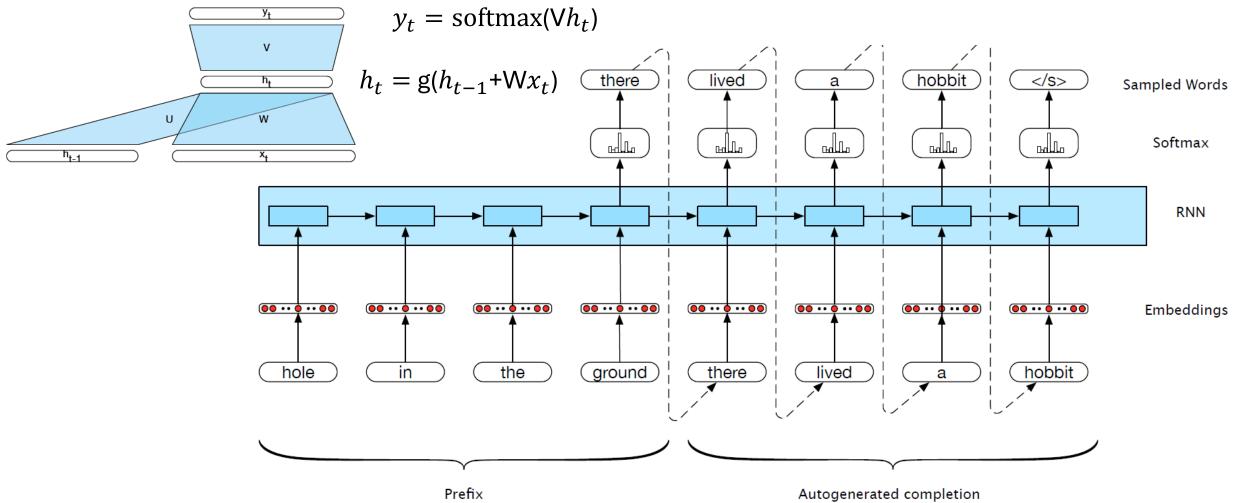
• •



Xa

xn

Sentence Completion using an RNN



- Trained Neural Language Model can be used to generate novel sequences
- Or to **complete** a given sequence (until end of sentence token <\s> is generated)

Extending (autoregressive) generation to Machine Translation

Autoregressive: word generated at each time step is conditioned on word from previous step.

• Training data are parallel text e.g., English / French

there lived a hobbit vivait un hobbit

••••

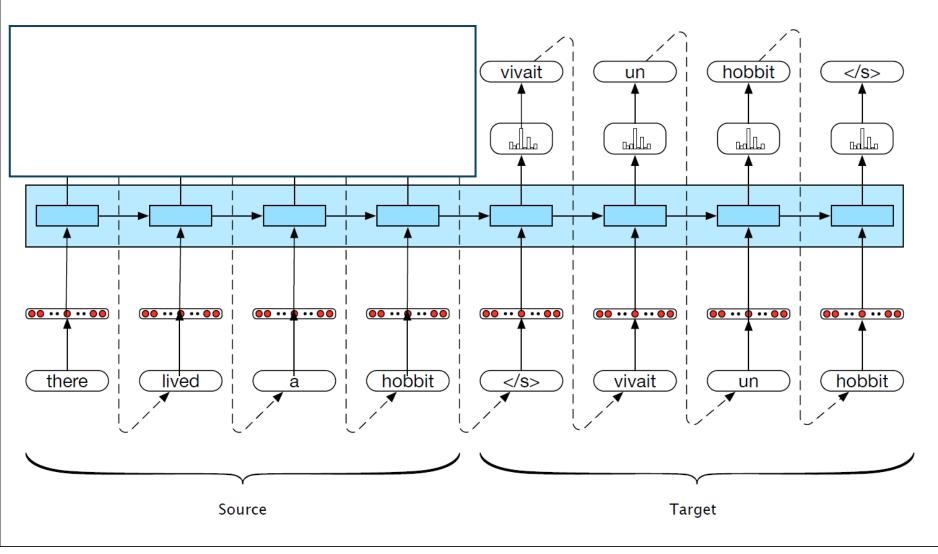
 Build an RNN language model on the concatenation of source and target

there lived a hobbit <\s> vivait un hobbit <\s>

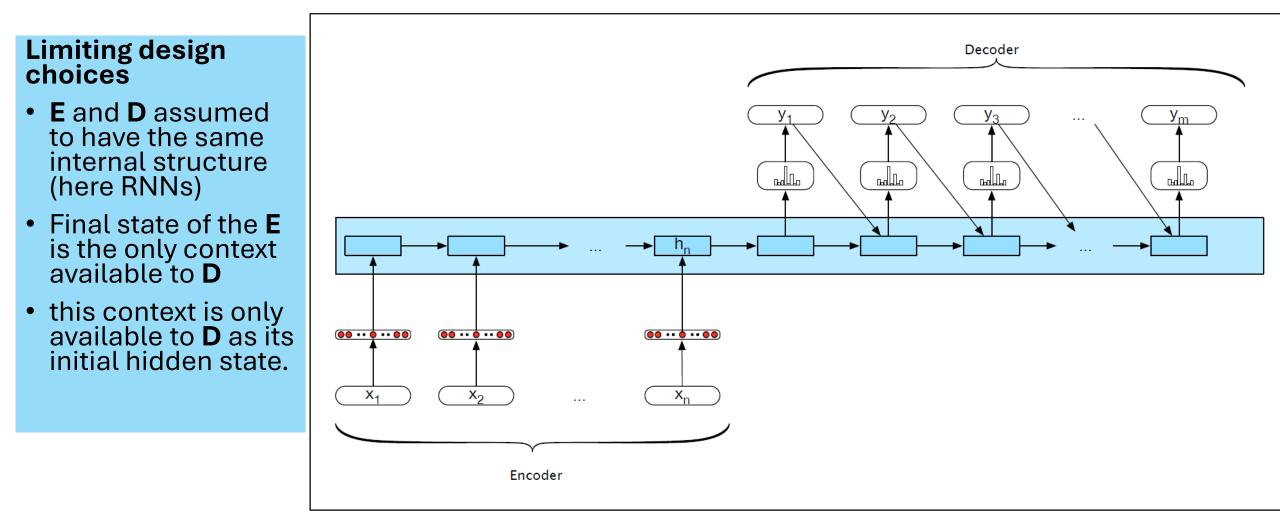
••••

Extending (autoregressive) generation to Machine Translation

 Translation as Sentence Completion



(simple) Encoder Decoder Networks



- Encoder generates a contextualized representation of the input (last state).
- Decoder takes that state and autoregressively generates a sequence of outputs

General Encoder Decoder Networks

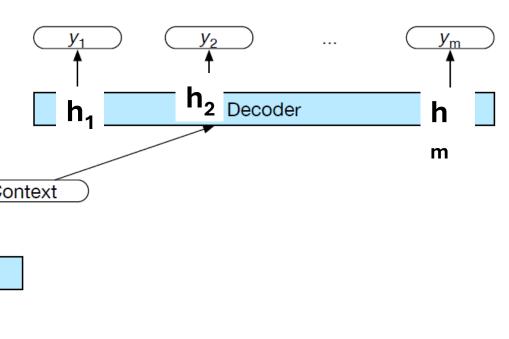
Abstracting away from these choices

- Encoder: accepts an input sequence, x_{1:n} and generates a corresponding sequence of contextualized representations, h_{1:n}
- 2. Context vector c: function of $h_{1:n}$ and conveys the essence of the input to the decoder.
- **3.** Decoder: accepts **c** as input and generates an arbitrary length sequence of hidden states $h_{1:m}$ from which a corresponding sequence of output states $y_{1:m}$ can be obtained.

h₂

 X_{2}

Encoder

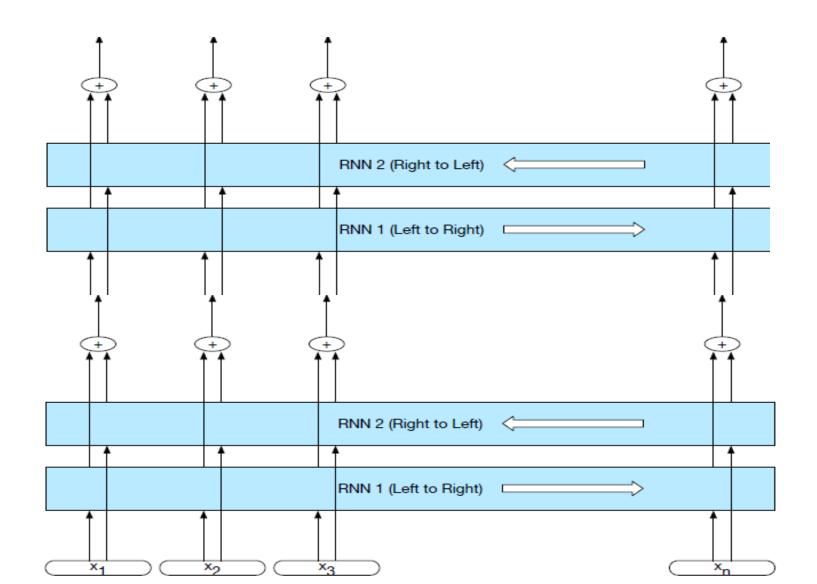


h_n

Popular architectural choices: Encoder

Widely used encoder design: **stacked Bi-LSTMs**

 Contextualized representations for each time step: hidden states from top layers from the forward and backward passes

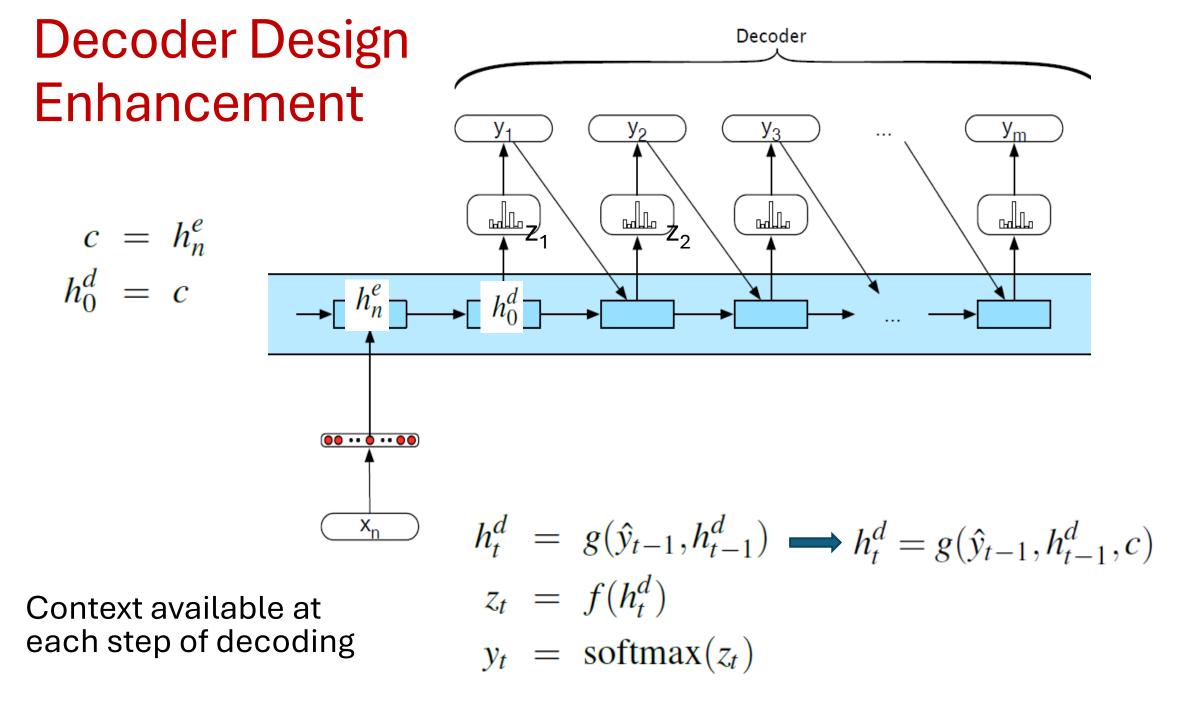


Decoder Basic Design

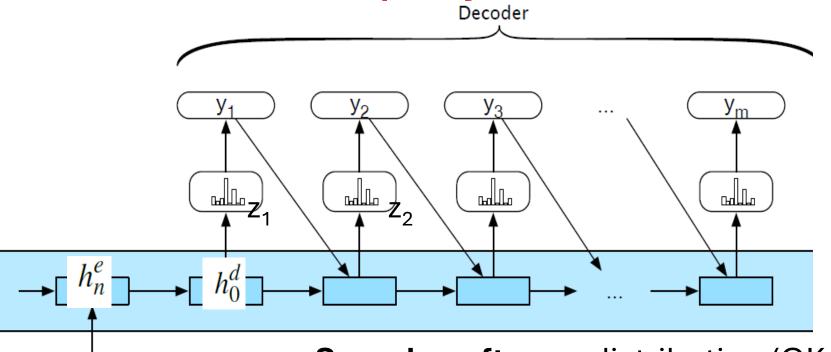
 produce an output sequence an element at a time

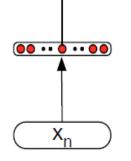
 $c = h_n^e$ $h_0^d = c$

Decoder У_m Уз ... Last hidden z₂ سلله z′1 state of the տեհ encoder h_n^e _► First hidden state of the $h_t^d = g(\hat{y}_{t-1}, h_{t-1}^d)$ decoder $z_t = (Vh^{\tilde{d}}_t)$ Xn $y_t = \operatorname{softmax}(z_t)$

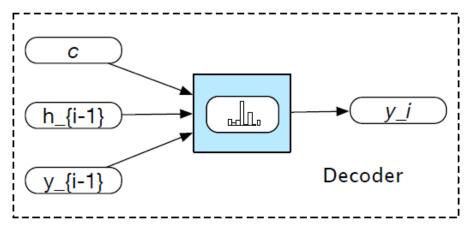


Decoder: How output y is chosen

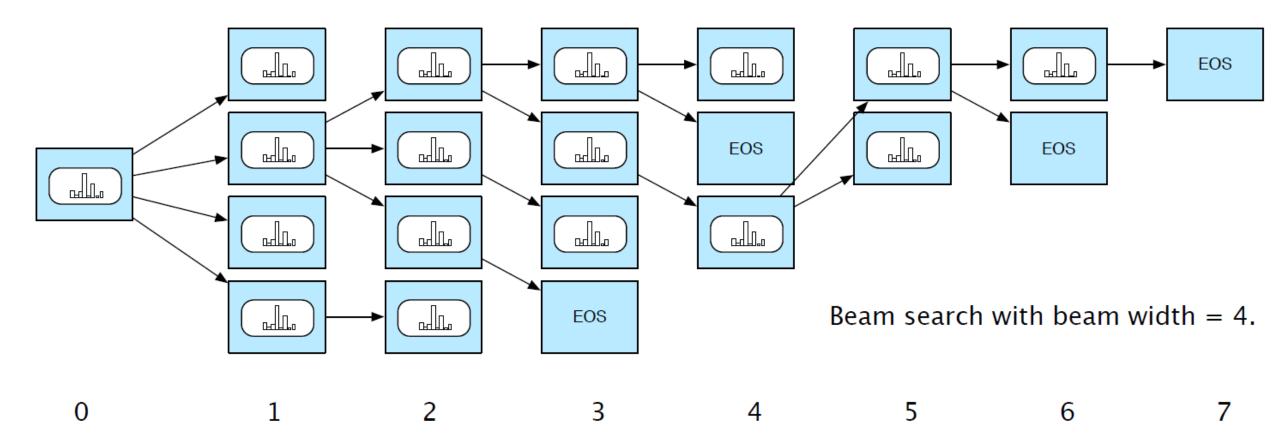




- Sample soft-max distribution (OK for generating novel output, not OK for e.g. MT or Summ)
- Most likely output (doesn't guarantee individual choices being made make sense together)



- 4 most likely "words" decoded from initial state
- Feed each of those in decoder and keep most likely 4 sequences of two words
- Feed most recent word in decoder and keep most likely 4 sequences of three words
- When EOS is generated. Stop sequence and reduce Beam by 1

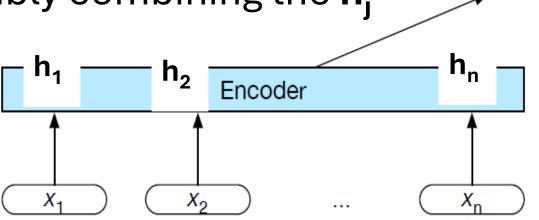


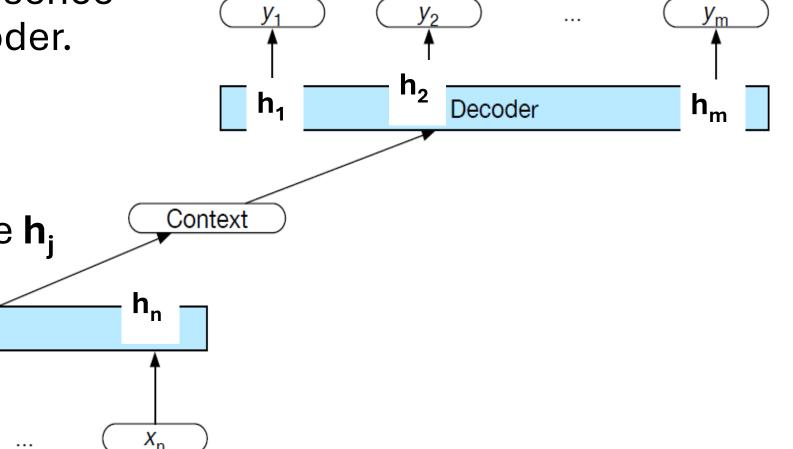
Flexible context: Attention

Context vector c: function of $h_{1:n}$ and conveys the essence of the input to the decoder.

Flexible?

- Different for each h_i
- Flexibly combining the h_j





Attention (1): dynamically derived context

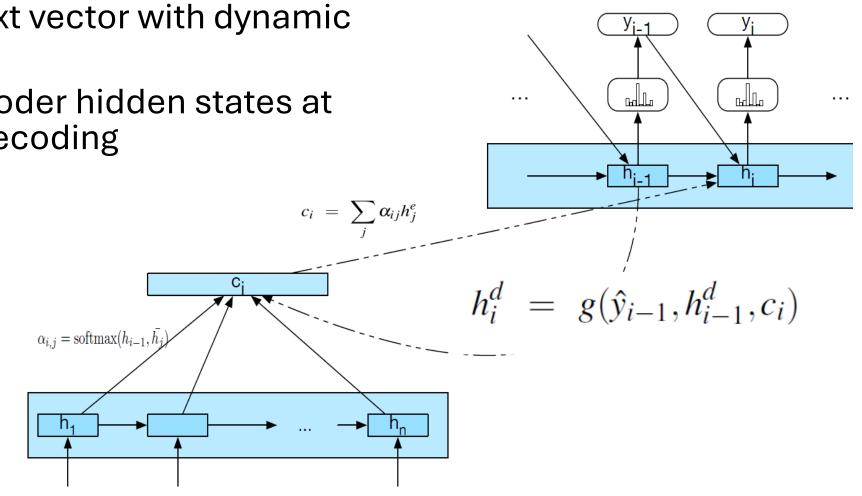
- Replace static context vector with dynamic
 c_i
- derived from the encoder hidden states at each point *i* during decoding

Ideas:

 should be a linear combination of those states

$$c_i = \sum_j \alpha_{ij} h$$

• α_{ij} should depend on ?



Attention (2): computing c_i

• Compute a vector of scores that capture the relevance of each encoder hidden state to the decod h_{i-1}^d ate

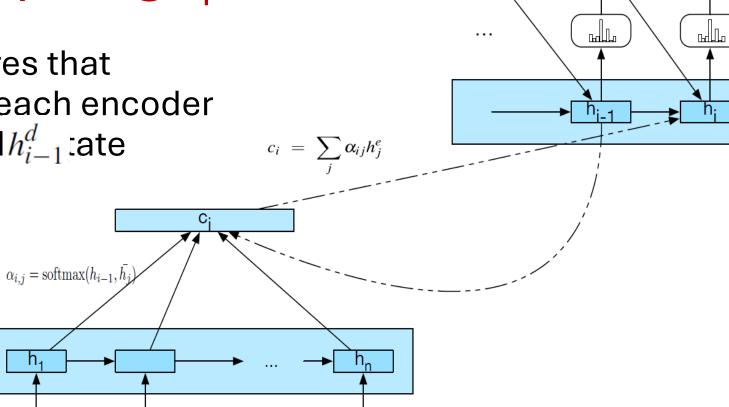
 $score(h_{i-1}^d, h_j^e)$

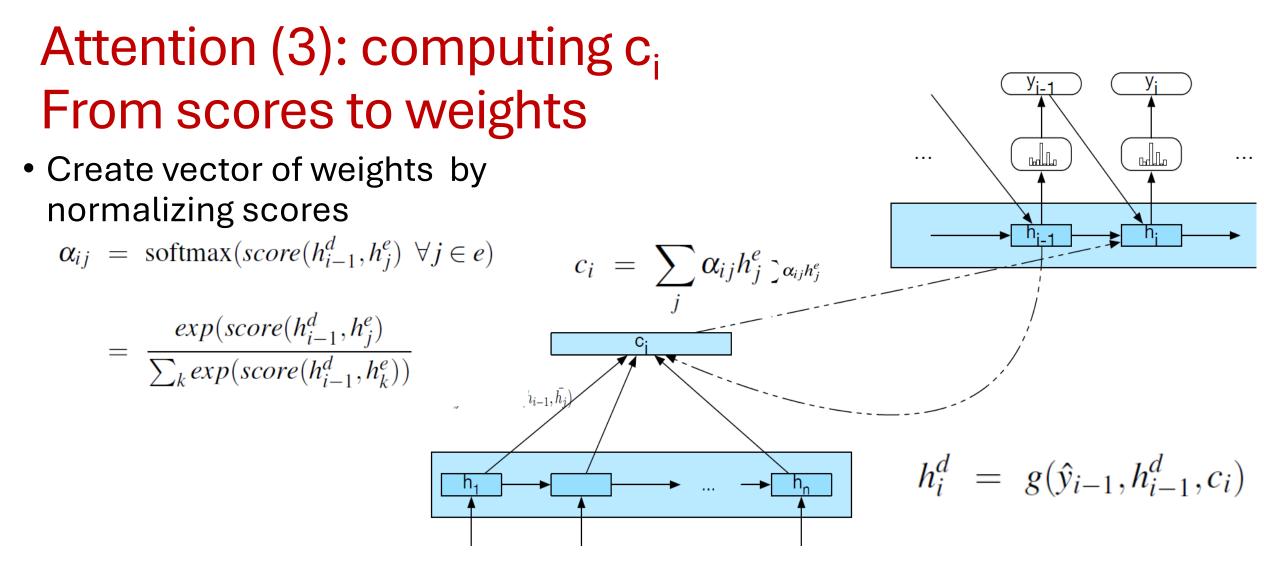
• Just the similarity

 $score(h_{i-1}^d, h_j^e) = h_{i-1}^d \cdot h_j^e$

• Give network the ability to learn which aspects of similarity between the decoder and encoder states are important to the current application.

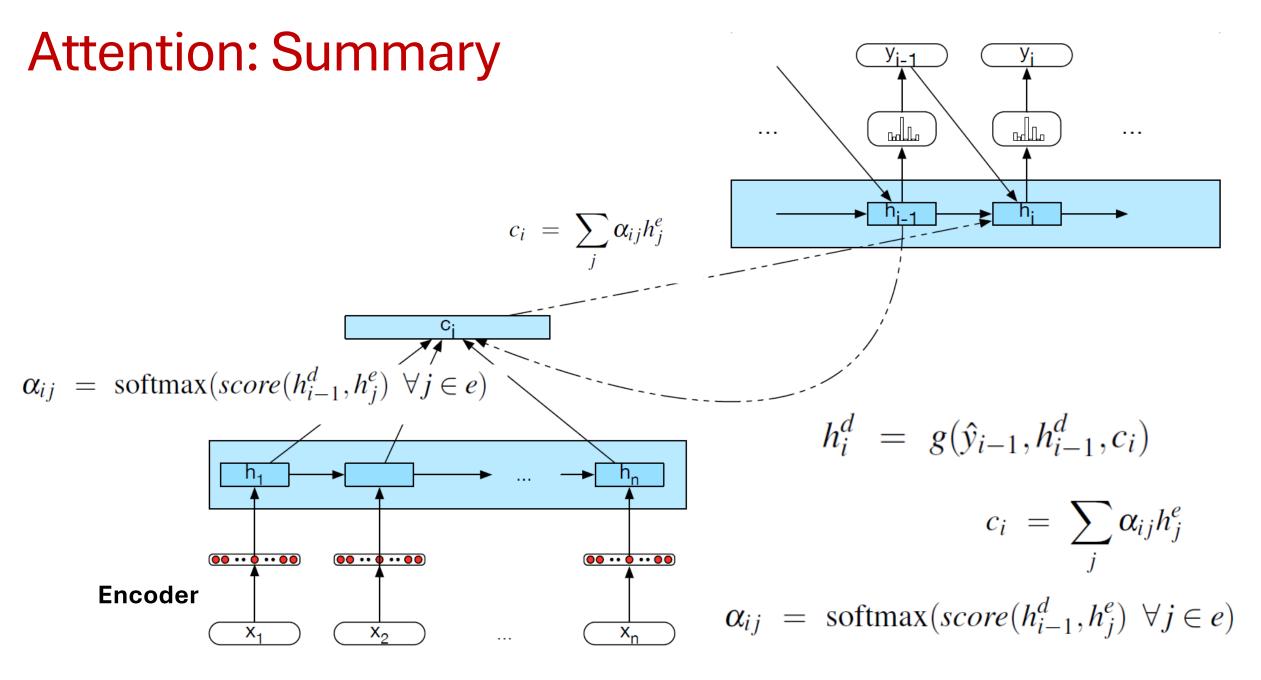
$$score(h_{i-1}^d, h_j^e) = h_{t-1}^d W_s h_j^e$$





• **Goal achieved**: compute a fixed-length context vector for the current decoder state by taking a weighted average over all the encoder hidden states.

Decoder



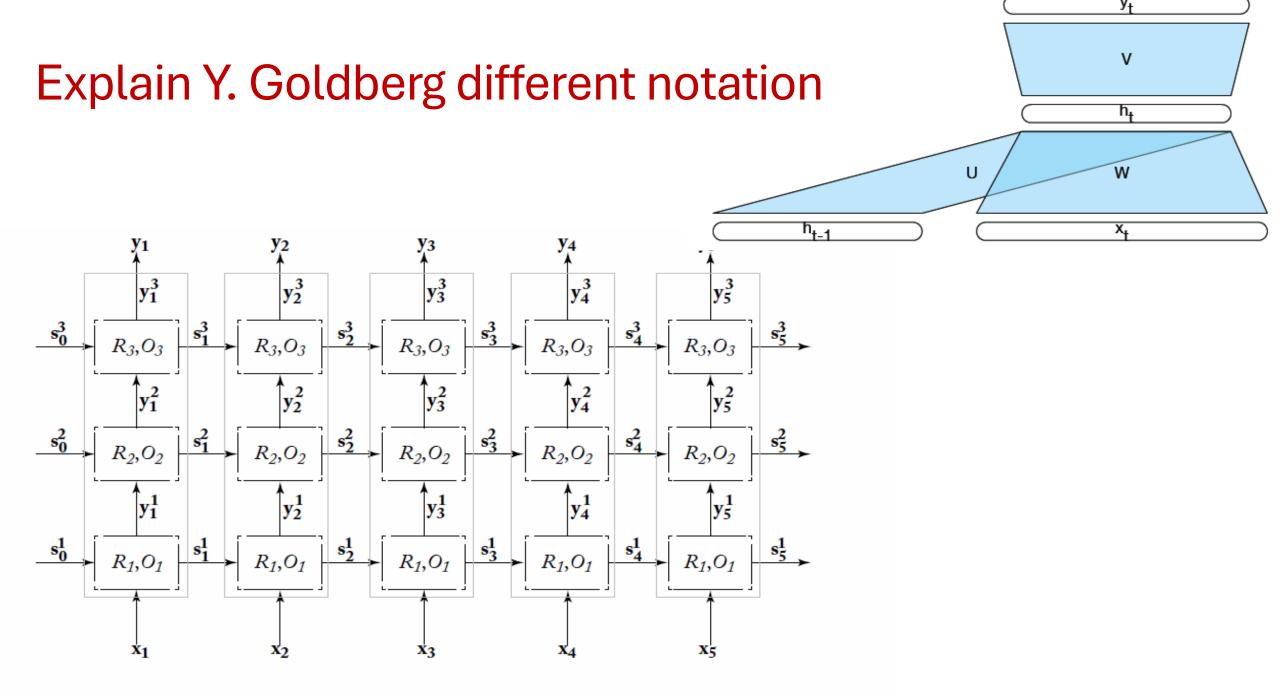
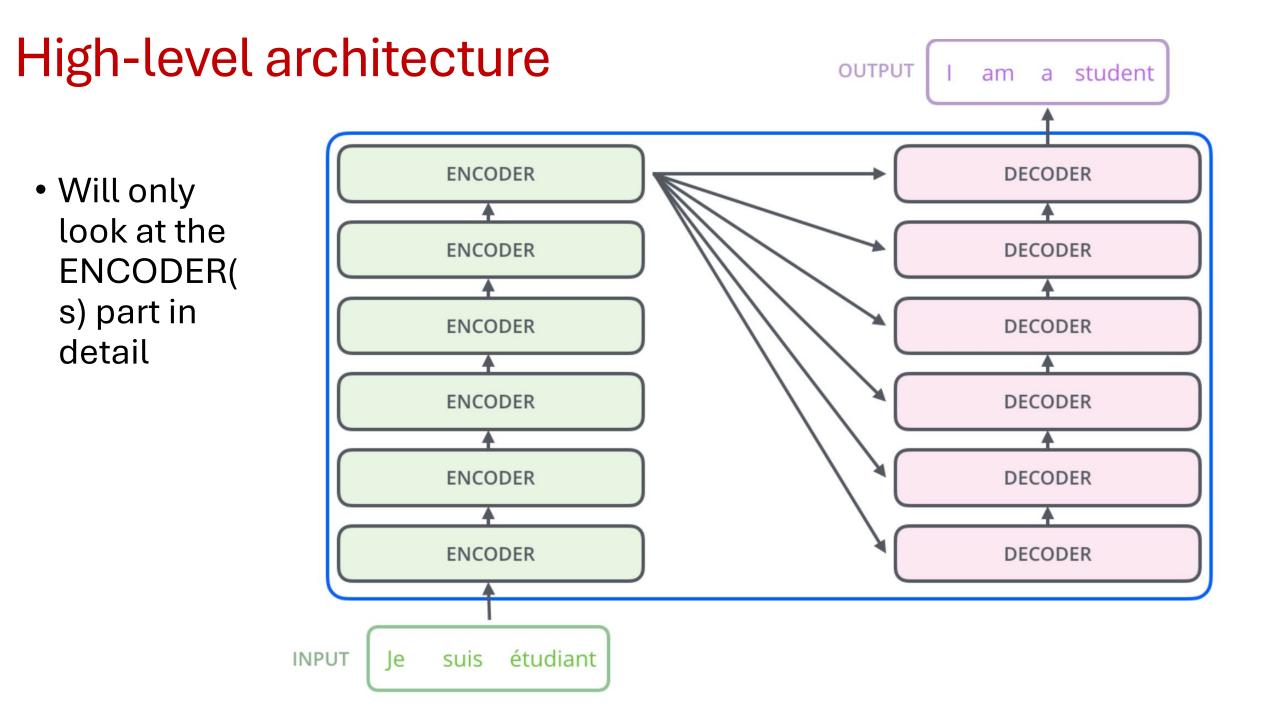
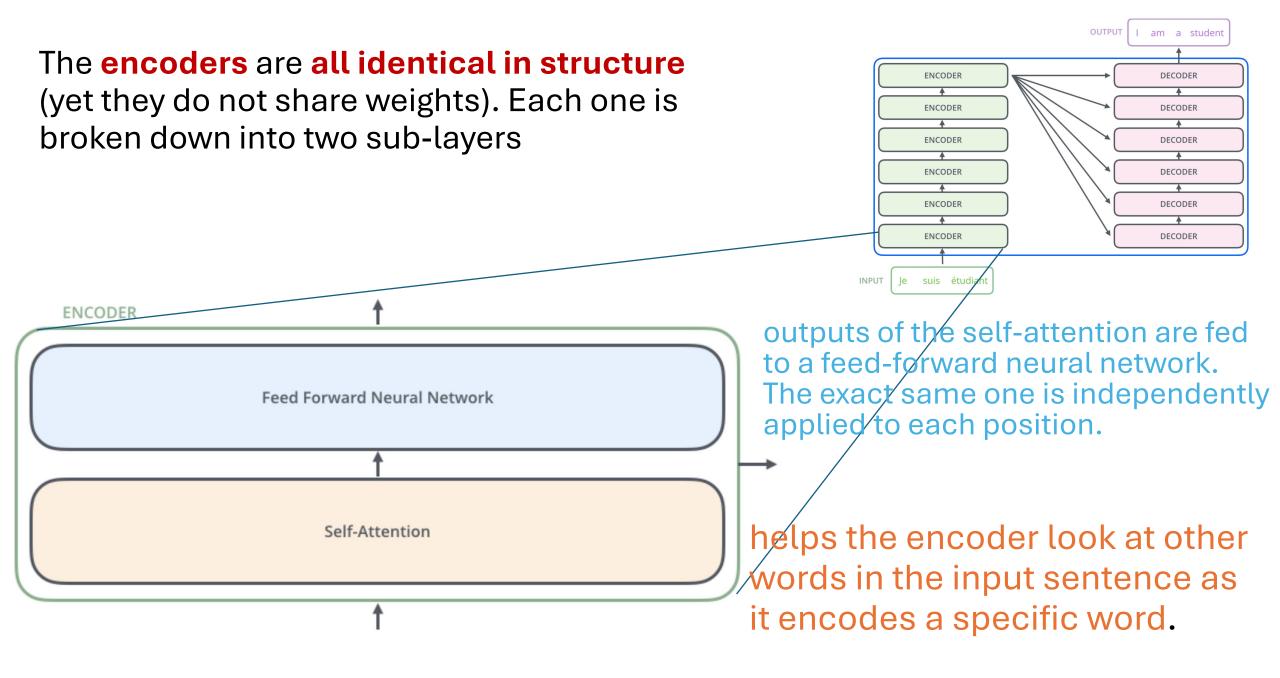


Figure 14.7: A three-layer ("deep") RNN architecture.

Transformers (Attention is all you need 2017)

- Just an introduction: These are two valuable resources to learn more details on the architecture and implementation
- <u>https://nlp.seas.harvard.edu/annotated-transformer/</u>
- <u>https://jalammar.github.io/illustrated-transformer/</u> (slides come from this source)

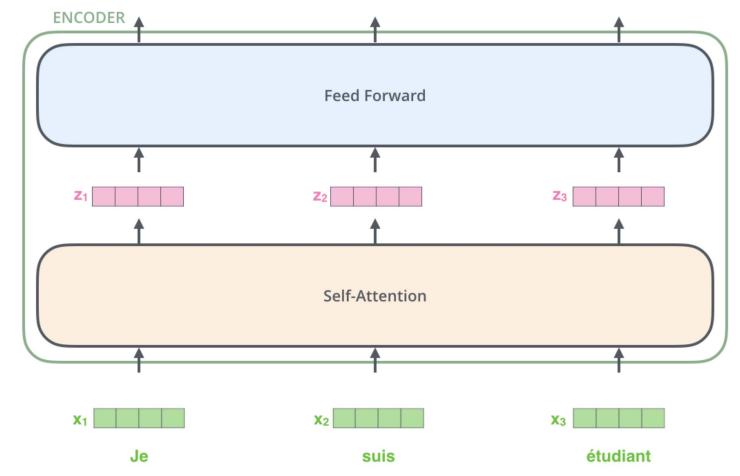




Key property of Transformer:

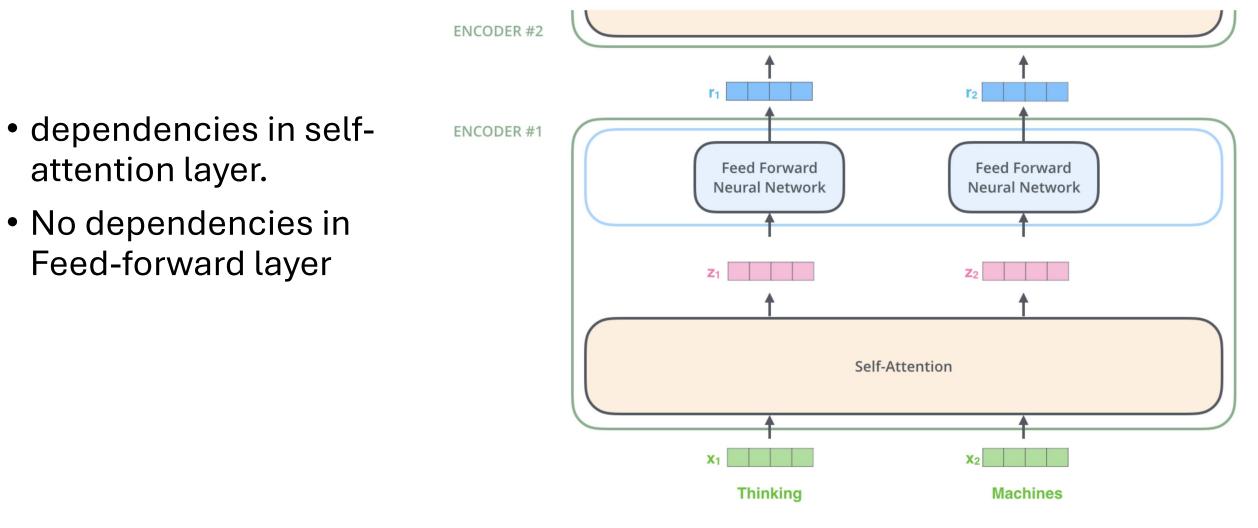
word in each position flows through its own path in the encoder.

- There are dependencies between these paths in the self-attention layer.
- Feed-forward layer does not have those dependencies => various paths can be executed in parallel !



Word embeddings

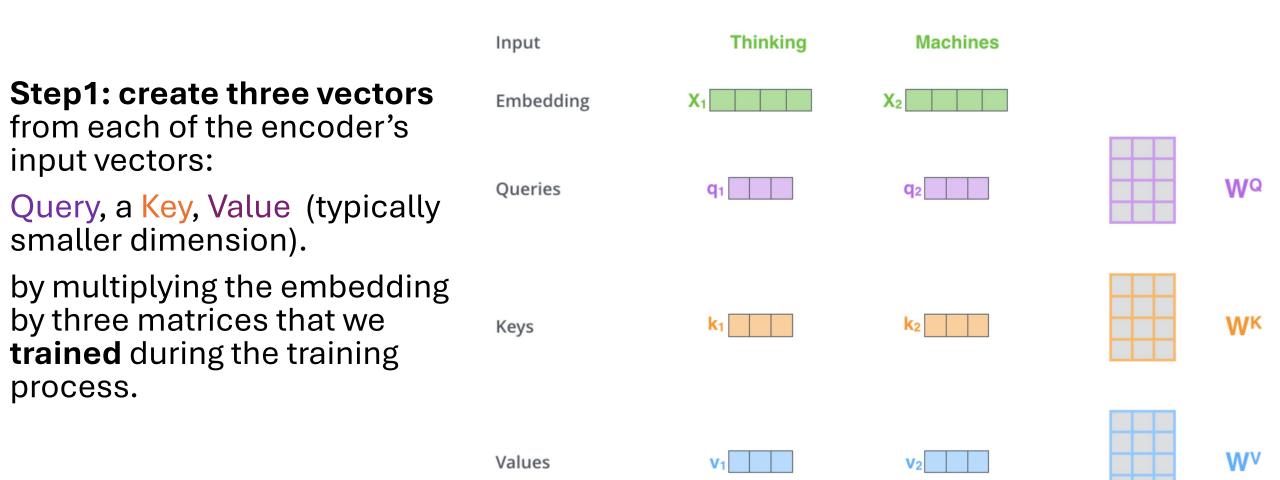
Visually clearer on two words



Word embeddings

Self-Attention

While processing **each word** it allows to look at other positions in the input sequence for clues to build a better encoding for **this word**.

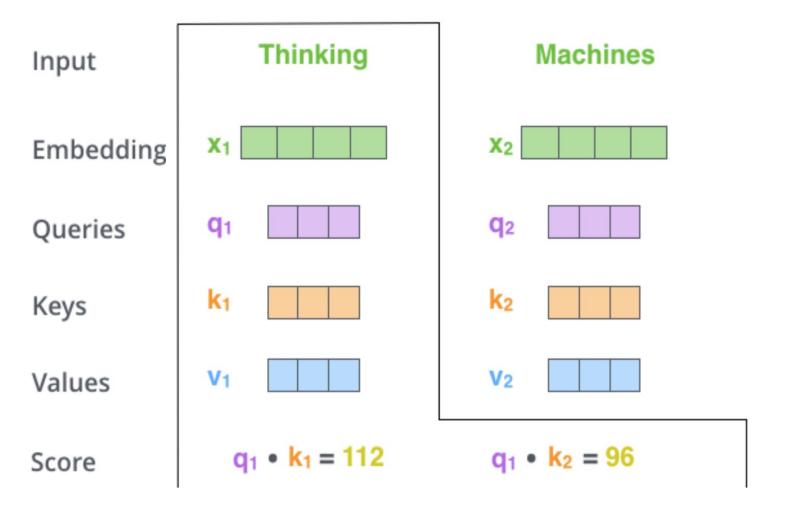


Self-Attention

Step 2: calculate a score

(like we have seen for regular attention!) how much focus to place on other parts of the input sentence as we encode a word at a certain position.

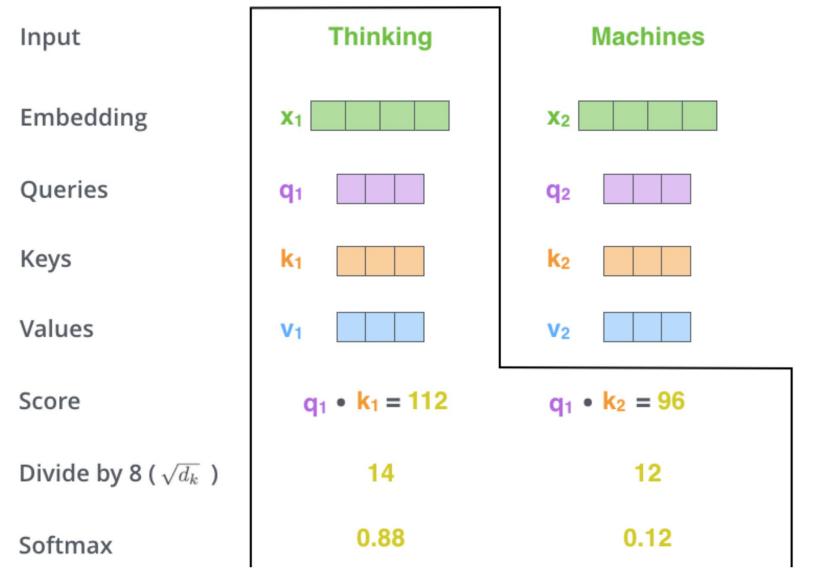
Take dot product of the query vector with the key vector of the respective word we're scoring.



E.g., Processing the self-attention for word "Thinking" in position #1, the first score would be the dot product of q1 and k1. The second score would be the dot product of q1 and k2.

Self Attention

- **Step 3** divide scores by the square root of the dimension of the key vectors (more stable gradients).
- **Step 4** pass result through a SoftMax operation. (all positive and add up to 1)



Intuition: SoftMax score determines how much each word will be expressed at this position.

Self Attention

• **Step6** : sum up the weighted value vectors. This produces the output of the selfattention layer at this position More details:

Input

Embedding

Queries

Keys

Values

Score

Softmax

Softmax

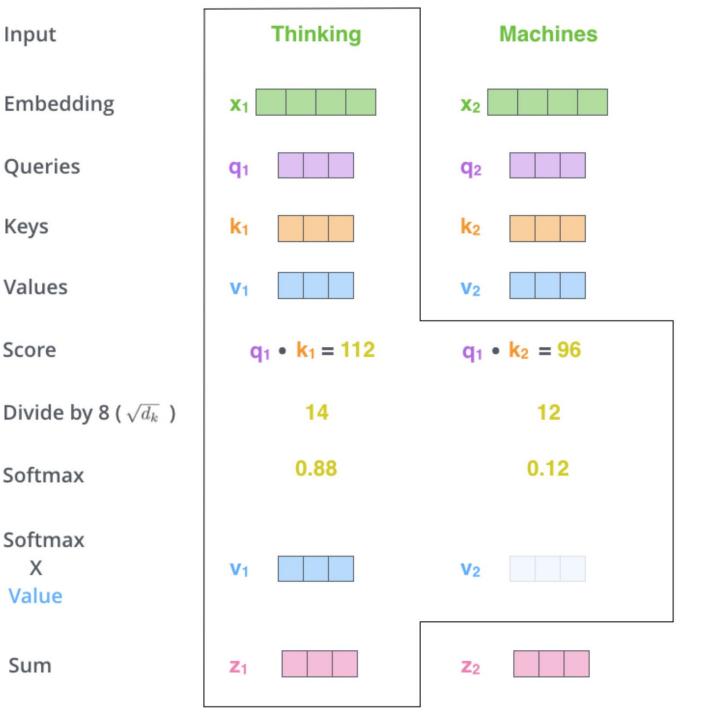
Х

Value

Sum

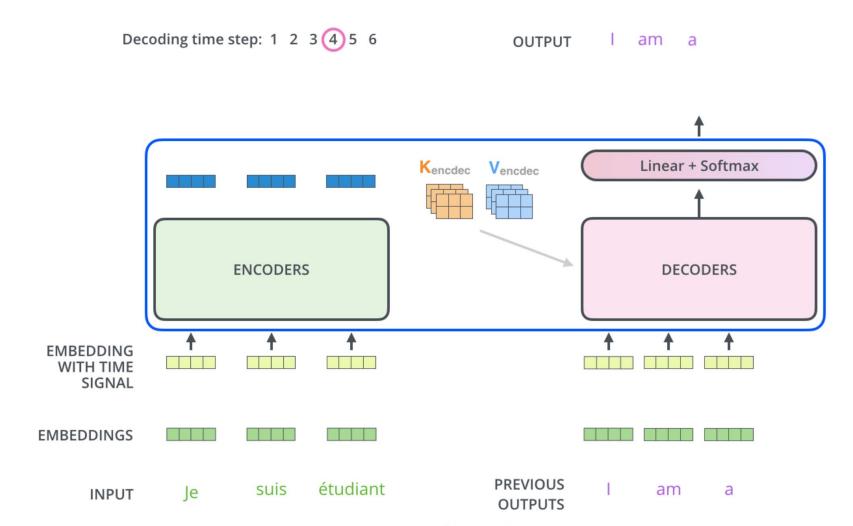
- What we have seen for a word is done for all words (using matrices)
- Need to encode position of words
- And improved using a mechanism called "multi-headed" attention (kind of like multiple filters for CNN)

see https://jalammar.github.io/illustratedtransformer/



The Decoder Side

- Relies on most of the concepts on the encoder side
- See animation on https://jalammar.github.io/illustrated-transformer/



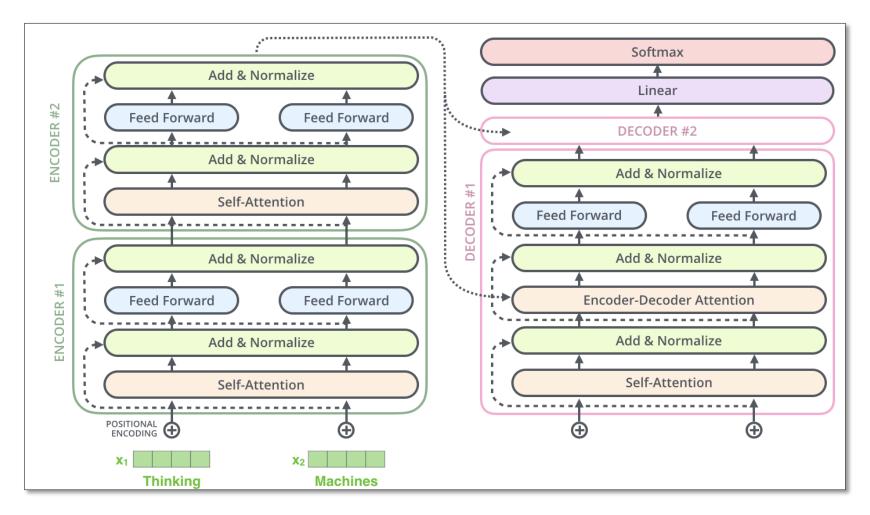
Stack for Decoder only and Stack for Encoder only

- The **RNN** and **LSTM** neural models were designed to process language and perform tasks like classification, summarization, translation, and sentiment detection
 - RNN: Recurrent Neural Network
 - LSTM: Long Short Term Memory
- In both models, layers get the next input word and have access to some previous words, allowing it to use the word's left context
- They used word embeddings where each word was encoded as a vector of 100-300 real numbers representing its meaning

Stack for Decoder only and Stack for Encoder only

- Transformers extend this to allow the network to process a word input knowing the words in both its left and right context
- This provides a more powerful context model
- Transformers add additional features, like <u>attention</u>, which identifies the important words in this context
- And break the problem into two parts:
 - An encoder (e.g., Bert)
 - A decoder (e.g., GPT)

Transformer model



Encoder (e.g., BERT)

Decoder (e.g., GPT)

Transformers, GPT-2, and BERT

- A transformer uses an encoder stack to model input, and uses decoder stack to model output (using input information from encoder side)
- 2. If we do not have input, we just want to model the "next word", we can get rid of the encoder side of a transformer and output "next word" one by one. This gives us **GPT**
- 3. If we are only interested in training a language model for the input for some other tasks, then we do not need the decoder of the transformer, that gives us **BERT**

Training a Transformer

- Transformers typically use semi-supervised learning with
 - Unsupervised pretraining over a very large dataset of general text
 - Followed by supervised **fine-tuning** over a focused data set of inputs and outputs for a particular task
- Tasks for pretraining and fine-tuning commonly include:
 - language modeling
 - next-sentence prediction (aka completion)
 - question answering
 - reading comprehension
 - sentiment analysis
 - paraphrasing

Pretrained models

- Since training a model requires huge datasets of text and significant computation, researchers often use common pretrained models
- Examples (circa December 2021) include
 - Google's <u>BERT</u> model
 - Huggingface's various <u>Transformer models</u>
 - OpenAI's and <u>GPT-3 models</u>

Huggingface Models

C' ☆ 🔒 huggingface.co/models 💀 💏 📅 🔟 🖺 👙 🗳 🛆 ◆ ୬ 🗘 ♀	û 🗠 me 🗅 🥲 📄 N 📄 G 😼 csee 📄 R 📄 util 📄 U 📄 S 📄 T 📄	w w	 Comparison of the sector of the
😕 Hugging Face 🔍 📦 Mo	odels 🗧 Datasets 🔠 Spaces 🗮 Resources 🚔 Solutions	s Pricing	Log In Sign Up
Tasks	Models 23,887 Search Models		1↓ Sort: Most Downloads
🔁 Fill-Mask 🎦 Question Answering			
Summarization	bert-base-uncased ☺ Fill-Mask • Updated May 18 • ↓ 24.9M • ♡ 72		
Table Question Answering	En mask - opuaced may 10 + + 24.5m + + 12		
Text Classification	sentence-transformers/paraphrase-multilin	gual-Mini	iLM-L12-v2
🐶 Text Generation	$\frac{OID}{\Delta IC}$ Sentence Similarity \circ Updated Nov 2 $\circ \downarrow$ 12.2M $\circ \heartsuit$ 10		
S Text2Text Generation			
📽 Token Classification	roberta-base Fill-Mask • Updated Jul 6 • \downarrow 5.21M • \heartsuit 9		
☆ _A Translation			
💥 Zero-Shot Classification	distilbert-base-uncased		
Sentence Similarity + 12	Fill-Mask ■ Updated Aug 29 ■ ↓ 5.01M ■ ♥ 30		
Libraries	gpt2		

OpenAl Application Examples

🗧 🎴 🍪 Examples - OpenAl API	× +	~
\leftarrow \rightarrow C \triangle \triangleq beta.openai.com/e	examples/	🖞 🖈 G 🛈 🛧 😋 🗔 뵭 👗 🗊 🌔 🗄
🏥 Apps 🔀 📬 🗵 🖺 👹 🌾 🛆	🔶 🎐 😚 🎧 🛅 me 🗎 🖖 🛅 N 🚍 G	😻 csee 📄 R 📄 util 📄 U 📄 S 🧼 📄 Other Bookmarks 🛛 📰 Reading List
S Overview Documentation E	kamples	Log in Sign up
—	Chat Open ended conversation with an AI assist	Q&A Answer questions based on existing knowle
2	Grammar correction Corrects sentences into standard English.	Summarize for a 2nd grader Translates difficult text into simpler concep
	Natural language to OpenAI API Create code to call to the OpenAI API usin	Text to command Translate text into programmatic commands.
•	English to French Translates English text into French.	Natural language to Stripe API Create code to call the Stripe API using nat
Θ	SQL translate Translate natural language to SQL queries.	Parse unstructured data Create tables from long form text
	Classification Classify items into categories via example.	# Python to natural language Explain a piece of Python code in human un
•	Movie to Emoji Convert movie titles into emoji.	Calculate Time Complexity Find the time complexity of a function.
Ż a	Translate programming languages	# Advanced tweet classifier

Text Representations

- Co-occurrence statistics
 - Brown Clusters
 - Count vectors, TF-IDF vectors, co-occurrence matrix decomposition
- Predictive
 - word2vec, GloVe, CBOW, Skip-Gram, etc
- Contextualized language models
 - Representation of word *changes* based on context
 - CoVE, ELMo, GPT, BERT, etc

Who is BERT?

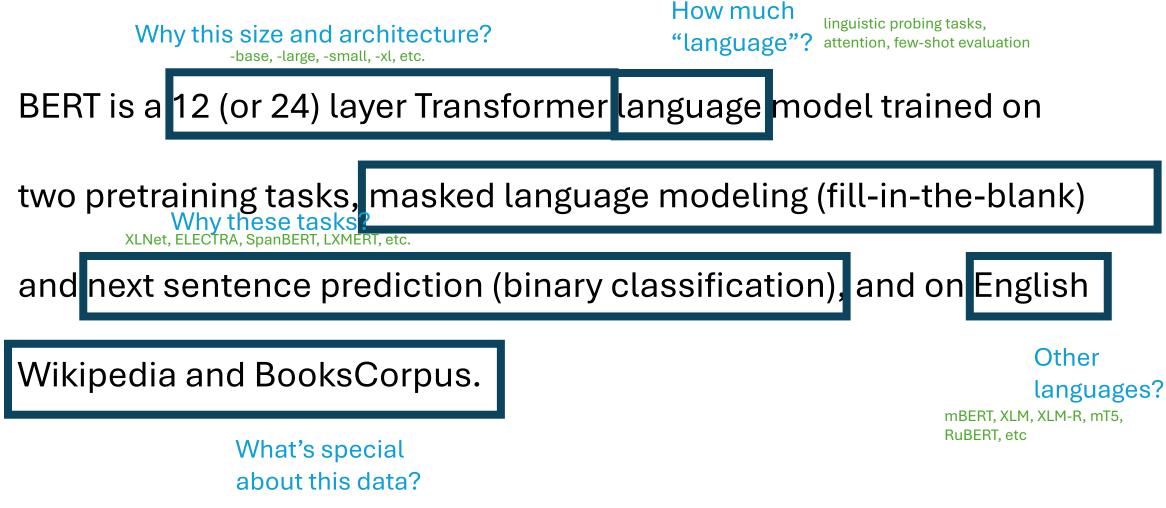
BERT is a 12 (or 24) layer Transformer language model trained on two

pretraining tasks, masked language modeling (fill-in-the-blank) and next

sentence prediction (binary classification), and on English Wikipedia and

BooksCorpus.

About BERT and friends

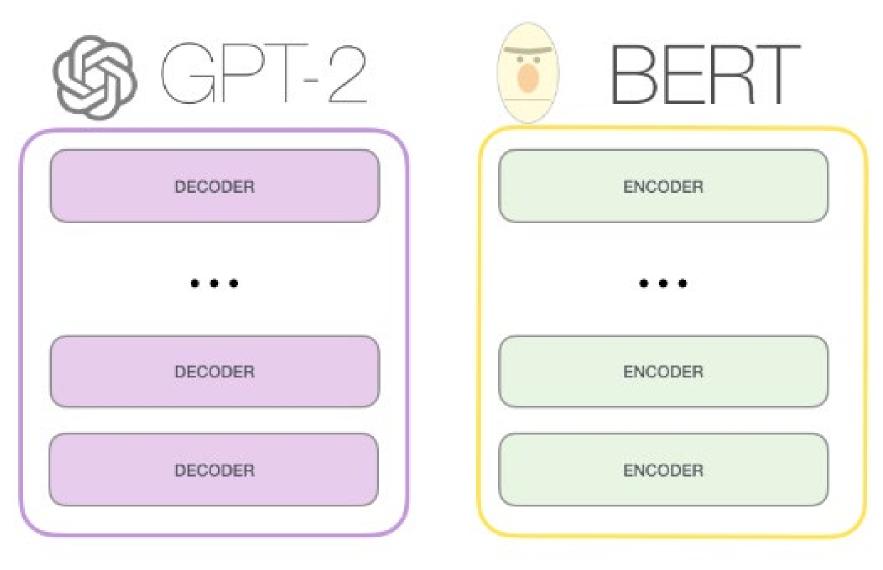


BioBERT, Covid-Twitter-BERT, etc

Using these *BERTs

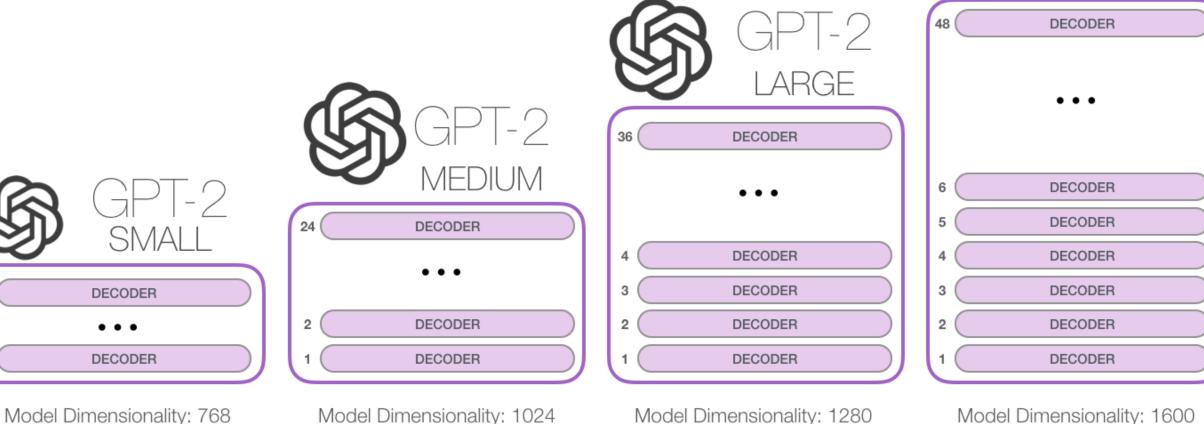
- Pretrain \rightarrow finetune
 - Pretrain encoders on pretraining tasks (high-resource/data, possibly unsupervised)
 - Finetune encoders on target task (low-resource, expensive annotation)
- Primary method of evaluation: Natural Language "Understanding" (NLU)
 - Question Answering and Reading Comprehension
 - Commonsense
 - Textual Entailments

GPT-2 - BERT



GPT released June 2018 GPT-2 released Nov. 2019 with 1.5B parameters GPT-3 released in 2020 with 175B parameters





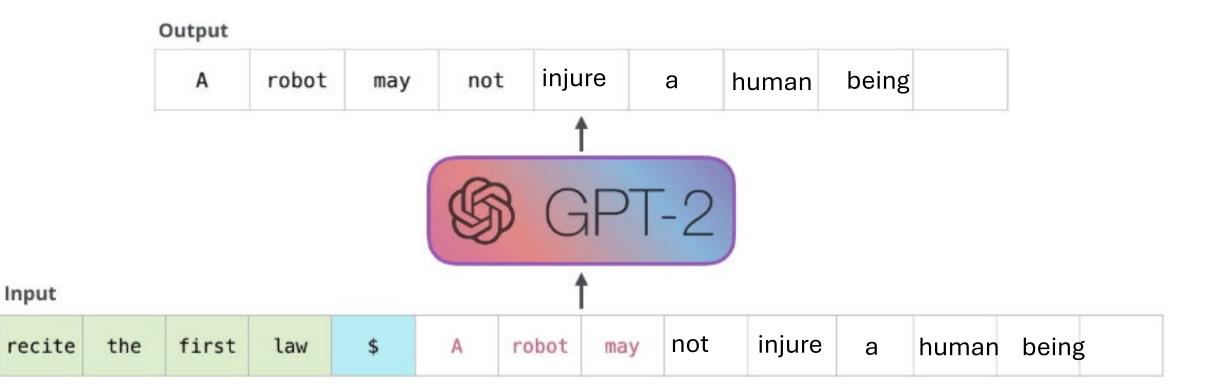
117M parameters

345M

762M



GPT-2 in action



Byte Pair Encoding (BPE)

Word embedding sometimes is too high level, pure character embedding too low level. For example, if we have learned

- old older oldest
- We might also wish the computer to infer
 - smart smarter smartest

But at the whole word level, this might not be so direct. Thus the idea is to break the words up into pieces like er, est, and embed frequent fragments of words.

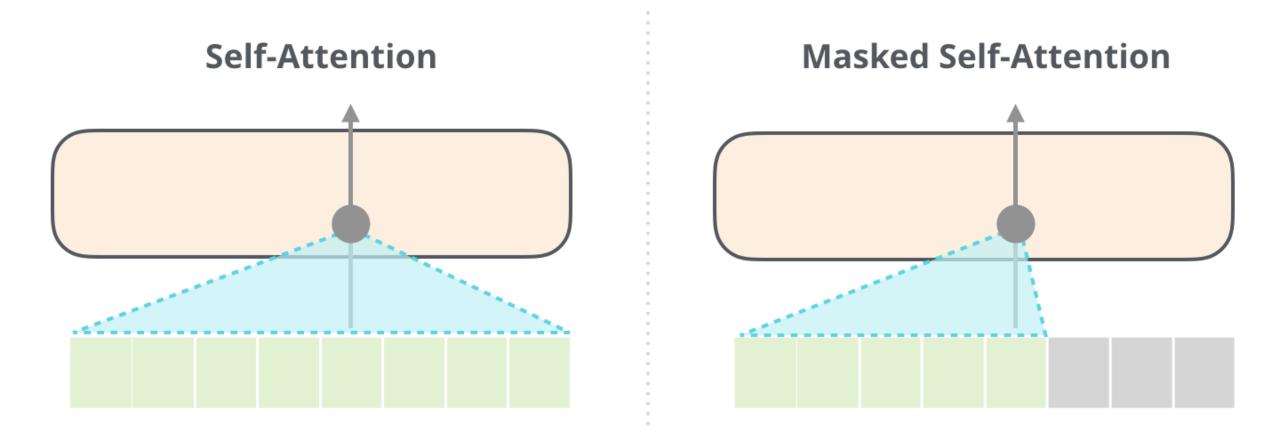
GPT adapts this BPE scheme.

Byte Pair Encoding (BPE)

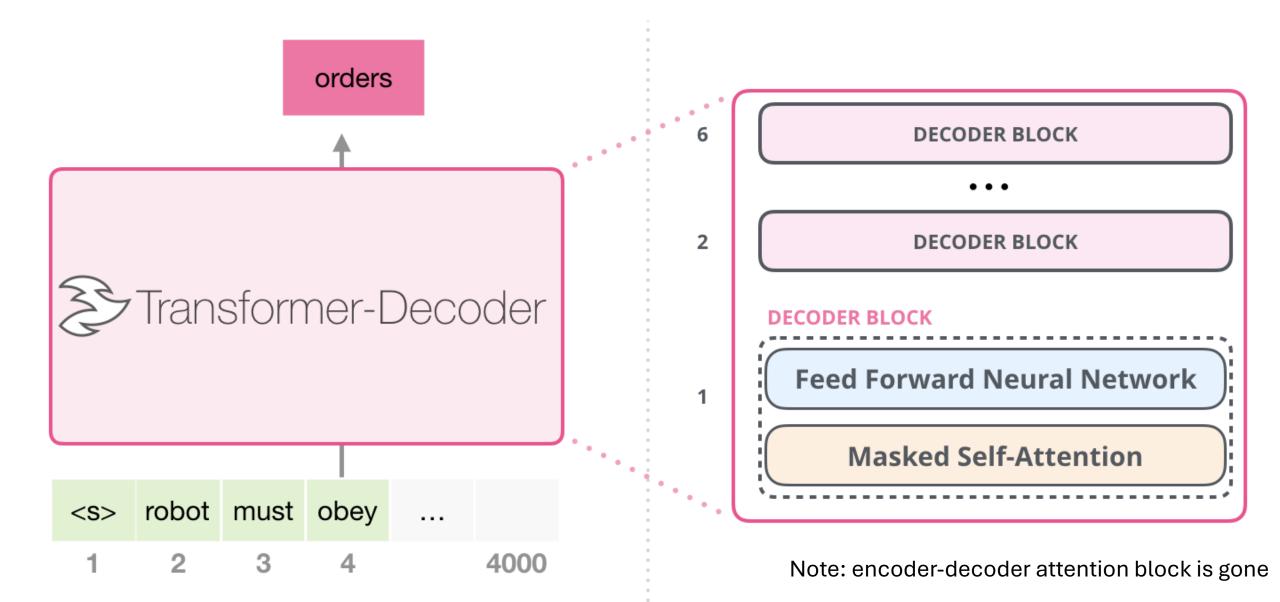
GPT uses BPE scheme. The subwords are calculated by:

- 1. Split word to sequence of characters (add </w> char)
- 2. Joining the highest frequency pattern.
- 3. Keep doing step 2, until it hits the pre-defined maximum number of subwords or iterations.

Example (5, 2, 6, 3 are number of occurrences) {'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w e s t </w>': 6, 'w i d e s t </w>': 3 } {'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w es t </w>': 6, 'w i d es t </w>': 3 } {'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w est </w>': 6, 'w i d est </w>': 3 } (lost freq. 9) {'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w est</w>': 6, 'w i d est</w>': 3 } (lo freq 7) Masked Self-Attention (to compute more efficiently)



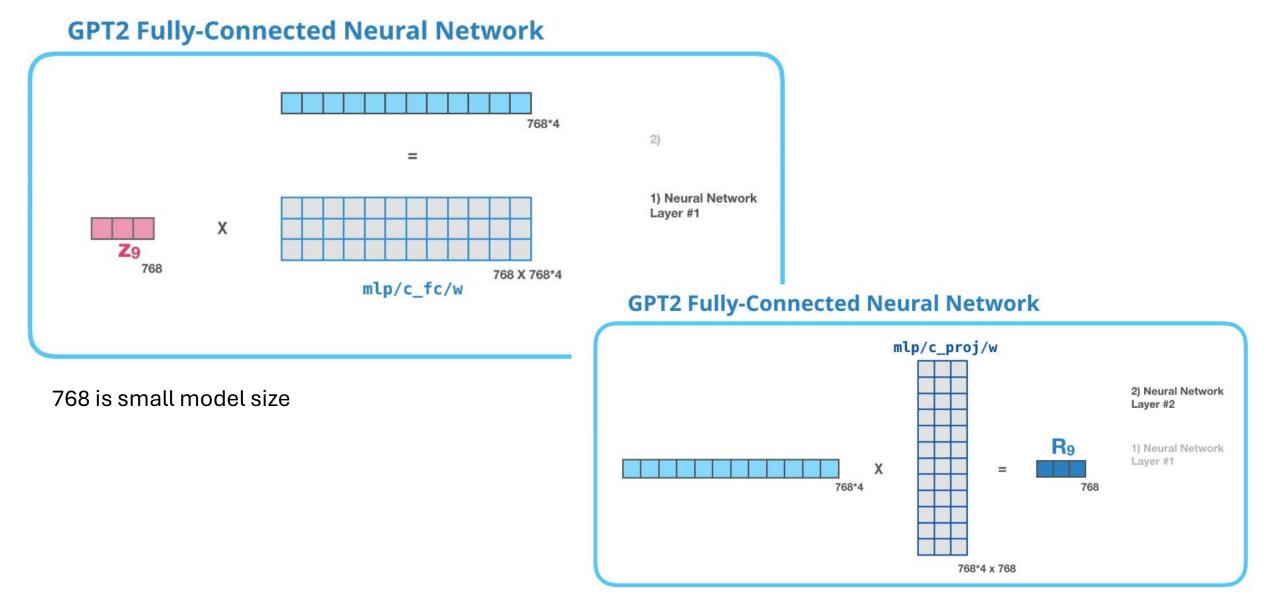
Masked Self-Attention



Masked Self-Attention Calculation

Re-use previous computation results: at any step, only need to results of q, k, v related to the new output word, no need to re-compute the others. Additional computation is linear, instead of quadratic.

GPT-2 fully connected network has two layers (Example for GPT-2 small)



GPT-2 has a parameter top-k, so that we sample words from top k (highest probability from SoftMax) words for each output

	Decoder #12, Position #1 output vector
	DECODER
	Decoder #2, Position #1 output vector
	DECODER
DECODER	Decoder #1, Position #1 output vector
	Feed Forward Neural Network
	 Masked Self-Attention

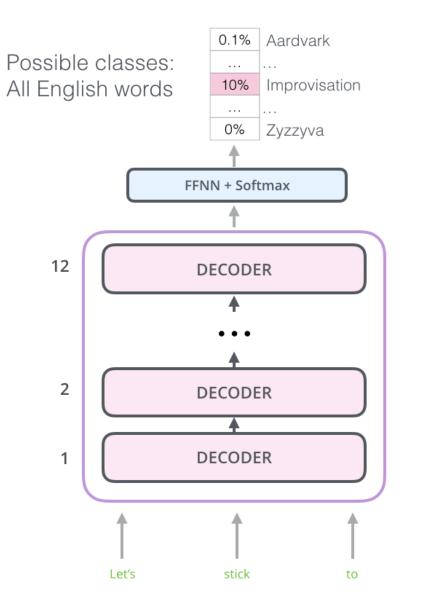


GPT Training

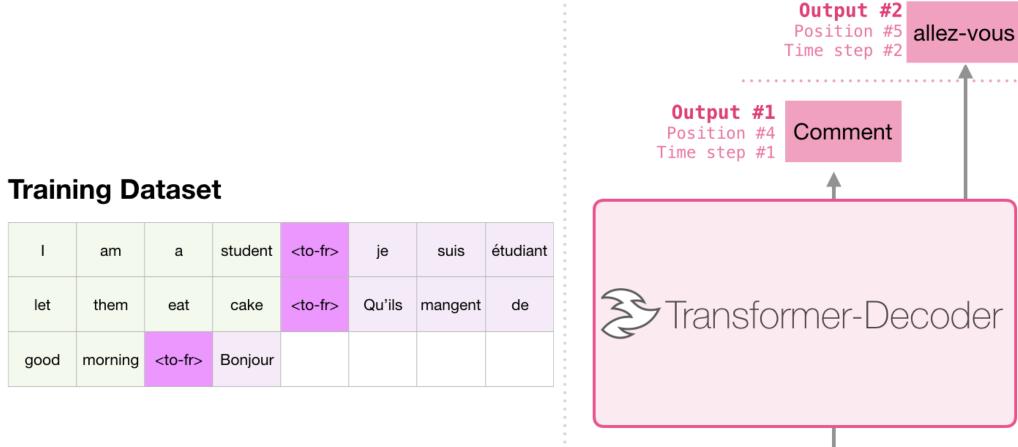
GPT-2 uses unsupervised learning approach to training the language model.

There is no custom training for **GPT-2**, no separation of pre-training and fine-tuning like **BERT**.

Transformer / GPT prediction



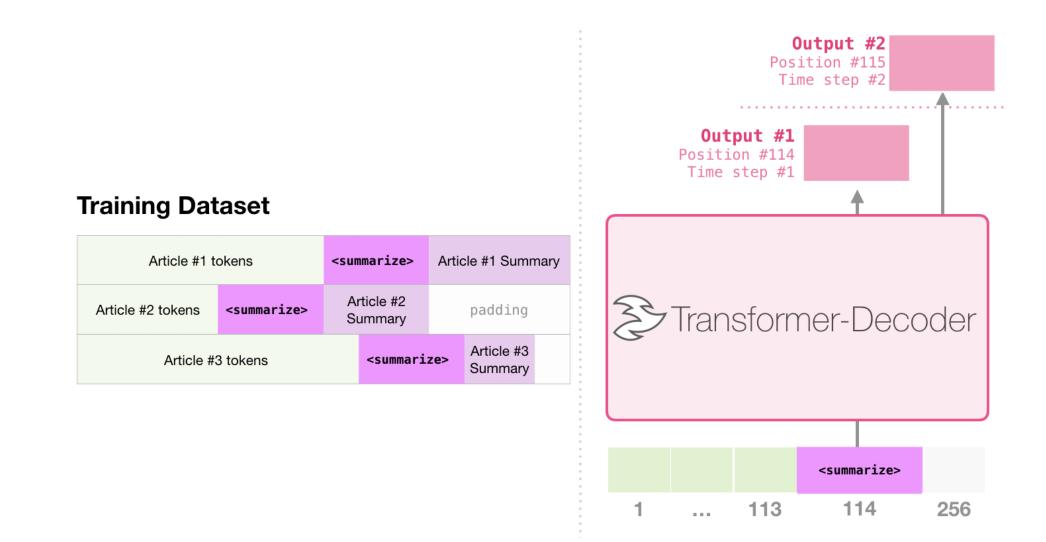
GPT-2 Application: Translation



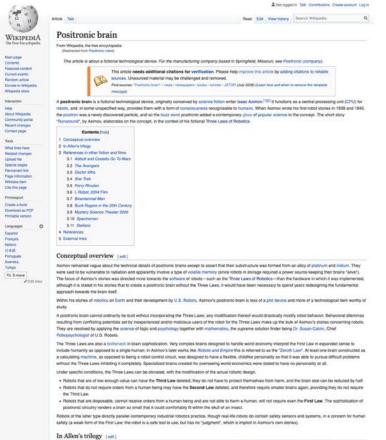
how	are	you	<to-fr></to-fr>	
1	2	3	4	1024

.

GPT-2 Application: Summarization



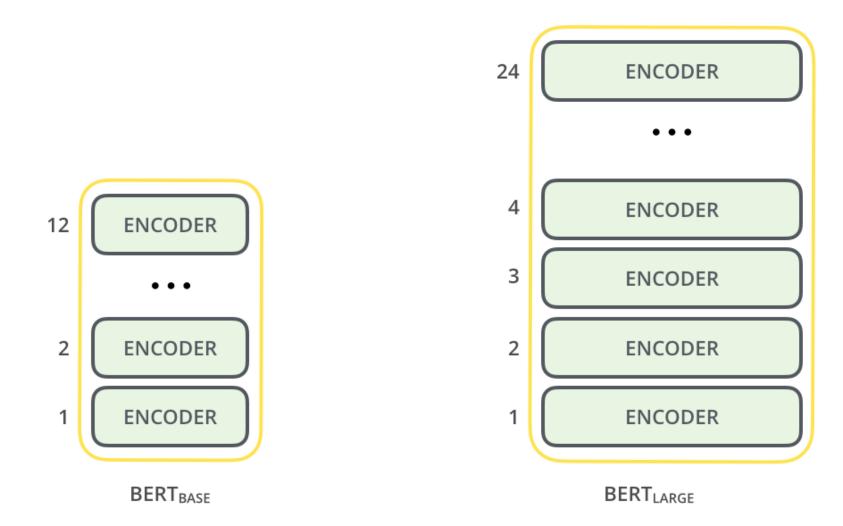
Using Wikipedia data



Several indext staries have been written by other authors following Asimon's death. For example, in Roger MacBride Alern's Catilian trilogy, a Spacer indextes that Gaber Antalian invents the granitable brain. It dens speed and capacity projectionets over traditional positive designs, but the strong imburse of handling traditional positive catilians and, College in endoted, Infordate Unity, chooses ta adopt granitence, beauties of them is the optimise the advance of the strong imburse of tradition traditional positive designs, but the strong imburse of the strong imburse beauties and endoties the strong imburse beauties of the strong imburse beauties and endoties in the strong imburse beauties of the strong imburse beauties and endoties in the strong imburse beauties of the strong imburse beauties and endoties in the strong imburse beauties of the strong imburse beauties and endoties in the strong imburse beauties to the trans and antiburse in the strong imburse beauties that in the strong imburse beauties of the strong imburse beauties of the strong imburse beauties and antiburse in the strong imburse beauties of the strong imburse beauties that beauties the strong imburse beauties that in the strong imburse beauties that in the strong imburse beauties that the strong imburse beauties the strong imburse that the strong imburse beauties that the strong imburse beauties the strong imburse the strong im

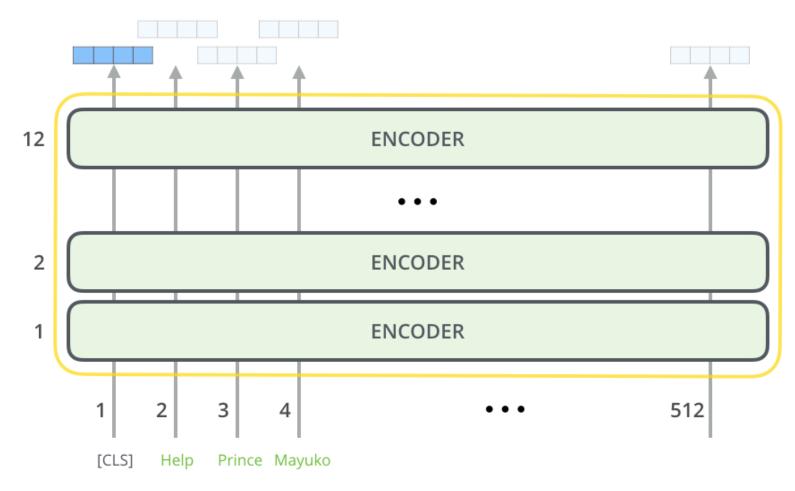
Positronic brain	
From Wikipedia, the free encyclopedia (Redirected from PostPonic robot)	
This article is about a fictional technological device.	For the manufacturing company based in Springfield, Missouri, see Positronic (company).
	nal citations for verification. Please help improve this article by adding citations to reliable
	ial may be challenged and removed.
Find sources: 'Positionic brain message)	- news - newspapers - books - scholar - JETOR (July 2008) (Learn how and when to remove this template
A positronic brain is a fictorial technological device, or robots, and, in some unspecified way, provides them wi the positron was a newly discovered particle, and so the	graphy conserved by source factor anter laser Assess ¹⁰⁰ Is functions as a control processing unit (CPU) for a 1 CI INANAA OV When Assess wrote the first robot stores in 1939 and 1949
"Funanciant", by Asimov, elaborates on the concept, in	he context of his fictional Three Laws of Pocotics.
Contents (hor)	
1 Conceptual overview 2 In Alleria Misory	
3 References in other fiction and films	
9.1 Abbett and Cestello Ge To Mars	
3.2 The Avengers 3.3 Dector Whe	
3.4 Star Trek	
3.5 Perry Rhodan	
3.6 J. Robot, 2004 Film 3.7 Dicentennial Man	
3.8 Buck Ropers in the 25th Century	
3.9 Mystery Science Theater 3000 3.10 Spectroman	
3.11 Stellaris	
4 References	
5 External links	
Conceptual overview [mit]	
	whome brains except to assert that their substructure was formed from an alloy of plantum and aldum. They verve a type of versiliar manney sance mbots in storage required a power source keeping their brains "alwa").
	a the software of robots - such as the Three Laws of Robotics - than the hardware in which it was implemented
	on beam without this. These Laws, it would have been necessary to spend years redesigning the fundamental
approach forwards the brain taelt.	
shutter for allowing of rootable on party and their develops shutter.	nerr by U.S. Hotots, Aamon's polymeric brain is less of a plot device and more of a technological item worthy
A positionic train carries understry be built without their	pointing the Three Laws, any motification invited would displicatly modify robot behavior. Behavioral elemente
resulting from conflicting potentials set by inexperiences	and/ter mainistus users of the robot for the Three Laws make up the bulk of Astmon's states concerning robots
They are resolved by applying the asteries of logic and a histocenychologist of U.S. Robots	eachology together with mathematics, the economic solution (index being Dr. Susan Calver, Chief
	mon. Wery controllery braining designed to handle wante economic interpret the Fund Law is conservable active to
include numerity as opposed to a single human; in Allin	ition. Wery complex brains designed to handle world economy interpret the Firld Law in sepanded sense to consister works like Rodoth and Empire this is referred to as the "Zamth Law". At least one brain constructed a
include numerity as opposed to a single human; in Allin	over later works like Robots and Empire this is referred to as the "Zamith Law". At asset one brain constructed a
Include humanity as opposed to a single human; in Adm a saturating machine, as opposed to being a robot con without fire Three Laws inhibiting is completely. Openant	or the end of the finance and finance for a second se
Include numerity as opposed to a single numerity to Astr a calculating mathine, as opposed to bring a robot con- without the Three Laws inhibiting it comprisity. Appendix Under specific conditions, the Three Laws can be obvio	ARTICLE
Include humanity iss opposed to a single humani; in Allo a saturating machine, as opposed to bring a noted core whold its Three Laree inholding it completely. Openial Under specific conditions, the Three Lares table to store • Reboti has an of two enough value can have the T	or the end of the finance and finance for a second se
Include insummity as opposed to a single human, is has a satisfiability mathem, as opposed to being a solution con- sistent of the time of how comparing a solution con- sistent and the solution of the solution of the solution - Redokt that are of how conceptive using the token - Redokt that are of how conceptive using the token are Third Law.	Contract of the second seco
Include homestry as opposed to a engine homes in Advance a sensitivity emitters, as a possible of basis, a reflect con- without fire Three Caree children's competing deposition Under section conditions, the Three Caree can be onlywed. • Reduct that and only an engine streters from a furname oper mit Three Laree. • Reduct that are all reacy and the three three Three mit Three Laree.	on taken were the fiber band hard frame their indexed as a the "dom't law". A least can be an experimentation of the standard of the standard law
Nucleah newarity is apposable as a engine herman. In Adv a destinating empirica, in a sposable able and a strated com- without in the Tweek Laws inhibiting transportant, disposati ables sposition contained, the "Threek Laws to be strate in Holdons that all of level encodes intern a turner to the "Holdons that all of level encodes intern a turner in the "Holdons that all of level encodes intern a turner in the "Holdons that all of level encodes intern a turner in the participation that all an encodes using turners in the second participation of the second second second models and that is to participation of the second second second models and that is to participation of the second second second second models and that is to participation of the second second second second models and that is to participation of the second second second second models and that is to participation of the second second second second models and that is to participation of the second second second second models and that is to appending complexity and the second second second models and that is to appending complexity and the second secon	The same is the fiber and frame the indexed at a strend "same". At least case have meanwhile at the same and the same is the same the indexed at a strend "same". At least case have meanwhile at the same and the same same is the same is the same and the same same is the same is the same and the same is the same is the same i
Include Innersity as opposited as a engine terms in 14 and a data Mark membra with a cooperation and any index con- writeduct free Tarves Laws inheriting it compares to the data there is a data and the second posterioric different previous a lower so would then 5 no become of the selecting devices and compares on the second and compared compares and compares on the second the second and the se	on taken were the fiber band hard frame their indexed as a the "dom't law". A least can be an experimentation of the standard of the standard law
Inclusion homewrity as opgosof to a engrin homewrit A had a claniadhr generity as a topposof to any a robot con- writoud frei Three Laws inheliting 't companie, figurial Unier seeking conditions, the Three Laws Law be laws - Riboth that are to encologit ware as there the T - Riboth that are information, commit means on their any Three Laws - Riboth that are information, commit means on their partners due to the partner address a commit means on their high back on the allow information, commit means on their high back on the allow information communication safety, caused barries of the Their Laws the robot is a safe	on later was the fiber and fiber that indexed at a the "death late". At least case that presented at its indexed by the second
Industry memory is appointed to a entry's herman. In Adv a similarly employee, in appointed to simple, in their com- when of the Theve Laws inhibiting is compared, flogolati- thether second constraints. The There Laws is the similar entry of the simple is and the entry of value and is the simple different law. In the entry of the simple is and the simple different law. The simple methods from the time simple that are allowed and the simple is a simple of the There Laws. The simple distribution is the simple partnerse, checking reaches, a benefit to a simple method law of the simple distribution is a simple simple of the simple reaction of the time of the simple distribution is a simple law of the simple simple distribution is a simple simple of the simple simple simple distribution is a simple simple in Allenn's trillogy. (simple)	our later even the fiber and figure the indexed to a the "dentities". At least one have meanwhile at it can be appreciately the second second of the CARCENCE : we want second s
Include humanity as opposed to a engine human is had a solution examines a cooperate to the solution of the humanity of the solution is a cooperate to the solution of the solution occurring the comparise, disposite the solution occurring the solution is the solution of the solution occurring the solution is not a format the solution occurring the solution is a solution for a solution occurring the solution is a solution for a format the solution occurring the solution is a solution for a format solution to be because written by other authors the format authors have because written by other authors of the format authors have because written by other authors of the format authors have because written by other authors of the format authors have because written by other authors observed resolution of the solution occurring the solution occurring the solution and the solution occurring the generation occurring the solution occurring the authors and the solution occurring the generation occurring the authors and the solution occurring the generation occurring the solution occurring the authors occurring the solution occurring the solution occurring the authors and the solution occurring the generation occurring the authors and the solution occurring the solution occurring the authors a	on-later waters the fiberbarran fibers that indexide ta a the "dead task. At least one have meanerships of the start of the start of t
evides humanly as opposed to a engine terms in A shall estimated results, a cooperate to long a school coordinate of the Three Laws inheling it compares, foreign the appendix conditions. The Three Laws can be devide the shade the condition with the Three Laws are been only a solution to the three the terms of the monophy wave marks the three three the terms of the monophy wave marks the term three terms that an or the monophy wave marks the term three terms that an orthogon and the term terms to the term three terms that any three terms that any three terms that any three terms three terms that any three terms three terms that any three terms terms to terms the term terms to terms three terms terms to terms the term terms to terms three terms terms to term terms terms three terms terms terms terms terms to term terms term	on later rest life floods and fraver that indexed to a string "Jambi Later. At later can be are presented at in the string of the string of the string that indexed to a string "Jambi Later. At later can be are presented at an use of the string of the string that the string of the

BERT (Bidirectional Encoder Representation from Transformers)



Model input dimension 512

Input and output vector size



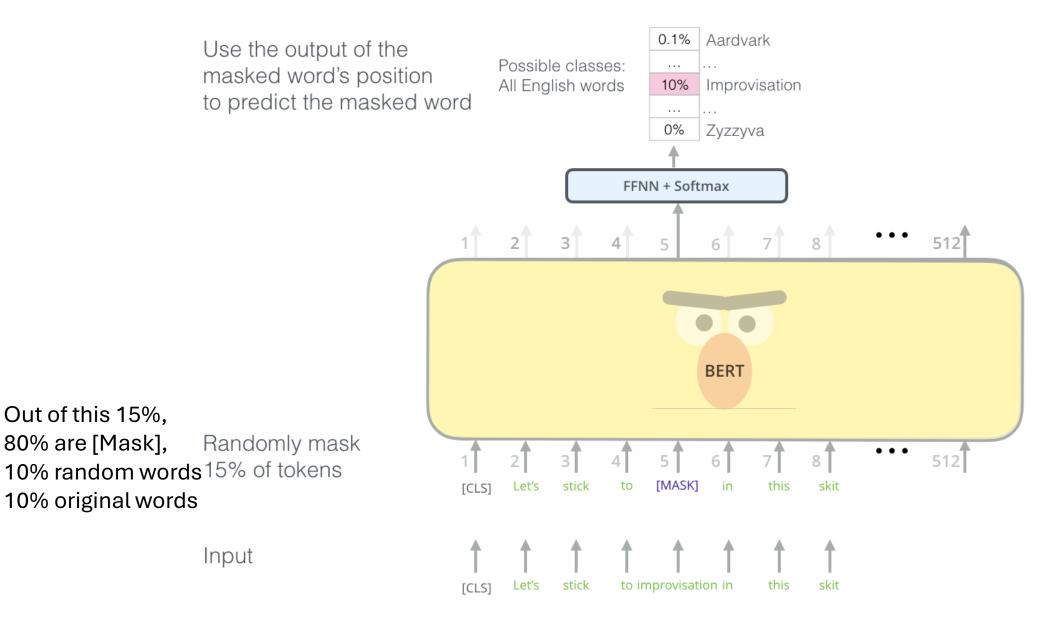
BERT

BERT pretraining

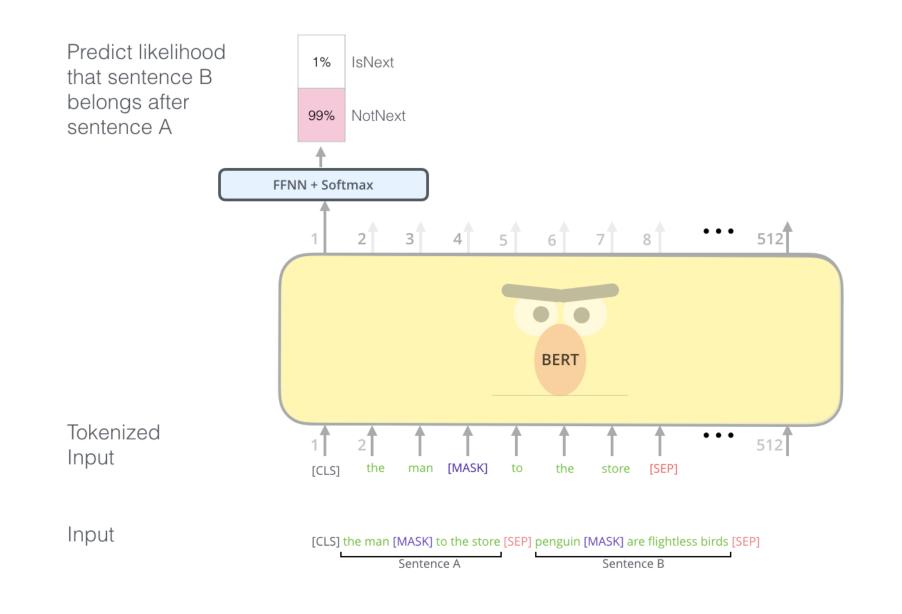
- ULM-FiT (2018): Pre-training ideas, transfer learning in NLP. ELMo: Bidirectional training (LSTM)
- **Transformer:** Although used things from left, but still missing from the right.
- GPT: Use Transformer Decoder half.

BERT: Switches from Decoder to Encoder, so that it can use both sides in training and invented corresponding training tasks: masked language model

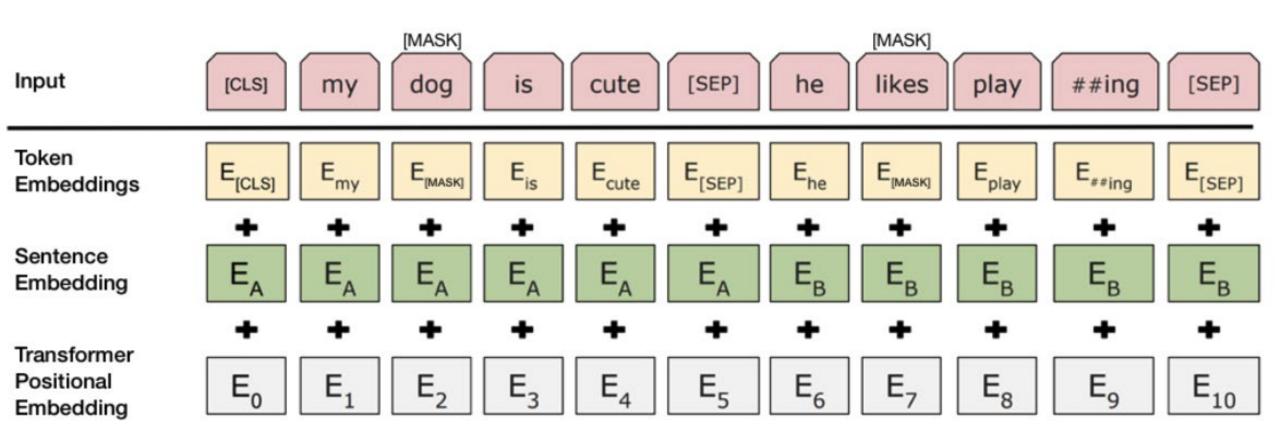
BERT Pretraining Task 1: masked words



BERT Pretraining Task 2: two sentences



BERT Pretraining Task 2: two sentences

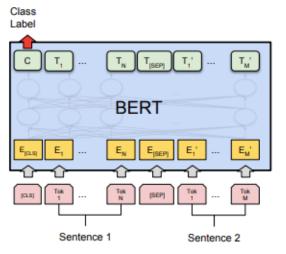


50% true second sentences 50% random second sentences

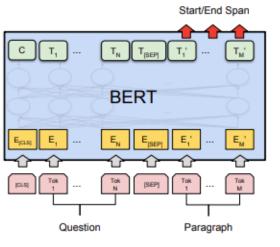
Fine-tuning BERT for other specific tasks

MNLI

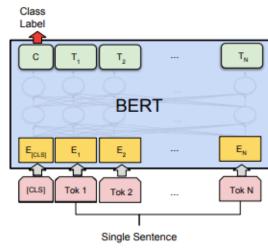
- QQP (Quaro Question Pairs) Semantic equivalence) QNLI (NL inference dataset) STS-B (texture similarity) MRPC (paraphrase, Microsoft) RTE (textual entailment) SWAG (commonsense inference) SST-2 (sentiment) CoLA (linguistic acceptability
- SQuAD (question and answer)



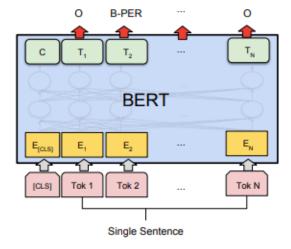
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(c) Question Answering Tasks: SQuAD v1.1



(b) Single Sentence Classification Tasks: SST-2, CoLA



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER SST (Stanford sentiment treebank): 215k phrases with finegrained sentiment labels in the parse trees of 11k sentences.

NLP Tasks: Multi-Genre Natural Lang. Inference

MNLI: 433k pairs of examples,	Met my first girlfriend that way.	FACE-TO-FACE contradiction	I didn't meet my first girlfriend until later.
	8 million in relief in the form of emergency housing.	C C N C Government neutral N N N N	The 8 million dollars for emergency hous- ing was still not enough to solve the prob- lem.
labeled by entailment, neutral or contraction	Now, as children tend their gardens, they have a new ap- preciation of their relationship to the land, their cultural heritage, and their community.	LETTERS neutral N N N N	All of the children love working in their gardens.
	At 8:34, the Boston Center controller received a third transmission from American 11	9/11 entailment E E E E	The Boston Center controller got a third transmission from American 11.
	I am a lacto-vegetarian.	Slate neutral n n e n	I enjoy eating cheese too much to abstain from dairy.
	someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	Telephone contradiction c c c c c	No one noticed and it wasn't funny at all.

Table 1: Randomly chosen examples from the development set of our new corpus, shown with their genre labels, their selected gold labels, and the validation labels (abbreviated E, N, C) assigned by individual annotators.

NLP Tasks (SQuAD -- Stanford Question Answering Dataset):

Sample: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24– 10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Which NFL team represented the AFC at Super Bowl 50?

Ground Truth Answers: Denver Broncos

Which NFL team represented the NFC at Super Bowl 50?

Ground Truth Answers: Carolina Panthers

Add indices for sentences and paragraphs

SegaTron/SegaBERT



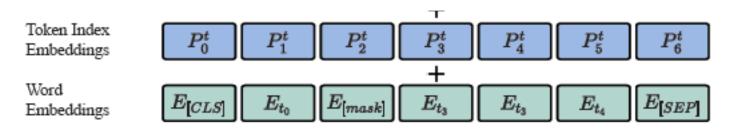


Figure 1: Input Representation of SegaBERT

H. Bai, S. Peng, J. Lin, L. Tan, K. Xiong, W. Gao, M. Li: SgaTron: Segment-aware transformer for language modeling and understanding. AAAI'2021

Conversion speed much faster:

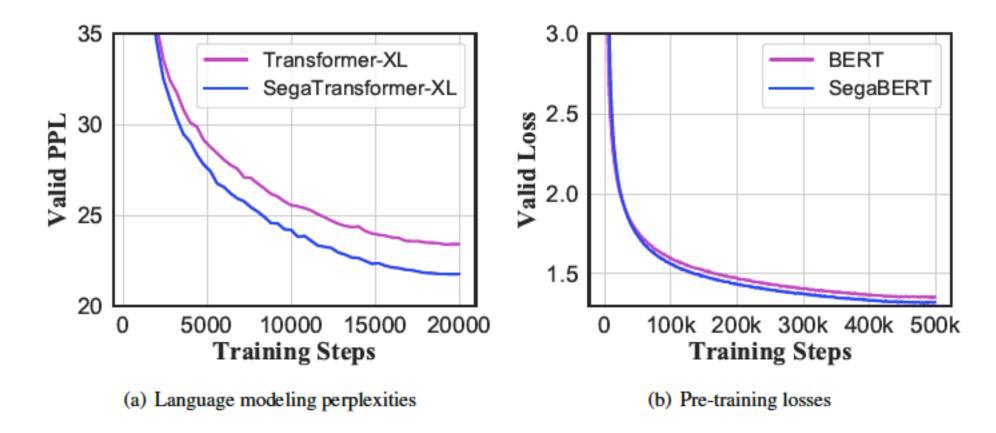


Figure 2: Valid perplexities and losses during the training processes of language modeling and pre-training.

Testing on GLUE dataset

Task(Metrics)	BASE model(wikipedia 500K steps)			LARGE model(wikibooks 1000K steps)				
	dev		test		dev		test	
	BERT	SegaBERT	BERT	SegaBERT	BERT	SegaBERT	BERT	SegaBERT
CoLA (Matthew Corr.)	55.0	54.7	43.5	50.7	60.6	65.3	60.5	62.6
SST-2 (Acc.)	91.3	92.1	91.2	91.5	93.2	94.7	94.9	94.8
MRPC (F1)	92.6	92.4	88.9	89.3	-	92.3	89.3	89.7
STS-B (Spearman Corr.)	88.9	89.0	83.9	84.6	-	90.3	86.5	88.6
QQP (F1)	86.5	87.0	70.8	71.4	-	89.1	72.1	72.5
MNLI-m (Acc.)	83.2	83.8	82.9	83.5	86.6	87.6	86.7	87.9
MNLI-mm (Acc.)	83.4	84.1	82.8	83.2	-	87.5	85.9	87.7
QNLI (Acc.)	90.4	91.5	90.1	90.8	92.3	93.6	92.7	94.0
RTE (Acc.)	68.3	71.8	65.4	68.1	70.4	78.3	70.1	71.6
Average	82.2	82.9	77.7	79.2	-	86.5	82.1	83.3

Table 2: The results on GLUE benchmark. All base models are pre-trained by this work. Every result of the dev set is the average score of 4 times finetuning with different random seeds. Scores of BERT large dev are from (Sun et al., 2019) and scores of BERT large test are from (Devlin et al., 2018).

H. Bai, S. Peng, J. Lin, L. Tan, K. Xiong, W. Gao, M. Li: SgaTron: Segment-aware transformer for language modeling and understanding. AAAI'2021

Reading comprehension – SQUAD tasks

System	Dev			
e jetetit	EM	F1		
BERT base (Single)	80.8	88.5		
BERT large (Single	84.1	90.9		
BERT large (Single+DA)	84.2	91.1		
KT-NET	85.2	91.7		
StructBERT Large (Single)	85.2	92.0		
SegaBERT base (Single)	83.2	90.2		
SegaBERT large (Single)	85.3	92.4		

Table 3: Evaluation results of SQUAD v1.1.

System	Dev			
o j sterii	D EM 72.3 75.4 76.3 78.7 80.6 81.8	F1		
BERT base	72.3	75.6		
BERT base (ours)	75.4	78.2		
SegaBERT base	76.3	79.2		
BERT large	78.7	81.9		
BERT large wwm	80.6	83.4		
SegaBERT large	81.8	85.2		

Table 4: Evaluation results of SQUAD v2.0.

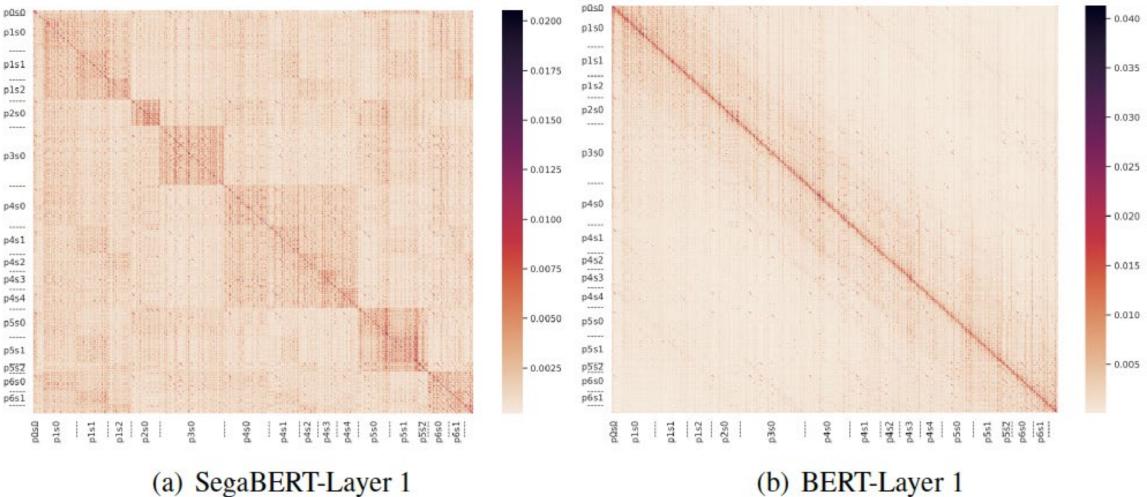
F1 = 2 (P*R) / (P+R), P is precision, R is recall, all in percentage, EM – exact match

Improving Transformer-XL

Model	#Param.	PPL
LSTM+Neural cache (Grave et al., 2017)	-	40.8
Hebbian+Cache (Rae et al., 2018)	-	29.9
Transformer-XL base, M=150 (Dai et al., 2019)	151M	24.0
Transformer-XL base, M=150 (ours)	151M	24.4
SegaTransformer-XL base, M=150	151M	22.5
Adaptive Input (Baevski and Auli, 2019)	247M	18.7
Transformer-XL large, M=384 (Dai et al., 2019)	257M	18.3
Compressive Transformer, M=1024 (Rae et al., 2020)	257M	17.1
SegaTransformer-XL large, M=384	257M	17.1

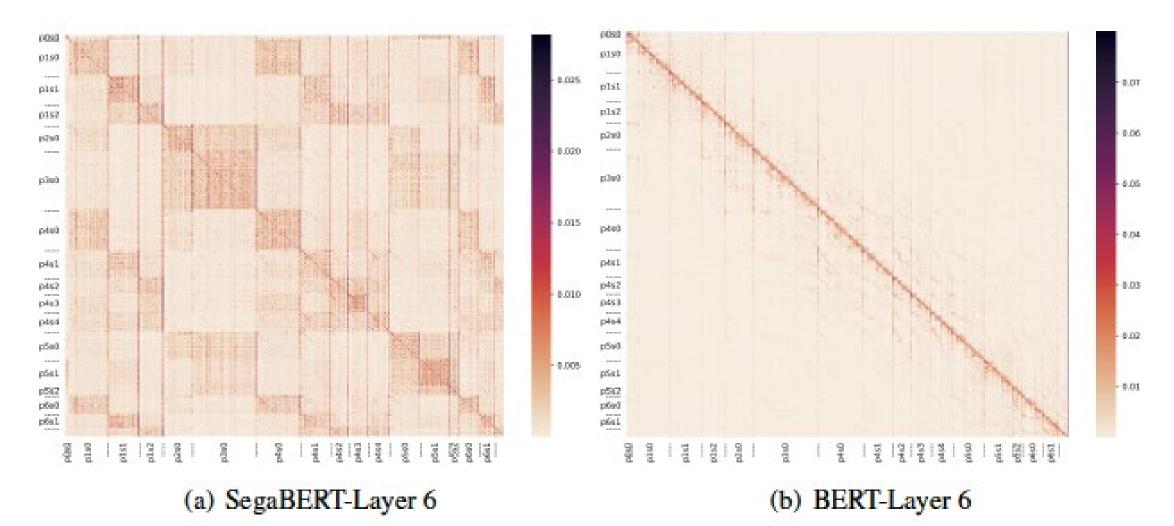
Table 1: Comparison with Transformer-XL and competitive baseline results on WikiText-103.

Looking at Attention

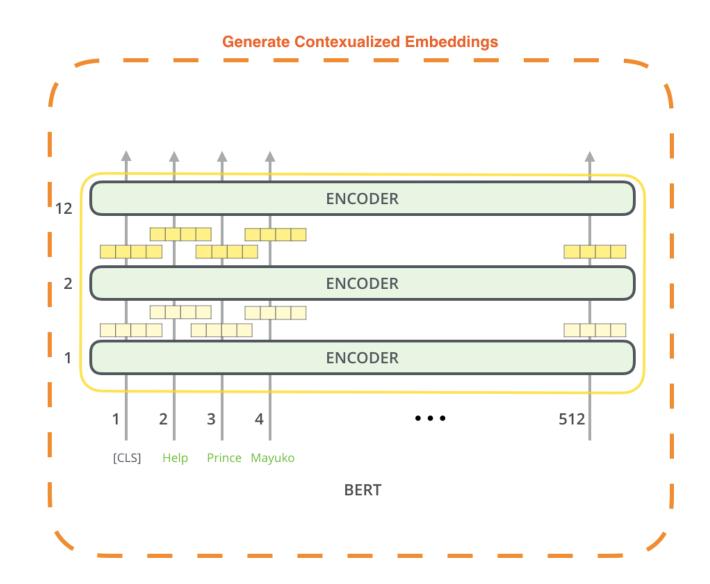


(a) SegaBERT-Layer 1

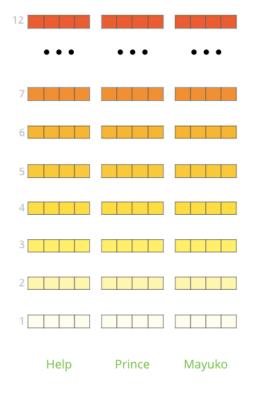
Looking at Attention



Feature Extraction



The output of each encoder layer along each token's path can be used as a feature representing that token.



We end up with some embedding for each word related to current input

We start with independent word embedding at first level

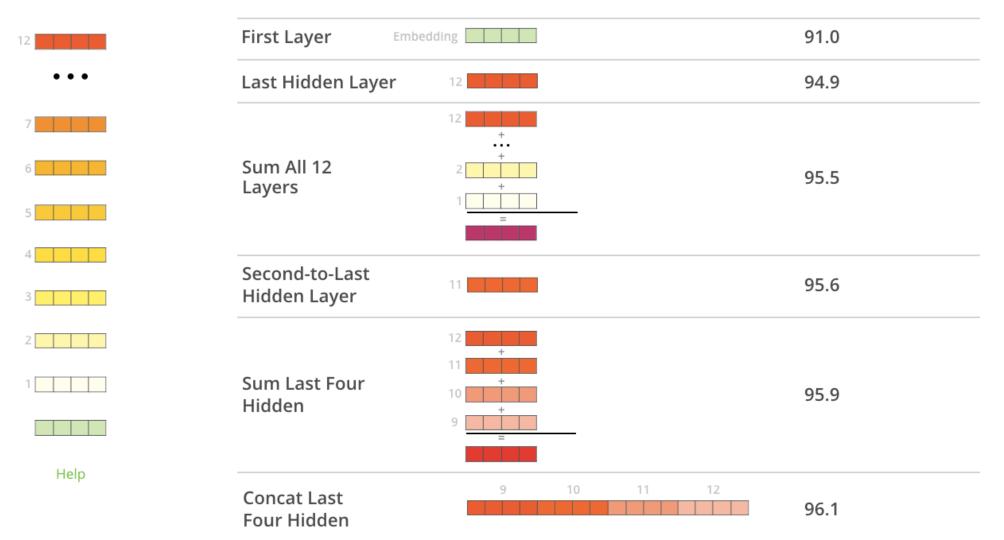
But which one should we use?

Feature Extraction, which embedding to use?

What is the best contextualized embedding for "Help" in that context?

For named-entity recognition task CoNLL-2003 NER

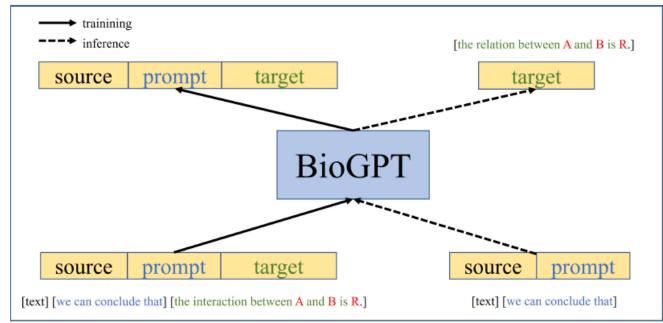
Dev F1 Score





BioGPT was fine-tuned and evaluated on several downstream tasks, including NER, QA, relation extraction, and document classification.

- Hard prompts: Hard prompts are manually designed discrete language phrases or templates that are prepended to the input text to guide the language model towards a specific task.
- Soft prompts, on the other hand, are continuous embeddings learned during the fine-tuning process.



"We have that [head entity] [relation] [tail entity]," "In conclusion, [head entity] [relation] [tail entity]," and "We can conclude that [head entity] [relation] [tail entity]."

BioBERT

- First, BioBERT is **intialized with weights from <u>BERT</u>**, which was pretrained on general domain corpora (English Wikipedia and BooksCorpus).
- Then, BioBERT is **pre-trained on biomedical domain corpora** (PubMed abstracts and PMC full-text articles).
- Finally, BioBERT is **fine-tuned** and evaluated on three popular biomedical text mining tasks (NER, RE and QA).

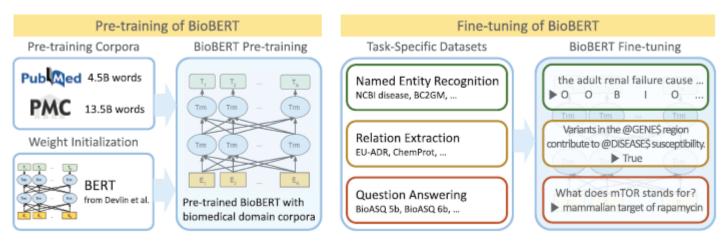


Fig. 1. Overview of the pre-training and fine-tuning of BioBERT

ClinicalBERT

The model is pre-trained on a large corpus of clinical notes from the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, which contains de-identified electronic health records of patients admitted to the intensive care unit.

The main goal of ClinicalBERT is to improve the prediction of hospital readmission within 30 days based on the information contained in clinical notes.

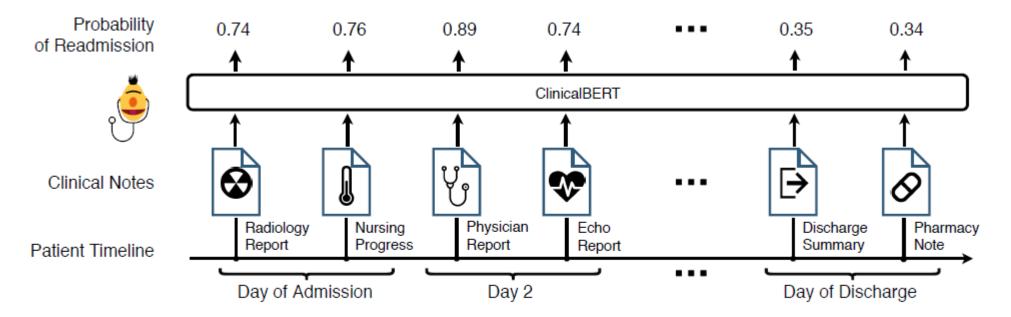


Figure 1: ClinicalBERT learns deep representations of clinical notes that are useful for tasks such as readmission prediction. In this example, care providers add notes to an electronic health record during a patient's admission, and the model dynamically updates the patient's risk of being readmitted within a 30-day window.

Thank you!

Literature & Resources for Transformers

Resources:

OpenAI GPT-2 implementation: <u>https://github.com/openai/gpt-2</u> BERT paper: J. Devlin et al, BERT, pretraining of deep bidirectional transformers for language understanding. Oct. 2018. ELMo paper: M. Peters, et al, Deep contextualized word representation, 2018 ULM-FiT paper: Universal language model fine-tuning for text classification. J. Howeard, S. Ruder., 2018 Jay Alammar, The illustrated GPT-2, <u>https://jalammar.github.io/illustrated-</u>

<u>gpt2/</u>