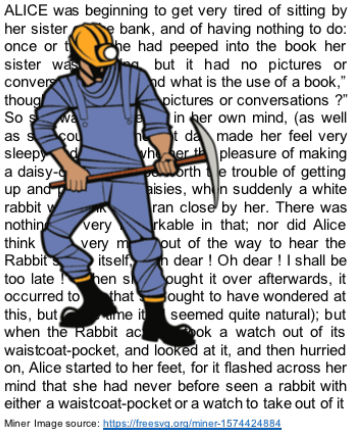


Miner Image source: https://freesvg.org/miner-1574424884

POLITECNICO
MILANO 1863

DIPARTIMENTO DI ELETTRONICA,
INFORMAZIONE E BIOINGEGNERIA

20.10.2021

Invited Lecture:
Deep Neural Techniques for Text Mining

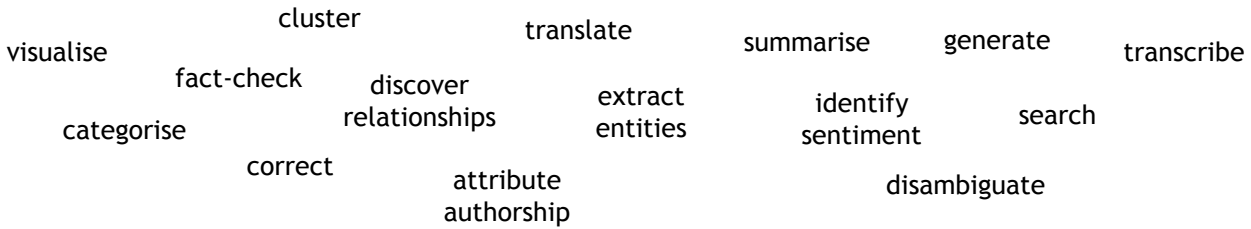
Mark Carman

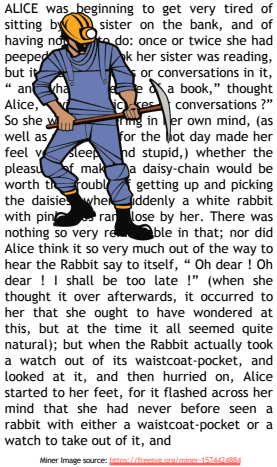
What is this talk about?

Mark Carman
20.10.2021

Text Mining

- process of **extracting useful knowledge** from **text data**
- *lots* of different things we can do with text data:




ALICE was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, " and what is the use of a book," thought Alice, " without pictures or conversations ?" So she was considering in her own mind, (as well as she could, for the hot day made her feel very sleepy and stupid,) whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a white rabbit with pink eyes ran close by her. There was nothing so very remarkable in that; nor did Alice think it so very much out of the way to hear the Rabbit say to itself, " Oh dear ! Oh dear ! I shall be too late !" (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually took a watch out of its waistcoat-pocket, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket or a watch to take out of it, and

Miner Image source: https://freerange.org/miner-1574424884

with **Deep Learning**

- machine learning models using deep neural networks
- which have revolutionised performance on ***all these tasks*** over last few years

Welcome

Who am I?

- Mark Carman, Politecnico di Milano
- Background:
  - Information Retrieval & statistical Natural Language Processing**
  - Machine Learning & Data Science
- Applications:
  - *Personalisation & Recommendation, Web Search, Social Media Analysis, Digital Forensics, Bioinformatics, …*
- Teaching:
  - Data Science and Artificial Intelligence
  - classes are more fun when there's interaction, so **help me out** by asking lots of questions!

🤔 Let's hope his teaching is better than his cooking …😂🤣

** Favourite NLP quote:
"Every time I fire a linguist, the performance of the speech recognizer goes up" Frederick Jelinek

a bit on my text-related research …

Mark Carman
20.10.2021

Here are some research problems we're working on:

- **Text Analysis for Bioinformatics**
  - translating text into structured data to facilitate search
  - model protein sequences for drug repurposing

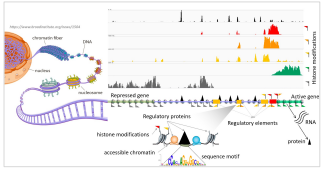
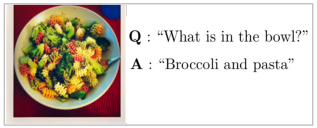


Image source: https://www.synapse.org/#!Portal:html/Synapse:syn17083707/wiki/588600

- **Text Analysis for Political Discourse**
  - identifying claims & finding evidence online
  - detecting fake and misleading news

Image source
https://www.npr.org/2018/04/11/601323233/6-facts-we-know-about-fake-news-in-the-2016-election

- **Text meets Images/Video**
  - visual question answering
  - explaining AI with text explanations



Want to try some **demos**?
Go to: http://131.175.120.138:61111/

So you know who I am …

Who are you?

- What sectors do you work in?
- How familiar are you with Machine Learning?
- What about text processing?
- And Deep Learning?

CC, source: https://commons.wikimedia.org/wiki/File:5th_Floor_Lecture_Hall_(cartoon).jpg

Text generation has become so good …

Mark Carman
20.10.2021

So good in fact, that I considered using it to generate a recent talk … 🤣😂

Workshop Title: How Deep Neural Techniques have revolutionised Information Retrieval and Text Mining

Abstract: The world of information retrieval and text mining is a global one, but the applications of deep learning have become highly specific and specialized. There are huge opportunities in this field. We examine how deep learning have revolutionized the problem of extracting and processing text. We demonstrate the importance of two major breakthroughs in the field: 1) The first, deep convolutional neural network (DNN), which can find hidden layer effects using very sparse input and a few simple operators. It is widely used for word embeddings and natural language processing. 2) The second, reinforcement learning, which can learn the hidden layer effect using a small amount of input. This technique is also widely used for image and speech recognition.

Workshop Title: How Deep Neural Techniques have revolutionised Information Retrieval and Text Mining

Abstract: Over the last few years, deep neural architectures have rewritten the rulebook in terms of the performance that can be achieved across a multitude of text processing tasks from sentiment analysis and sarcasm detection, to machine translation, web search, question answering, and dialog generation. In this workshop I will explain the language modelling technology behind these advances, discussing its evolution from shallow embeddings to modern transformer models composed of ever deeper self-attention networks. I will describe numerous applications of these deep models in information retrieval and text mining and then look to the future, to applications that seamlessly combine information across text and image modalities.

One of these texts was **generated automatically** by conditioning on the title

- the other is my abstract
- can you tell which is which?
- try the text generator yourself here: https://transformer.huggingface.co/

Mark Carman

- If you guessed that the first abstract was the automatically generated one
  - then you were right ;-)
- So what will we talk about in this lecture?
  - Here's the abstract again (my one, not the fake one ;-):

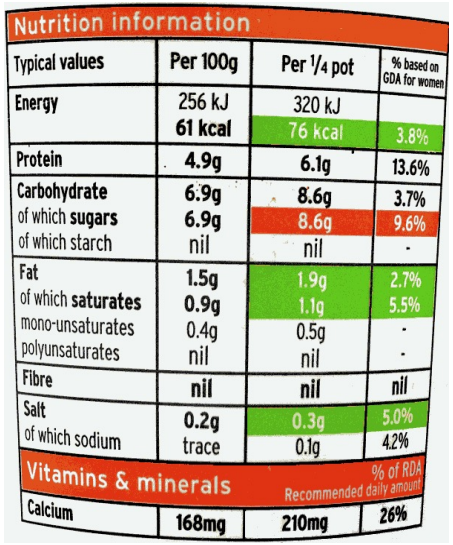
Workshop Title: How Deep Neural Techniques have revolutionised Information Retrieval and Text Mining

Abstract: Over the last few years, deep neural architectures have rewritten the rulebook in terms of the performance that can be achieved across a multitude of text processing tasks from sentiment analysis and sarcasm detection, to machine translation, web search, question answering, and dialog generation. In this workshop I will explain the **language modelling technology** behind these advances, discussing its evolution from shallow **embeddings** to modern **transformer models** composed of ever deeper self-attention networks. I will describe numerous **applications** of these deep models in **information retrieval** and **text mining** and then look to the future, to applications that seamlessly combine information across text and image modalities.

Workshop Contents**

Mark Carman
20.10.2021

1. **motivation**
   - What is language modelling and why should I care about it?
2. **brief history of language models**
   - Markov models, word embeddings, recurrent neural networks, attention
3. **deep learning for text**
   - what is deep learning, self-attention and transformers, BERT vs GPT-2, what can we do with deep models
4. **example applications**
   - Text classification, translation, summarisation, etc.
5. **research applications**



Nutrition information			
Typical values	Per 100g	Per 1/4 pot	% based on GDA for women
Energy	256 kJ 61 kcal	320 kJ 76 kcal	 3.8%
Protein	4.9g	6.1g	13.6%
Carbohydrate	6.9g	8.6g	3.7%
of which sugars	6.9g	8.6g	9.6%
of which starch	nil	nil	-
Fat	1.5g	1.9g	2.7%
of which saturates	0.9g	1.1g	5.5%
mono-unsaturates	0.4g	0.5g	-
polyunsaturates	nil	nil	-
Fibre	nil	nil	nil
Salt	0.2g	0.3g	5.0%
of which sodium	trace	0.1g	4.2%
Vitamins & minerals			% of RDA Recommended daily amount
Calcium	168mg	210mg	26%

Nutrition Information UK Label Yoghurt by Samataroo (CC0 1.0)

**all part of a balanced machine learning diet

Plus: hands on training of models in Google colab!

What is Language Modeling and why should I care about it?

Source: https://pixabay.com/photos/bored-female-girl-people-school-16811/

A language what?

Mark Carman
20.10.2021

According to Wikipedia:

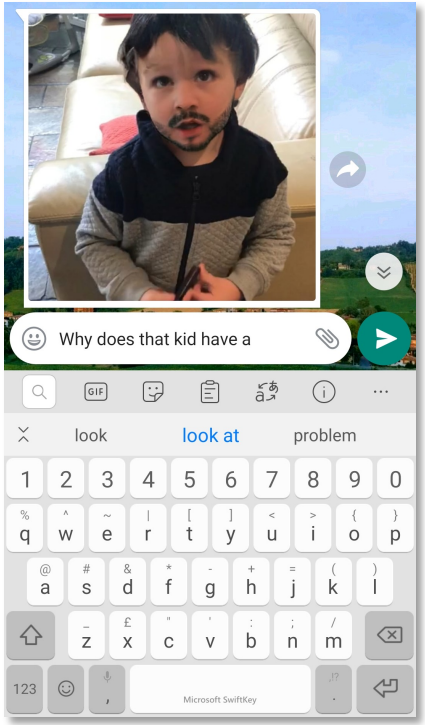
- *A **statistical language model** is a **probability distribution** over **sequences of words***

If we have a distribution over word sequences,

- we can **condition** the next word on the previous content,
- and **sample new sequences** from it

In other words a **language model**

- is general-purpose random **text generator**

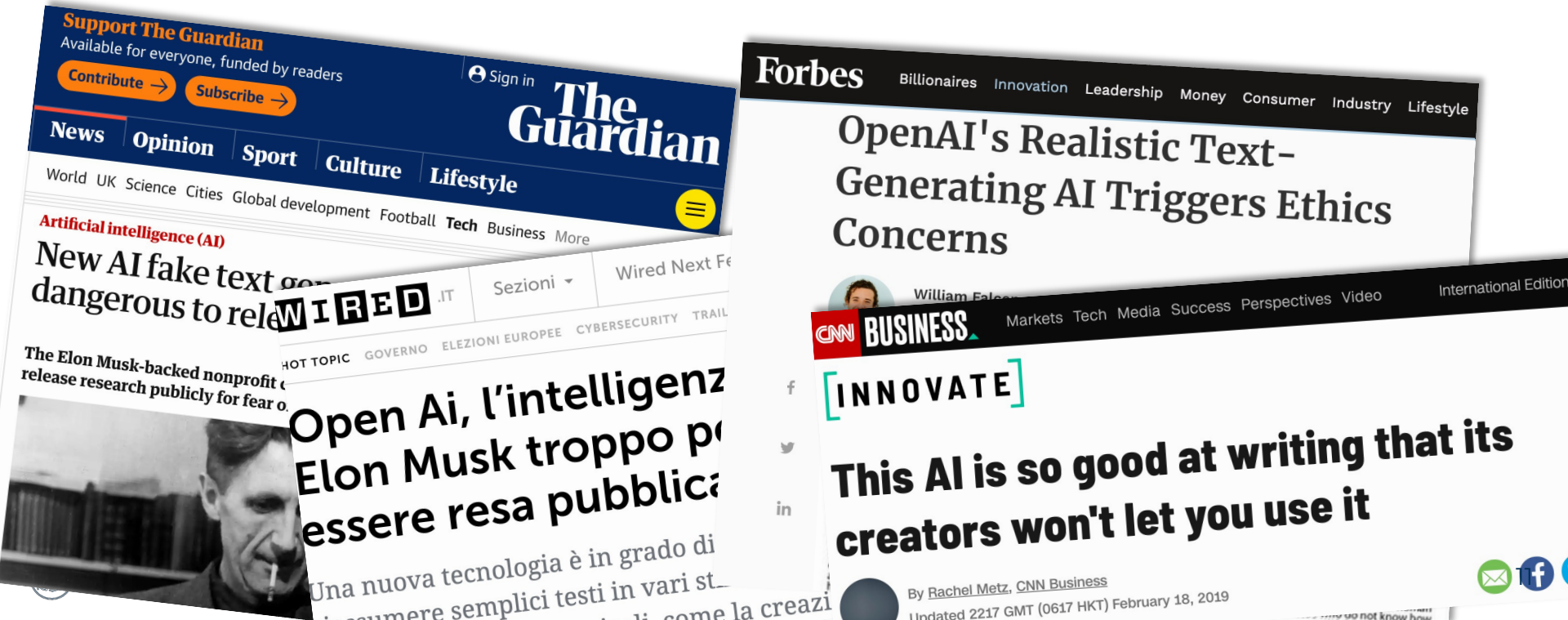


Buzz

Mark Carman
20.10.2021

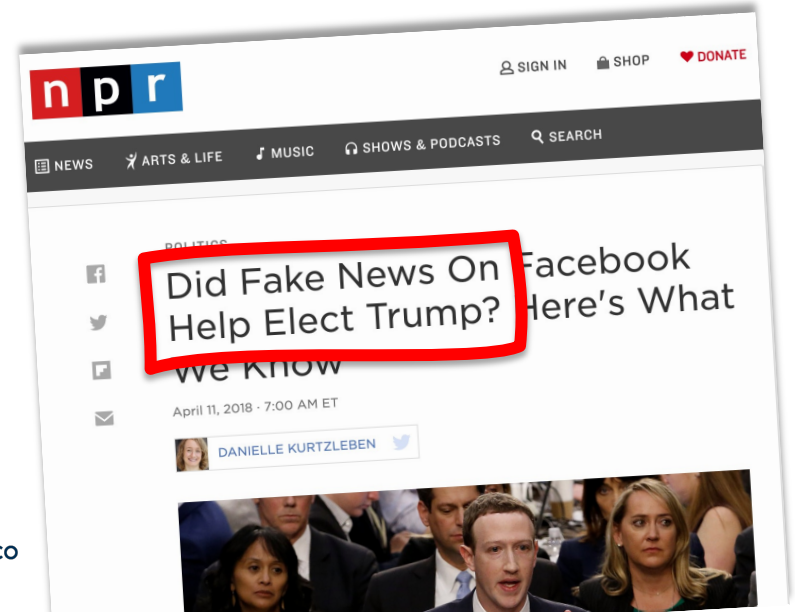
Lots of buzz about language models lately

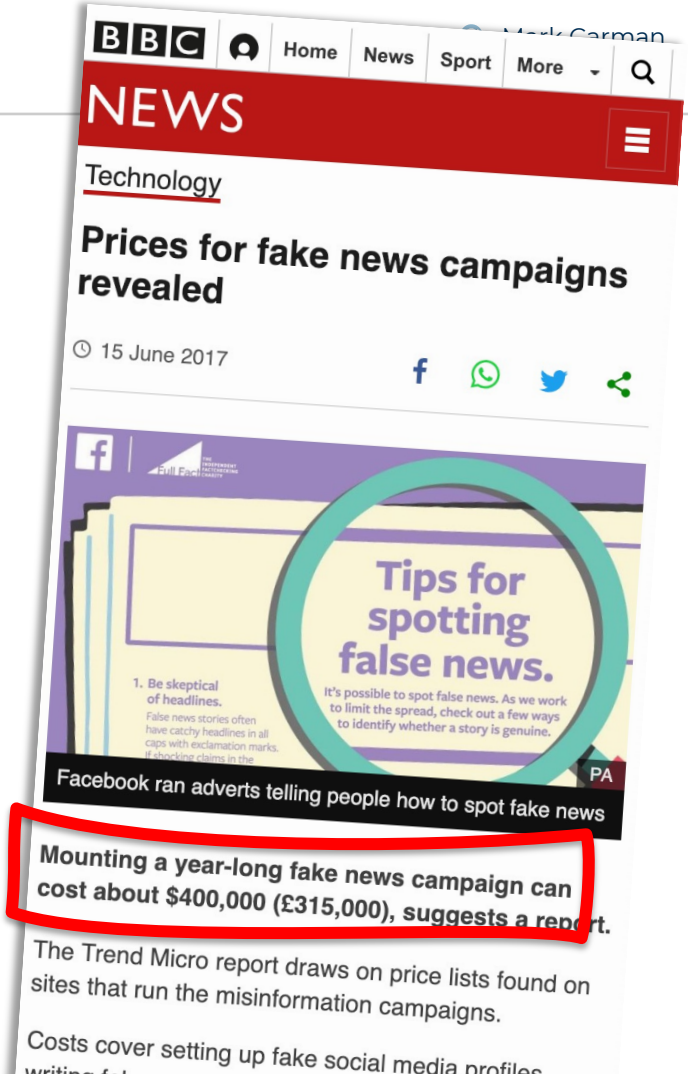
- Could **language models** really be that dangerous?



Impact …

- Language modeling may be a **lucrative business model** …
- and in global politics: *the **automated pen** may be mightier than the sword*





brief history of Language Models

Markov models
word embeddings
recurrent neural networks
attention

language models

Mark Carman
20.10.2021

model **statistical regularities** in text:

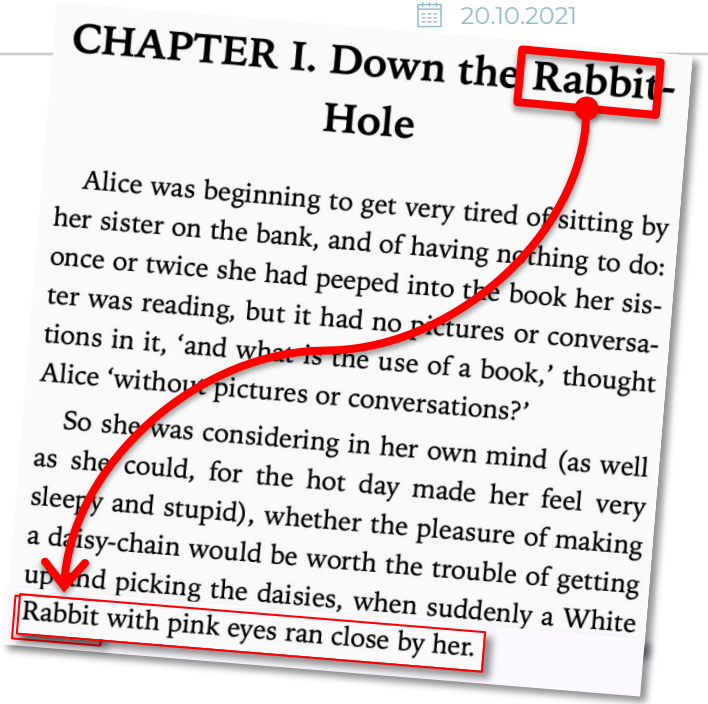
- to predict next word in sentence

example of a regularity?

- next token is often a repeat of a previous word

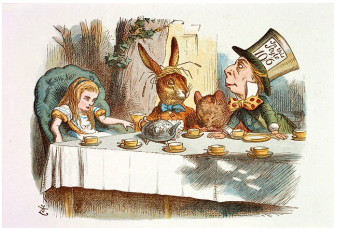
If we can predict next word:

- iterating allows us to predict entire sentence


CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversations?'

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.



Source: https://commons.wikimedia.org/wiki/File:John_Tenniel_-_Illustration_from_The_Nursery_Alice_(1890)_-_c03757_07.jpg

brief history of Language Models

> Markov models
word embeddings
recurrent neural networks
attention

Markov models

Mark Carman
20.10.2021

simplest models count **n-grams** in large corpus

- n-gram = sequence of *n* words
- longer n-grams give better predictions

> 'For the Duchess. An invitation from the Queen to play **croquet**.' The Frog-Footman repeated, in the same solemn tone, only changing the order of the words a little, 'From the Queen. An invitation for the Duchess to play
> ... ???

problem:

- as *n* gets big, chance of finding sequence in corpus drops dramatically
- must back off to shorter n-grams

Query	Google Hits
play station	16100000
play sport.	2680000
play gym	2430000
...	

Query	Google Hits
to play sport	403000
to play croquet	119000
...	

Query	Google Hits
Duchess to play croquet	11000
Duchess to play station	0
Duchess to play sport	0
Duchess to play gym	0
...	

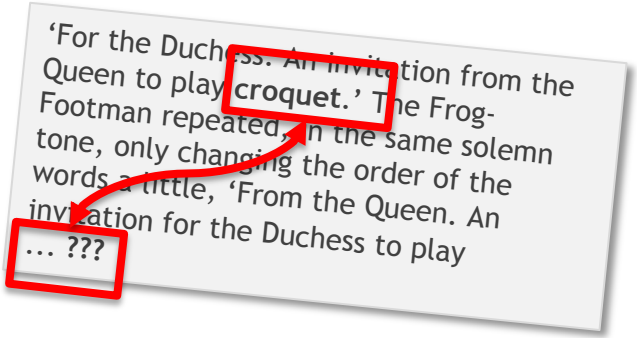
$$P(\text{croquet} \mid \text{play}) = \frac{N(\text{play croquet})}{N(\text{play})}$$

$$P(\text{croquet} \mid \text{to play}) = \frac{N(\text{to play croquet})}{N(\text{to play})}$$

$$P(\text{croquet} \mid \text{Duchess to play}) = \frac{N(\text{Duchess to play croquet})}{N(\text{Duchess to play})}$$

Markov models (cont.)

Mark Carman
20.10.2021



In order to generate reasonable language

- need to model **long distance dependencies**

Memory and data requirements:

- scale exponentially in length of observable dependency
- so **Markov models just don't scale**
  - nonetheless, they were still state-of-the-art not that long ago

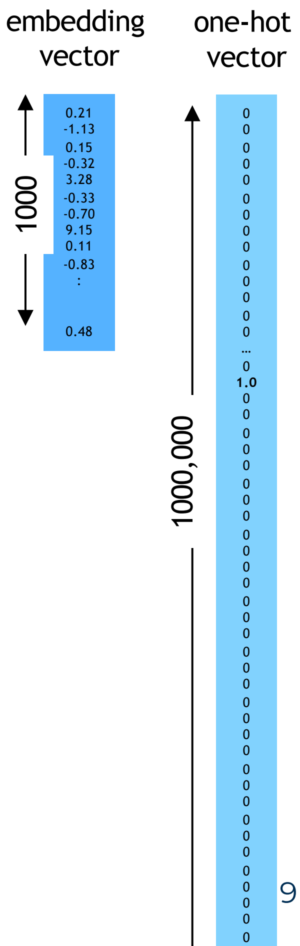
Need instead methods that can both

- **generalise** from limited data
- handle **longer dependencies**

brief history of Language Models

Markov models
> word embeddings
recurrent neural networks
attention

Word Embeddings - what are they?



Dense vectors representing **words** in a **high dimensional space**

- typically have between 100 & 1000 dimensions
- low dimensional compared to one-hot encoding of terms
  - typical vocabulary of document collection may be 100k to 1m items
- just like one-hot encodings they can be aggregated to represent sentences and documents

Word embeddings appeared around 2013 and improved performance on just about every NLP task

Motivating word embeddings

Mark Carman
20.10.2021

Your task is to fill in the blank in the sentence:

'Sure Sally, let's have a skype call at 3pm _____ the 3rd of June.'

What could fit?

- prepositions: on, by, before, around, near, …
- days of the week: Monday, Tuesday, Wednesday, …
- timezones: GMT, CET, EST, AEST, …

Note:

- very **few words fit**
- those that do come in **groups** that are **semantically related** to one other

Embeddings are produced by **supervised machine learning models**

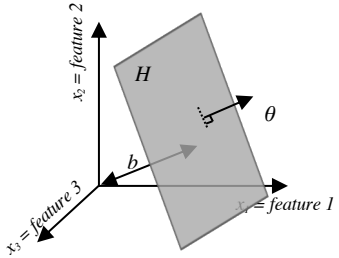
- models trained to **predict missing word** based on **surrounding context**
- context may include only previous words (causal models)
- or also future words (non-causal models)

Supervised Learning Problem

Mark Carman
20.10.2021

Predicting missing word:

- **Features:** words in current context:
  - "Sure", "Sally", "let's", "have", "a", "skype", "call", "at", "3pm"
- **Target:** missing word from sequence
  - multi-class problem (estimate probability for every word in vocab)



Issue:

- requires a very large number of parameters!
- example:
  - multi-class linear classifier (e.g. Logistic Regression) to predict all word in vocabulary
  - with bag-of-words feature vector (so ignoring word order)
  - requires parameters quadratic in the size of the vocab
  - if vocab=100,000, then we would have 10 billion parameters!!
  - which **used to be a lot** before deep learning came along 🤣😂

Word2Vec

Word2Vec developed in 2013 by Mikolov et al.
- following early work by Bengio et al. in 2003
- later GloVe in 2014 by Penington et al.

Word2Vec solved the parameter space issue by using:
1. bag-of-words representation
2. neural network with single (linear) hidden layer
3. training model in discriminative fashion
   - by inventing negative examples

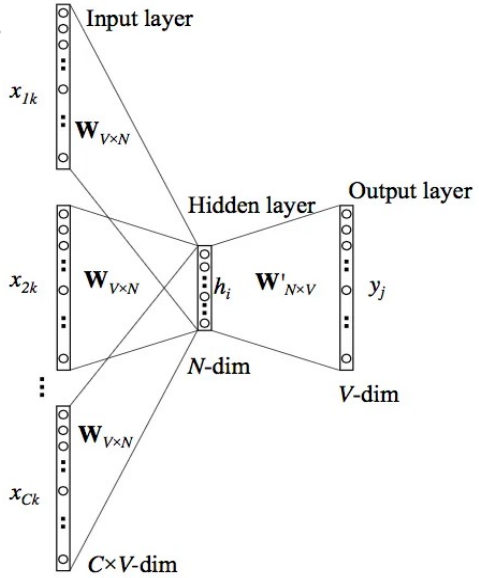


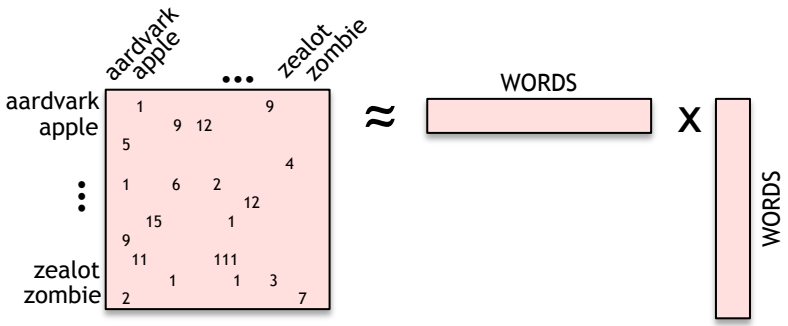
Image source: http://www.stokastik.in/understanding-word-vectors-and-word2vec/

Just matrix decomposition

Mark Carman
20.10.2021

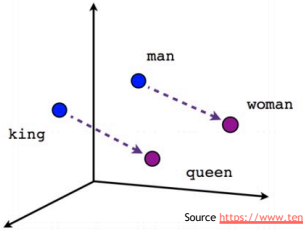
Word embeddings can be seen as a form of **matrix decomposition**

- square count matrix: vocabulary × vocabulary
- contains co-occurrences within a fixed-size context window
- factorizing **generalises** the information in those windows



Properties of Word Embeddings

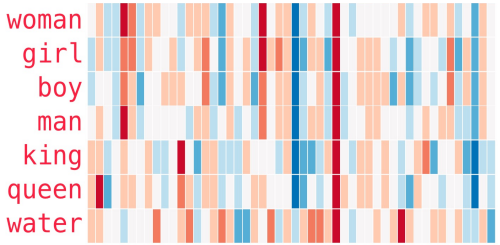
Mark Carman
20.10.2021



Source https://www.tensorflow.org/images/linear-relationships.png

Word embeddings have interesting properties

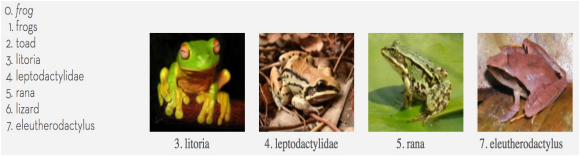
- translation in the space meaningful
- semantics is additive



Source: http://jalammar.github.io/illustrated-word2vec/

Semantic Clustering:

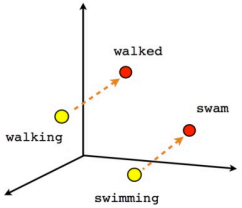
- neighbours in space are **semantically** related



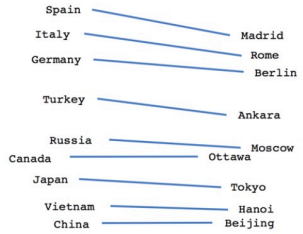
Source https://nlp.stanford.edu/projects/glove/

Discovers **relationships** between words:

- e.g. part-of-speech, type-of, geographic, etc.



Verb tense



Country-Capital

Source: https://www.tensorflow.org/images/linear-relationships.png

Mark Carman

Low dimensional representation causes similar terms to share similar descriptions

- allows model to generalize from semantically related examples

the **Duchess** to play … ??

Examples from corpus with similar contexts:

- the **Queen** to play croquet
- the **Duke** to play chess

- e.g. part-of-speech and hypernym (type-of) relationships implicitly encoded in embedding vector in additive manner

the **Duchess** to play … ??

Look for examples with any of these types

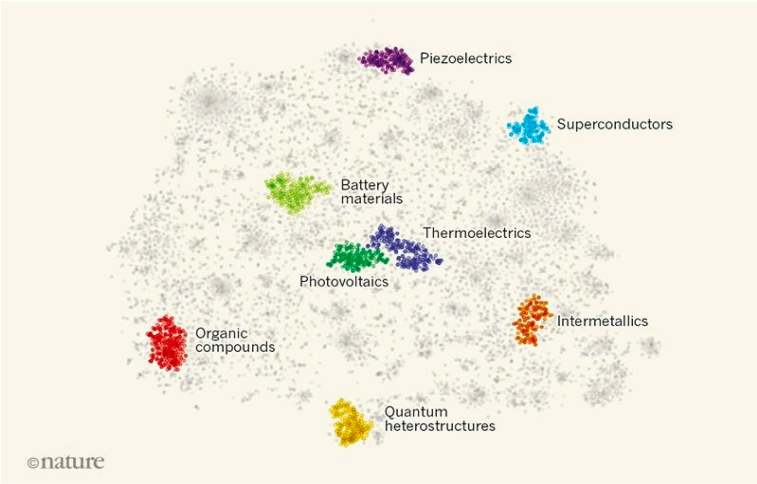
- the **[noun+person+female+royal]** to play croquet

Useful for Mining Text

Mark Carman
20.10.2021

Embeddings place similar concepts close together

- useful for discovering implied (but unknown) properties of them



Visualisation from news article "Text mining facilitates materials discovery" by Olexandr Isayev ,

Sub-word embeddings

Mark Carman
20.10.2021

Word embeddings work well if vocabulary is fixed

- so **no new words** in test set
- if we see a new word, don't have embedding for it!

Fasttext (2016 Bojanowski et al.)

- split words into character sequences
- learns embeddings for character n-grams
- combines the embeddings to form words

Advantage:

- deals nicely with morphologically related terms, so:
  - "**believ**e" and "**believ**ing" have similar representations
  - as do "**rain**" and "**rain**fall"

```
Embeddings are cool.
=>
<Em
 Emb
  mbe
   bed
    edd
     ddi
      din
       ing
        ngs
         gs>

            <ar
             are
              re>

                <co
                 coo
                  ool
                   ol>
```

brief history of Language Models

Markov models
word embeddings
> recurrent neural networks
attention

Recurrent Neural Networks (RNNs)

Mark Carman
20.10.2021

Now have way to represent words in semantic space ✔👍
- still need to aggregate information over longer contexts

RNNs provide general way to accumulate information
- by combining context from the previous words
- with the embedding of current word

RNNs are simply models which:
- take 2 vectors as input: <current input, previous state>
- produce 2 vectors as output: <current output, updated state>

They can be used to process arbitrarily long input contexts
- i.e. encode a sequence of text to a single embedding

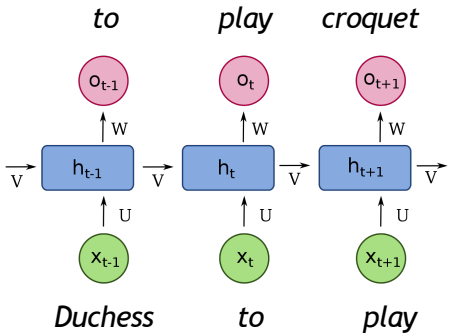
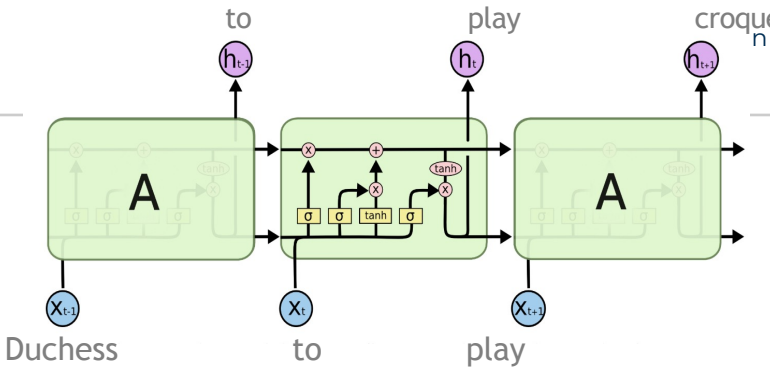


Image source:
https://en.wikipedia.org/wiki/Recurrent_neural_network

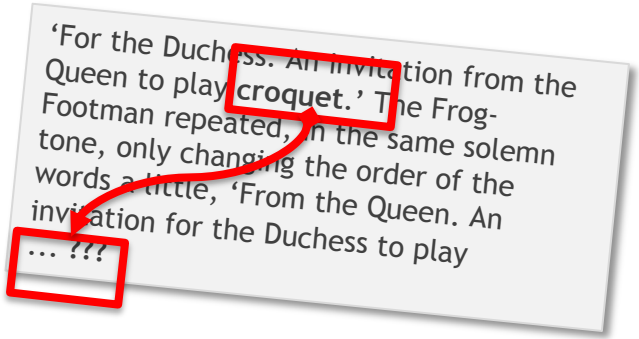
LSTMs



Images source: Understanding LSTM Networks by Christopher Olah
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Long Short Term Memory (**LSTM**) networks

- allow for **long term dependencies**
- gates allow model to **remember/forget** information
- model learns what type of information to keep and what to discard
- **default** operation is to pass information from one state to the next



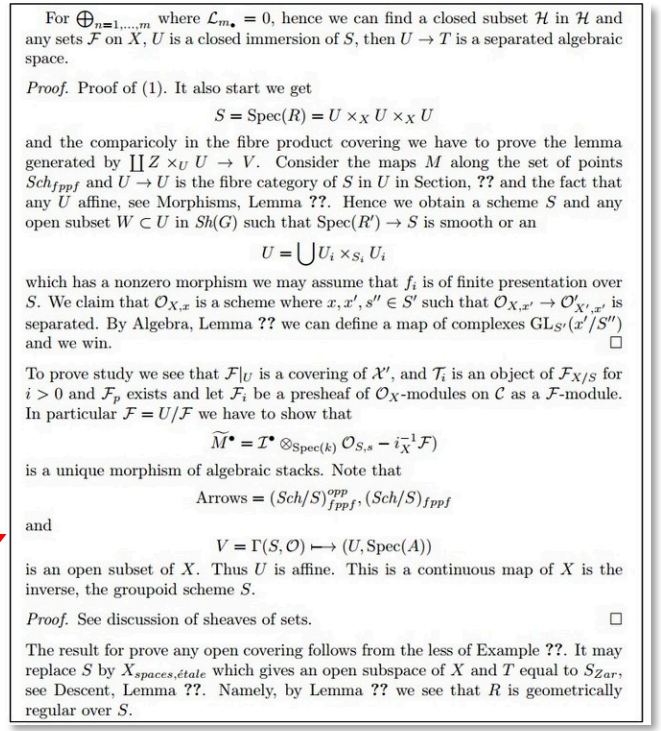
Stacked LSTMs and contexts

Mark Carman
20.10.2021

Stacked LSTMs: layers of LSTMs placed on top of each other

- have uncanny ability to remember **nested contexts**
- e.g. they don't forget to close the brackets at the end of a maths expression

PDF compiled from Latex source code that was generated by a multi-layer LSTM (by Andrej Karpathy)
http://karpathy.github.io/2015/05/21/rnn-effectiveness/



For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the compariccly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example **??**. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma **??**. Namely, by Lemma **??** we see that R is geometrically regular over S.

Contexts in natural language

Mark Carman
20.10.2021

The ability to deal with contexts is useful for natural language too

- Consider gender and possessive pronouns:
  - Fill in the gaps in the following sentences with one of {**he**, **she**, **his**, or **her**}

  My mother was taking on the phone to ___ friend Jim. Jim said that ___ favourite game was confusing students. Replying, ___ said that he should find a better hobby.
  - answer: (“her”, ”his”, “she”), because mother is feminine and Jim is masculine.
  - subject changes gender from one sentence to the next
  - LSTM is able to remove feminine subject and add masculine one, etc.
- Or consider sentiment and scoping of negation:
  - What’s more likely to come next in these sentences: “**friendly**” or “**self-absorbed**”?

  I get along well with her brother. He’s always ___

  I can’t get along well with her brother. He’s always ___

  I can’t help but get along well with her brother. He’s always ___

brief history of Language Models

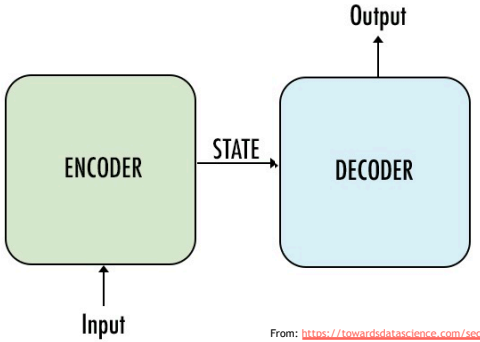
Markov models
word embeddings
recurrent neural networks
> attention

Sequence to sequence (seq2seq) models

Mark Carman
20.10.2021

LSTMs are so powerful that they were soon used for **translation** and **dialog systems**

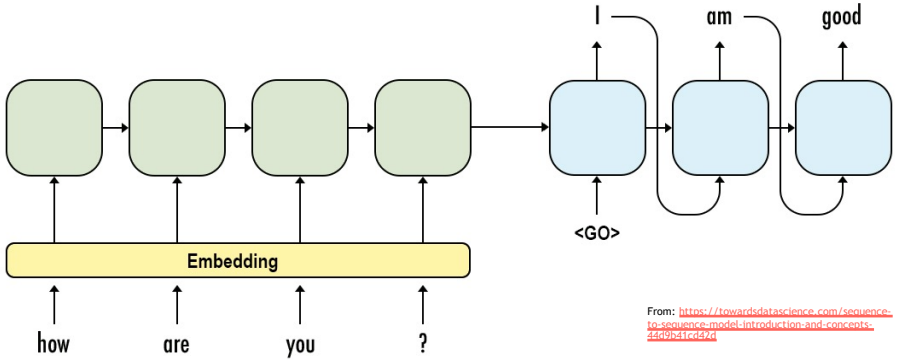
- How can one learn translation models with LSTMs?



From: https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d

Train 2 different RNN models

- **encoder**: reads in input text and generates a representation (an embedding) of entire sequence
- **decoder**: take output of encoder and serialise it into text

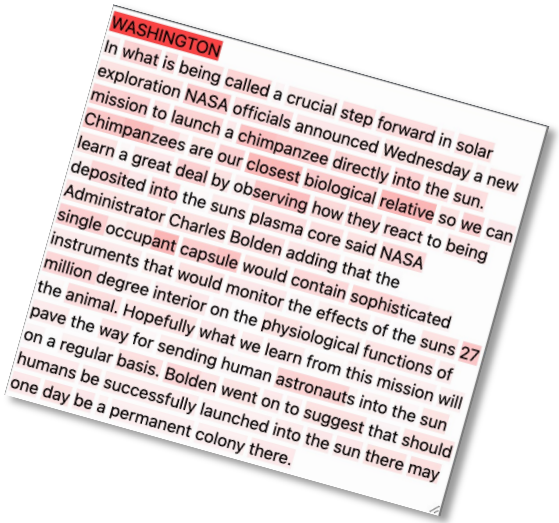


From: https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d

Attention!

Attention is a critical building bock for modern image and text processing

- what is attention?
- why it is implemented?
- how does it work?

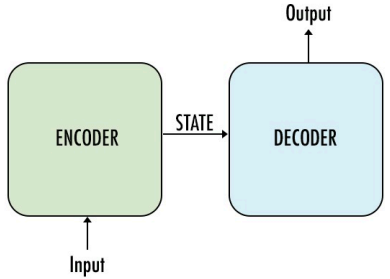

WASHINGTON
In what is being called a crucial step forward in solar exploration NASA officials announced Wednesday a new mission to launch a chimpanzee directly into the sun. Chimpanzees are our closest biological relative so we can learn a great deal by observing how they react to being deposited into the suns plasma core said NASA Administrator Charles Bolden adding that the single occupant capsule would contain sophisticated instruments that would monitor the effects of the suns 27 million degree interior on the physiological functions of the animal. Hopefully what we learn from this mission will pave the way for sending human astronauts into the sun on a regular basis. Bolden went on to suggest that should humans be successfully launched into the sun there may one day be a permanent colony there.

A simpler way …

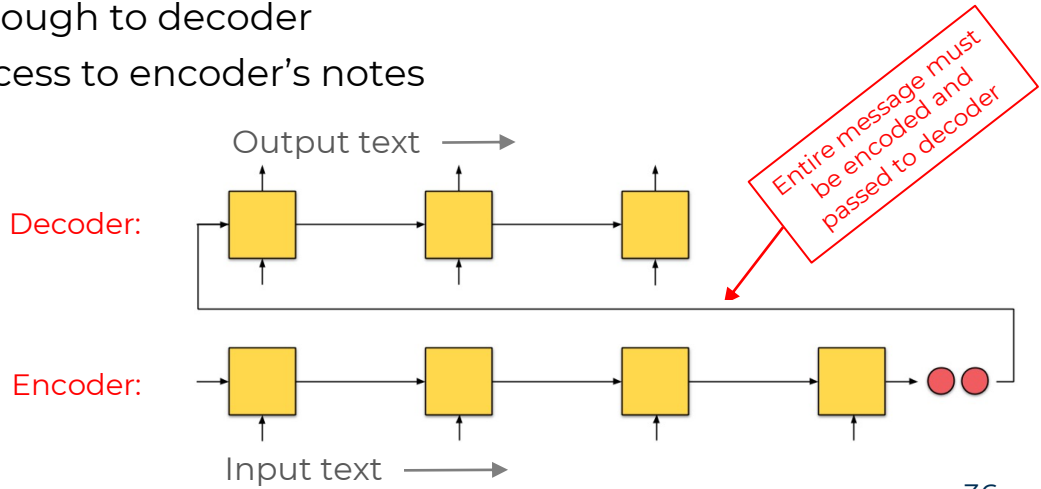
Mark Carman
20.10.2021

Ever thought about interpreters translating politicians during meetings?

- must wait for rambling politician to stop speaking before they start translating…
- that's a lot of stuff to remember!
- same problem for encoder-decoder architecture
  - too much to information to pass through to decoder
  - easier to translate if decoder has access to encoder's notes



From: https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d



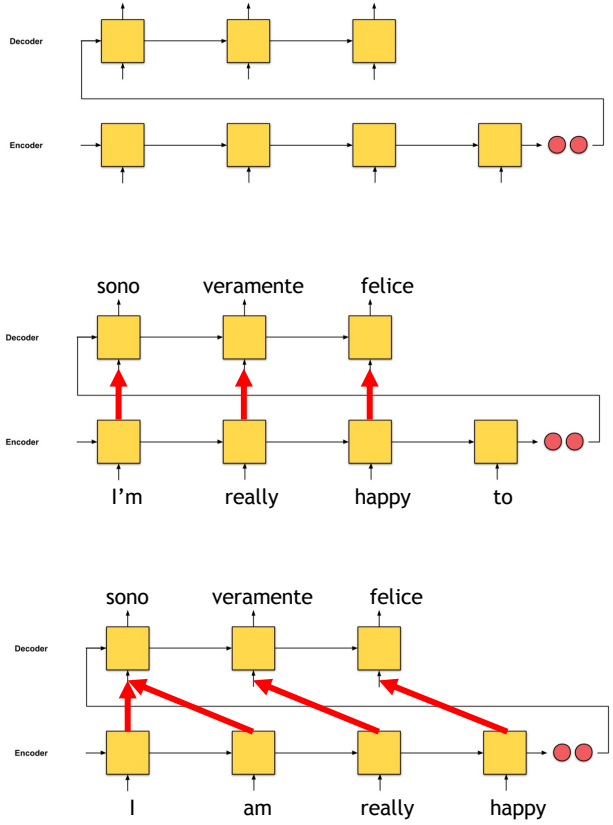
Source: https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Towards attention

Mark Carman
20.10.2021

Attention models

- make encoded input available to decoder
- provides a direct route for the information to flow from input to output
- why not just directly map input words to output words?
- varying number of tokens used to describe same concept and different word order across languages

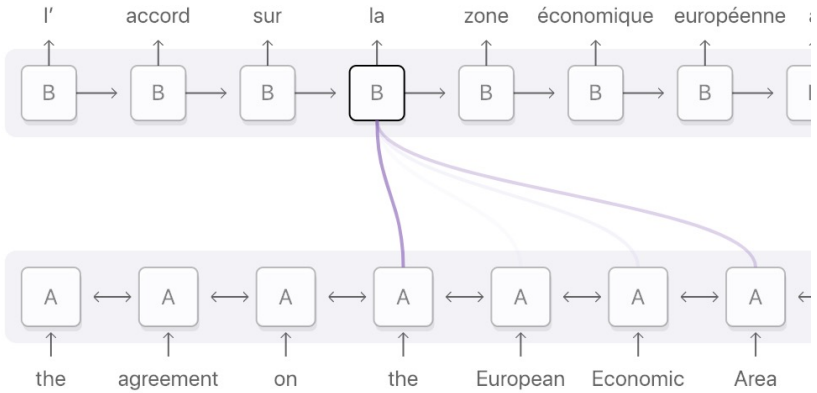


Adapted from: https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Towards attention

Mark Carman
20.10.2021

- Moreover generating the right output word often requires knowing more than just the current word in the input
- Indeed it can require knowing the value of a future word

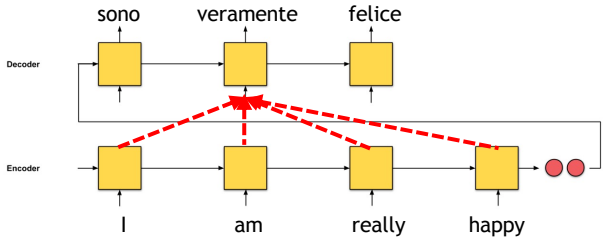


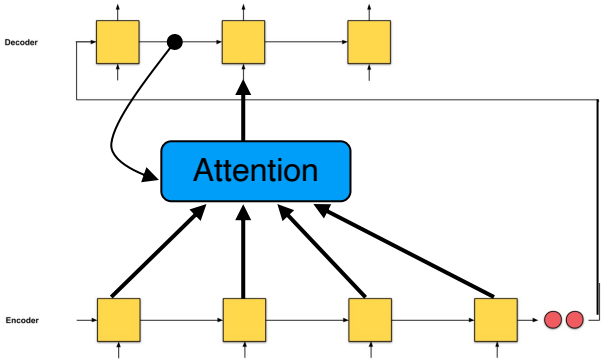
Source: https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Need for attention

Mark Carman
20.10.2021

- Need mechanism to pass information from embeddings of input words to corresponding output word
- Attention provides a direct route for the information to flow from input to output
- What information flows into the decoder is controlled by the previous state of the decoder





Source: https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Soft attention over input

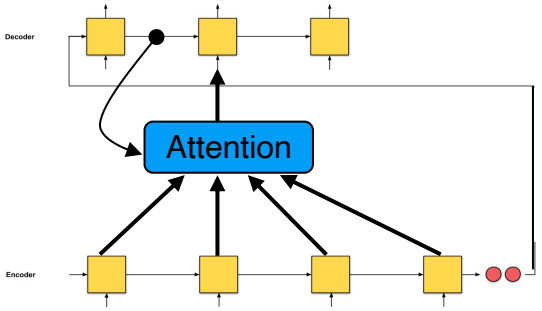
Mark Carman
20.10.2021

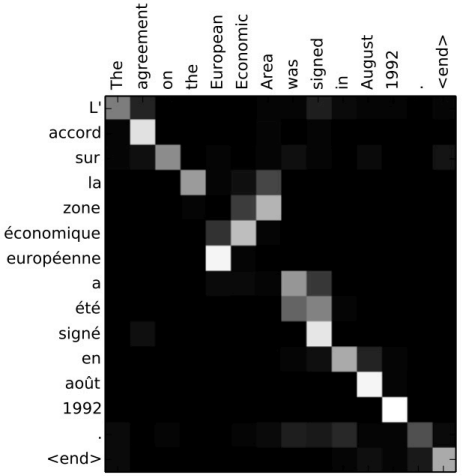
Similarity is computed between state of decoder and embedding of each term

- embedding of input term then weighted by score

soft-attention produces **weighted average** over input embeddings

- Example of Bahdanau attention for English to French translation shown





Source 2015 paper by Bahdanau et al. :https://arxiv.org/pdf/1409.0473.pdf

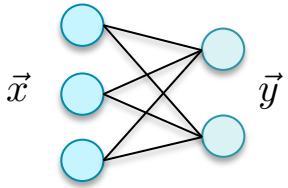
Deep Learning for text

> what is deep learning?
self-attention and transformers
BERT vs GPT-2
what can we do with deep models?

What is deep learning?

Mark Carman
20.10.2021

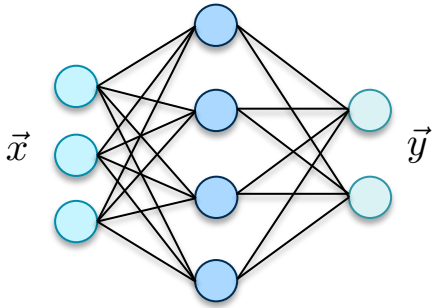
Deep learning is the learning of deep neural networks.



$$\vec{y} = tanh(W\vec{x} + \vec{b})$$

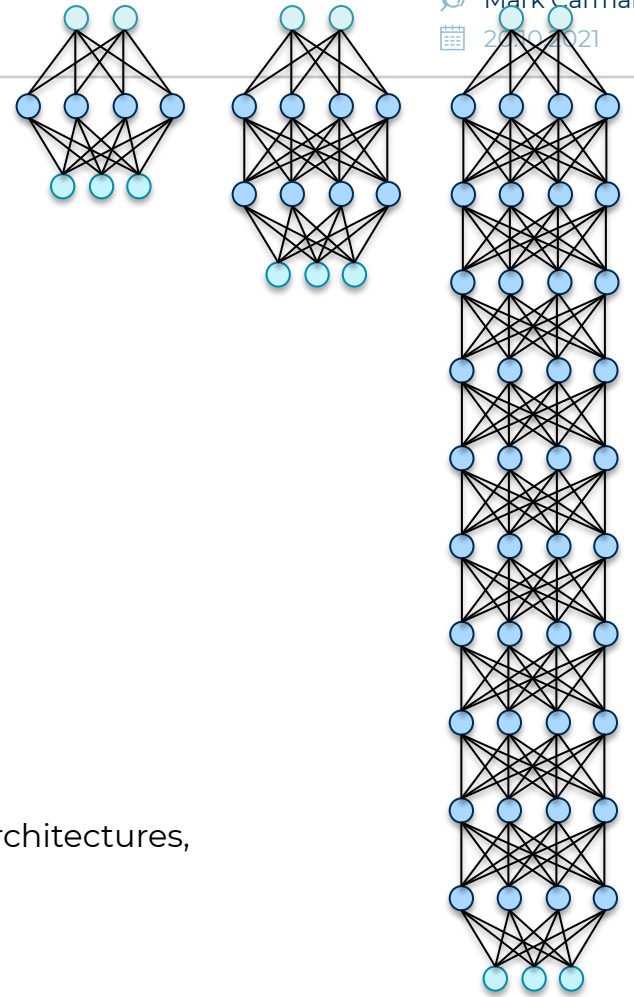
So what are neural networks?

- networks of nodes, where
  - each node acts as simple linear classifier (e.g. logistic regression)
  - previous layer of nodes corresponds to features for next layer
- networks with a hidden (internal) layer can learn any non-linear function



$$\vec{y} = tanh(W_2 \cdot tanh(W_1\vec{x} + \vec{b}_1) + \vec{b}_2)$$

What is deep learning?



Networks with **many** layers

- why do we want many layers?
  - **improved performance**
  - ability to **learn useful features** automatically
- what is the downside?
  - need **lots of training** data!
  - and large **computing resources** (GPUs)
  - can be unstable & much harder to train …

Why is deep learning so big now?

- provides **amazing performance** on text and images
- possible due to:
  - hardware advancements, huge data quantities, clever architectures, new training procedures, and specialised toolkits

Why is deep learning important for text?

Mark Carman
20.10.2021

State-of-the-art performance for most text processing tasks

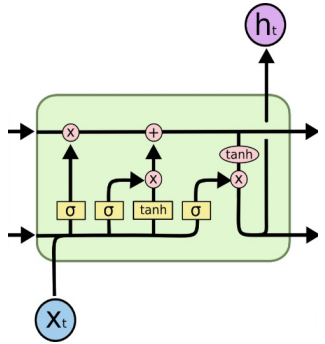
- including **classification, summarisation, generation, translation**

Up until recently deep architectures involved

- stacking LSTMs on top of each other

Over last couple of years new type of architecture has emerged

- called a Transformer
- makes use of **self-attention** networks



Images source:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

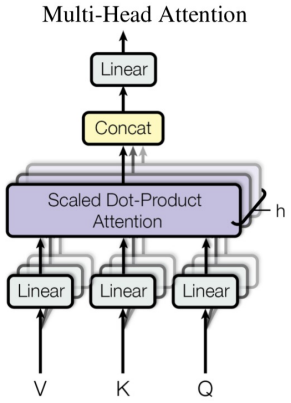


Image Source:
"Attention Is All You Need" by Vaswani et al.
https://arxiv.org/pdf/1706.03762.pdf

Deep Learning for text

what is deep learning?
> self-attention and transformers
BERT vs GPT-2
what can we do with deep models?

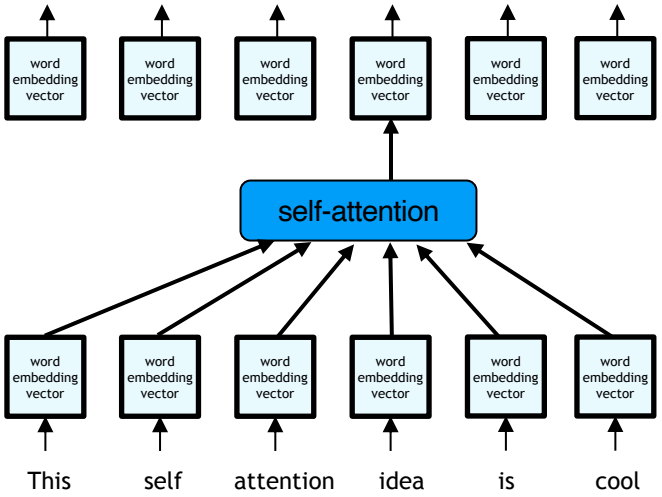
So what is self-attention?

Self attention is a mechanism for:

- combining word embedding vectors to produce new word embedding vectors
- each high-level embedding is **weighted average** of word embeddings below it

Weights are computed:

- based on similarity between embeddings in respective positions
- model parameters control process
  - learn how best to compute the weights

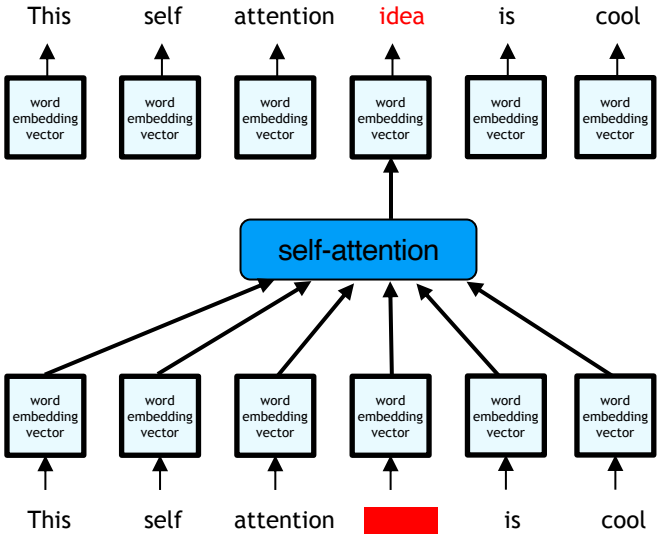


So what is self-attention?

Mark Carman
20.10.2021

Self attention models are trained to recover missing words from the input sentence

- i.e. to perform the language modeling task



Motivating self-attention

Mark Carman
20.10.2021

To understand why self-attention is useful for language models, consider that

- words take on different meanings depending on their context
- attention mechanism allows representation to **depend on context**
- learns **weighting function** over lower-level embeddings of context terms

I arrived at the bank after crossing the street.

I arrived at the bank after crossing the river.

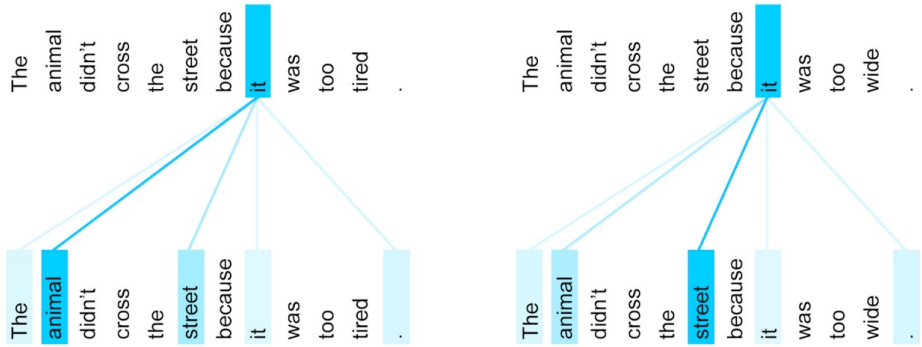


Image source: "Transformer: A Novel Neural Network Architecture for Language Understanding", by Uszkoreit et al. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

What makes transformers deep?

Basic self-attention module is stacked on itself **many** times

- allows semantics of each word to build up over multiple steps

Each transformer module contains:

- **multiple attention heads** working in parallel
- **feedforward network**, with **residual connections** and **normalisation**

Architecture is word position agnostic

- so **positional encoding** provided as additional input to bottom layer

Note: transformers are faster to train than stacks of RNNs

- in RNNs gradient must be iterated back along sequence

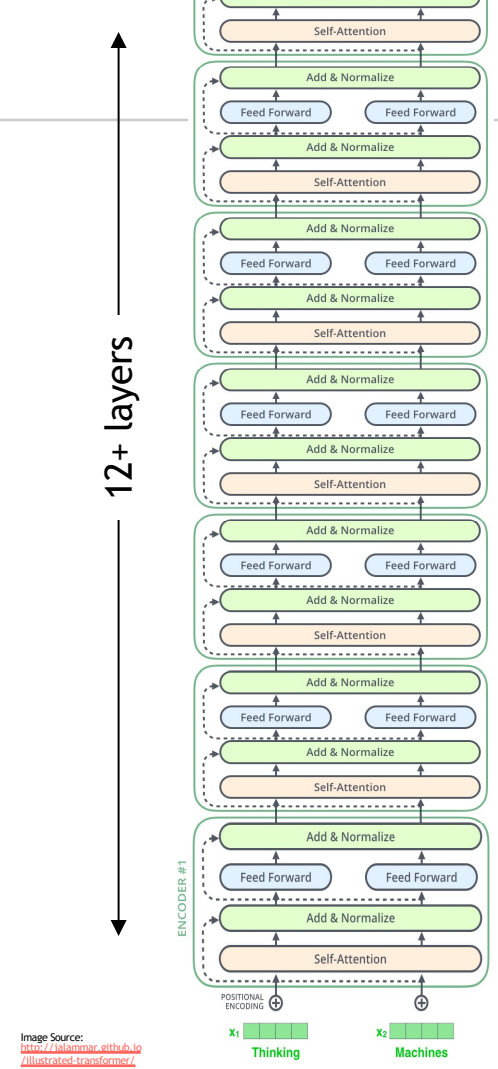


Image Source: http://jalammar.github.io/illustrated-transformer/

Transformer sizes

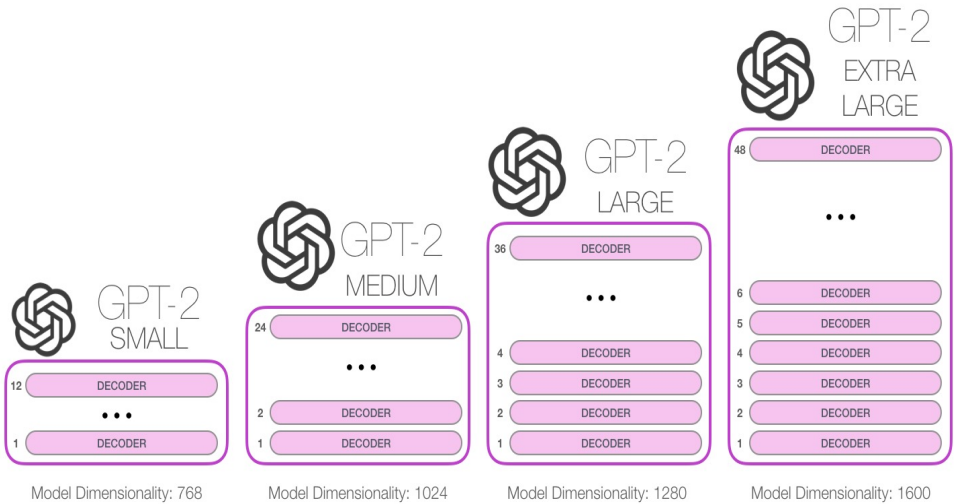
Mark Carman
20.10.2021

Transformers come in multiple sizes, depending on

- number of self-attention layers
- size of the embedding used at each layer
- number of parallel attention heads

More parameters results in:

- better performance
- but longer training times
- and larger memory requirements



Source: http://jalammar.github.io/illustrated-gpt2/

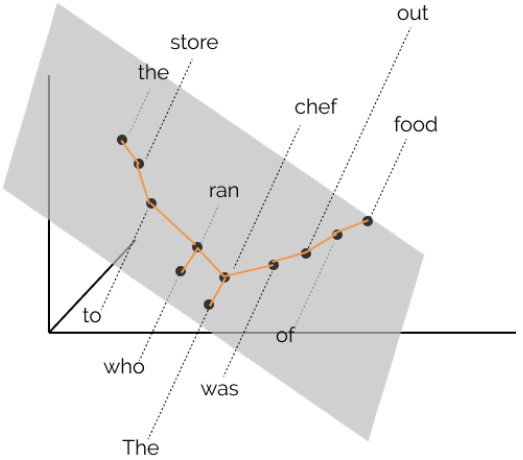
Why is stacked self-attention so useful?

Mark Carman
20.10.2021

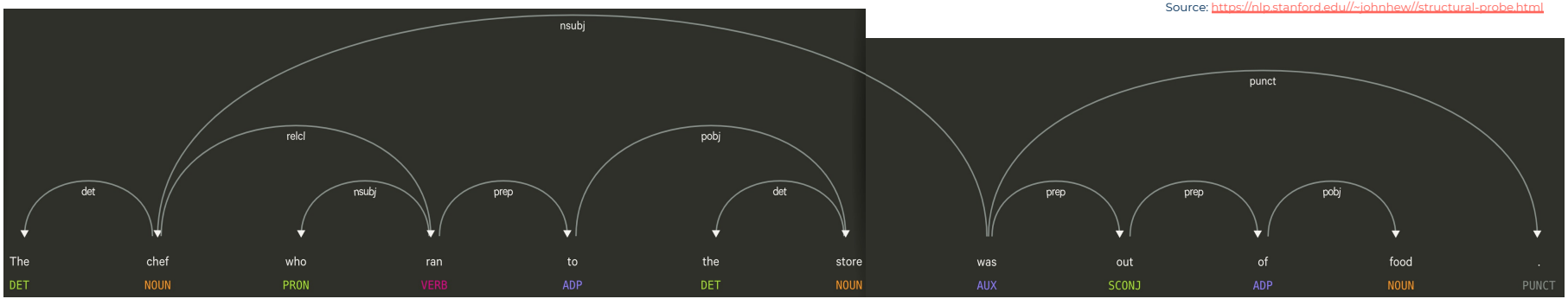
because transformers **effectively learn** how to build a **dependency parse tree** over concepts in the text

- Hewitt et al. 2019 https://nlp.stanford.edu/pubs/hewitt2019structural.pdf

- consider examples:
  - The store was out of food. ___
  - The chef who ran to the store was out of food. ___
- to predict next sentence, need to who is out of food, the store or the chef



Source: https://nlp.stanford.edu/~johnhew//structural-probe.html



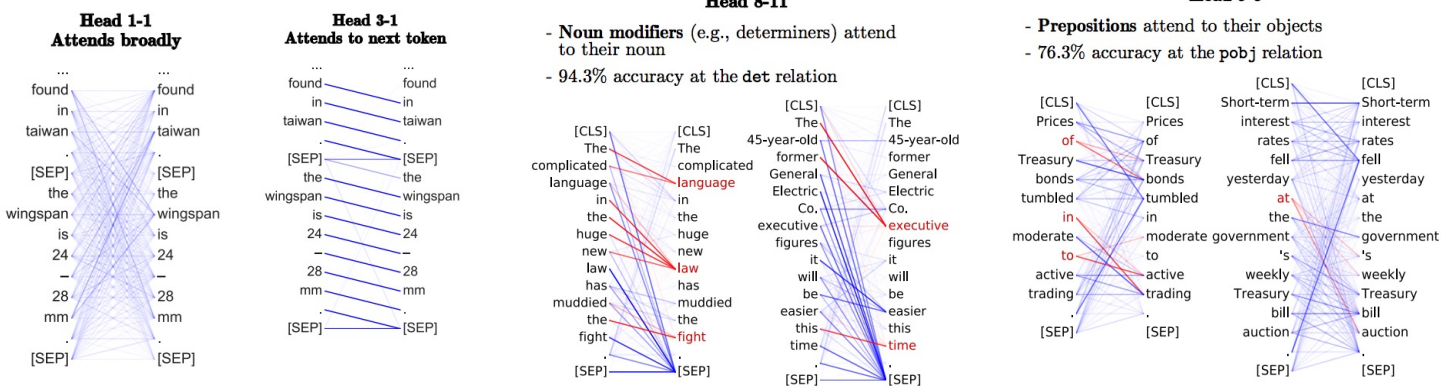
Source: https://explosion.ai/demos/displacy?text=The%20chef%20who%20ran%20to%20the%20store%20was%20out%20of%20food.&model=en_core_web_sm&cpu=0&cph=0

Why is stacked self-attention so useful?

Mark Carman
20.10.2021

Lots of visualisation is going on trying to interpret what is being learnt.

- some heads simply *aggregate information* or *attend to a previous token*
- others *learn language relationships* (2019 paper by Clark et al.)
  - see demo: https://colab.research.google.com/drive/1PEHWRHrvxQvYr9NFRC-E_fr3xDq1htCj



Source: https://arxiv.org/pdf/1906.04341.pdf

Deep Learning for text

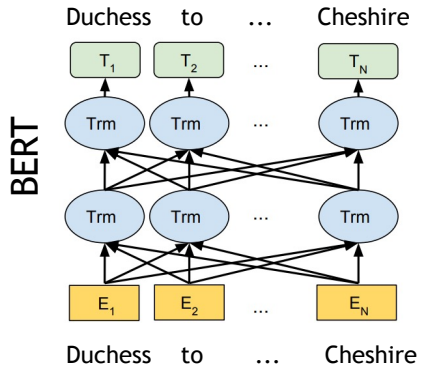
what is deep learning?
self-attention and transformers
> BERT vs GPT-2
what can we do with deep models?

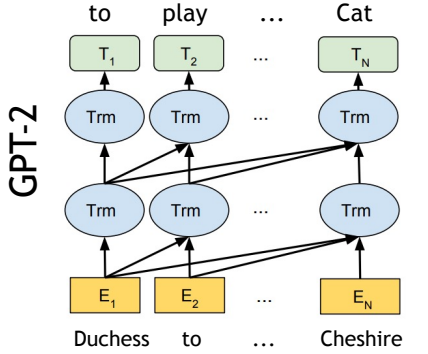
Story of two architectures

Mark Carman
20.10.2021

- **BERT** = **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
  - 2019 paper by Devlin et al. (Google)
  - **Autoencoder**: text in at bottom, same text comes out at top
  - Recovers input text
  - Great for **representing text** (e.g. for building classifiers)

- **GPT-2** = **G**enerative **P**retrained **T**ransformer (Version **2**)
  - 2019 paper by Radford et al. (OpenAI)
  - **Autoregressive**: text in at bottom, text shifted one to the left
  - Predicts the next token
  - Great for **generating text**





Images from: https://arxiv.org/pdf/1810.04805.pdf

aside: transformers use sub-word tokens

Mark Carman
20.10.2021

Find most common character sequences by performing a byte-pair encoding

- iteratively replace most frequent consecutive characters:

```
though they think that the thesis is thorough enough
     th → θ
θough θey θink θat θe θesis is θorough enough
     ou+g+h → ə
θə θey θink θat θe θesis is θorə enə
     θe → ψ
θə ψy θink θat ψ ψsis is θorə enə
```

- In this way, common prefixes/suffixes become vocabulary elements:

Intermediate tokenization: [“I", “like", “playing", “football", “."]

BPE tokenization: [“I", “like", “play", “ing", “foot", “ball", “."]

How are BERT and GPT trained?

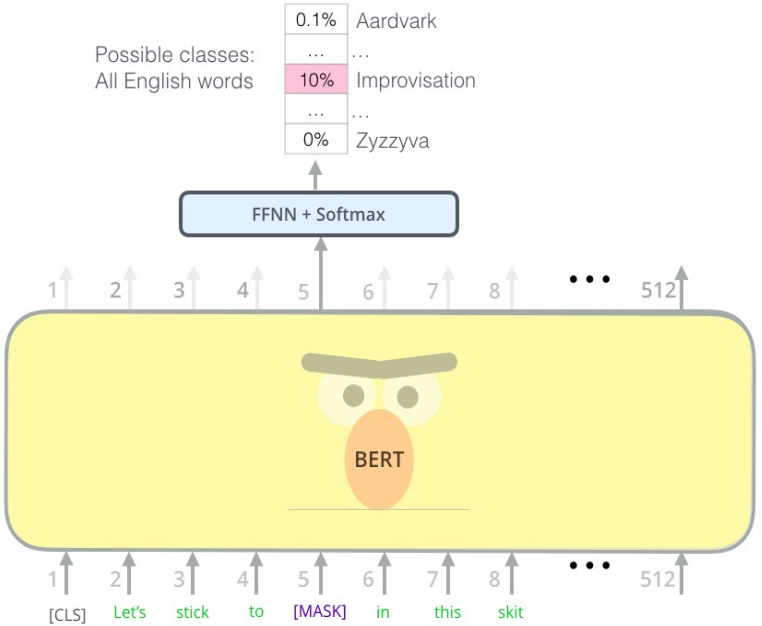
Mark Carman
20.10.2021

BERT:

- by **masking out** random words in the input using a special [MASK] token
- model must recover all words including the masked ones

GPT-2

- by simply masking the future words in the sequence and at each point predicting the next word



Source: http://jalammar.github.io/illustrated-bert/

What data were they trained on?

Garbage in => garbage out

- model will produce similar text to that which it was trained on

GPT-2:

- trained on 40GB of web text that Reddit users rated highly

Image source: https://en.wikipedia.org/wiki/Reddit

BERT:

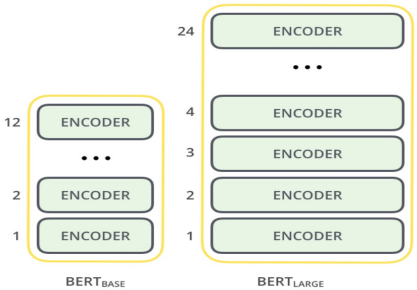
- trained on Wikipedia and a corpus of books

Image source: https://en.wikipedia.org/wiki/File:Wikipedia-logo-v2.svg

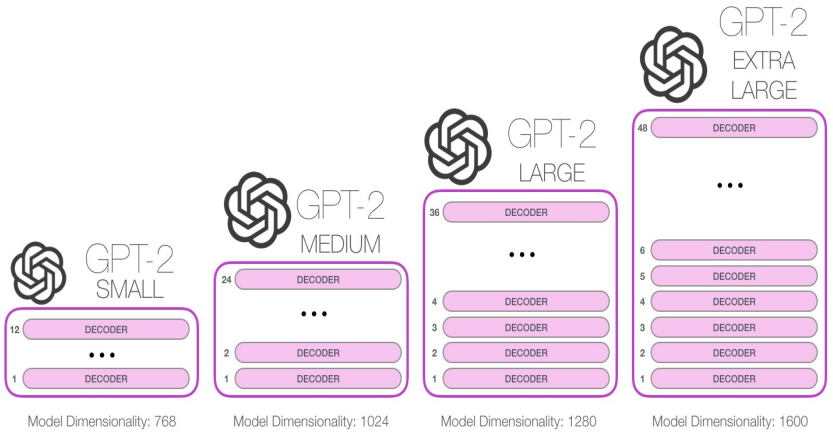
BERT & GPT-2 sizing

Mark Carman
20.10.2021

- Both models come in different sizes
  - number of self-attention layers
  - size of the embedding used at each layer
- more parameters => longer training time but better performance



Source: http://jalammar.github.io/illustrated-bert/



Source: http://jalammar.github.io/illustrated-gpt2/

How many parameters is that?

Mark Carman
20.10.2021

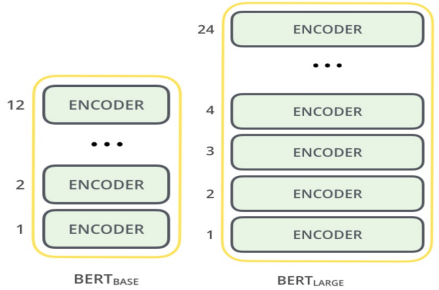
Sizing for BERT models:

- base model has 110M parameters
- large model has 340M parameters

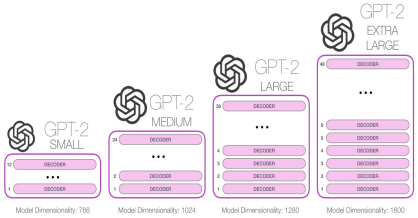
Sizing for the GPT-2 models:

- largest has **1.5 billion** parameters!
- vocabulary of 50,257

Most applications use a **context size** of up to **1000 tokens**



Source: http://jalammar.github.io/illustrated-bert/



Source: http://jalammar.github.io/illustrated-gpt2/

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

From: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Other related models

- **RoBERTa** (Facebook's version of BERT)
  - modifies slightly training objective
  - trained on more data with larger batches

- **XLNet** (BERT with some GPT-2)
  - introduces autoregressive modelling (GPT-2) into BERT training
  - was quite hyped for a while

- **DistilBERT** (a distilled version of BERT)
  - designed to be smaller (40%) and faster (60%) to fine-tune, while retaining 97% of accuracy

- **T5** (Text-To-Text Transfer Transformer)
  - uses encoder+decoder model, same as the original transformer paper
  - so particularly useful for translation or other text2text problems

ever BIGGER models

Mark Carman
20.10.2021

Competition at moment to build ever bigger models

- BERT has **340 million** parameters
- GPT-2 has **1.5 billion**
- Microsoft's Turing-NLG has **17 billion**
- OpenAI's GPT-3 has **175 billion** parameters!

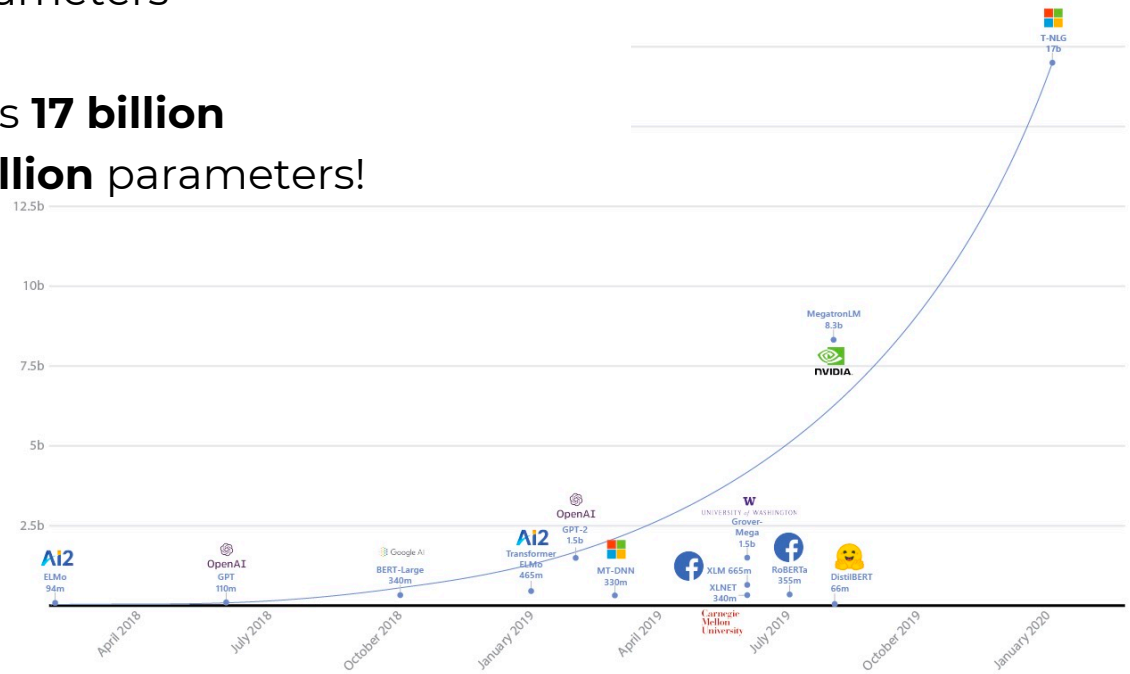


Image source: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Deep Learning for text

what is deep learning?
self-attention and transformers
BERT vs GPT-2
> what can we do with deep models?

what can we do with deep models?

Text Generation
Fine Tuning
One-shot Learning

GPT-2 is a text generator

Mark Carman
20.10.2021

Context (WebText test)

Corporal Michael P. Goeldin was an unskilled laborer from Ireland when he enlisted in Company A in November 1860. Goldein survived the war. Corporal Patrick O'Neal, also from Ireland, first enlisted in 1854 and served with Company L, 3d U.S. Artillery, in Oregon. He returned to the East Coast and enlisted in the company in 1860. O'Neal served until 1874, when he was named superintendent of the National Cemetery at Willets Point, New York. Corporal Benjamin Browne was a shoemaker from Orange County, New York. In August 1862, he enlisted in the newly formed 124th New York Volunteers, and was one of sixty-one men who transferred into Company A that October. Browne reenlisted in the company in February 1864 while it was camped at Brandy Station. He returned to civilian life after completing his enlistment in 1867.

On 10 June, Artificer William Collins was promoted to corporal, probably to fill a combat leadership void for the crossing of the James River. Collins's service record does not reflect the qualities he demonstrated to earn this promotion, but he had obviously overcome some serious problems. Born in Sacketts Harbor, New York, Collins enlisted in the company in December 1853 at the age of twenty-two, and reenlisted in December 1858. Just a month before the war began in April 1861, Collins went "over the hill" and was not caught until three years later. Returned to the company on 22 March 1864, he was tried

GPT-2 completion

in June of that year. Sentenced to be shot for desertion, he remained on active duty.

Captain James J. Dall was born in Virginia in 1829 during the Revolution and joined his father in the Army in 1836. He served with the 2d Virginia Infantry until 1845, when he went to the Western Reserve Regiment; and, on his discharge, he reenlisted with Company I, 25th New York Volunteers (which in 1845 numbered about 4,500 men). One of his brothers, James, served with his company in the same capacity in the same brigade. While the regiment remained in Connecticut through 1846, Captain Dall served with Company I at various times before his discharge. A veteran of the Mexican War, he had

Conditioned on **large** piece of text

- model can generate new content

Image source: "Language Models are Unsupervised Multitask Learners" by Radford et al.
https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

what can we do with deep models?

Text Generation
Fine Tuning
One-shot Learning

What else can we do with them?

Transformers allow for **transfer learning** with text

- model comes **pre-trained** on very large quantities of data
- **fine-tune** model on specific task for which little data is usually available

Transfer learning for text can lead to:

- big **improvements in performance**
  - get better performance from **small training corpus**
  - since model leverages unsupervised pretraining to learn language model
- **multi-linguality:**
  - Multilingual BERT has been pretrained on 104 languages!
  - **train** classification model **on English documents**, use it on Italian ones!
    - transfer learning across languages was technically possible but practically unheard of before
    - opens up amazing possibilities for resource poor languages, or leveraging larger datasets

text classification meets BERT

Mark Carman
20.10.2021

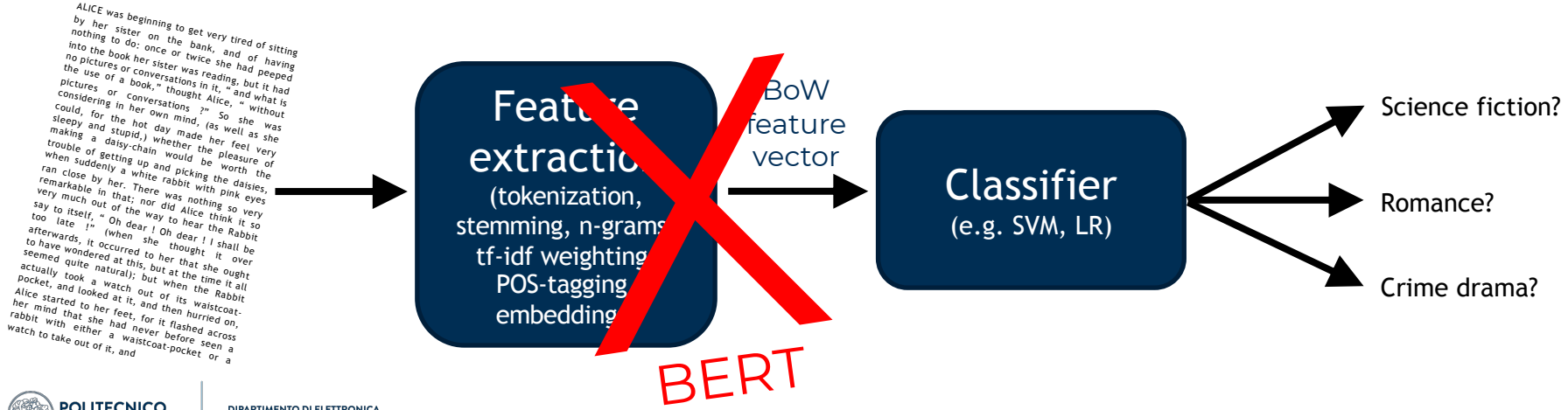
Traditionally, to build a text classifier:

- first decide on types of features to extract from text (e.g. n-gram counts)
- and how to process them (e.g. stemming, idf-weighting, add PoS tags, etc.)

BERT removes the feature extraction step

Moreover, performance improvements likely over count-based features

- since BERT leverages unsupervised pre-training (language modelling)
- and doesn't discard word order



Why would I want to classify text?

Improvements in text classification is a big deal

- because it is an **extremely common** task to need to perform
  - Email spam detection
  - Authorship identification
  - Sentiment analysis in product reviews
  - Offensive content detection
  - Web search query intent identification
  - Creating your news feed on Facebook/LinkedIn
  - Identifying criminal behaviour online (fraud, grooming,..)
  - Routing communications to the right person
  - Parsing requests to spoken interfaces (Alexa, Siri, ...)
  - ...
- in fact text classifiers control much of the content you see online

So how does it work in practice?

Download large **pre-trained language model**

- BERT model choose type
  - lowercase?
  - multilingual?
  - how big?

Fine tune the model for your task on your labelled training data

- spend time **tuning the learning rate**
  - to make sure your model is learning and doesn't collapse or overfit training data
- want to try it out yourself?
  - Open this Google Colab page, and create a copy: https://drive.google.com/file/d/19UcKYpcQeuZB3_T_D1QzhaltuXbpv_VH/view?usp=sharing

So what's the catch?

Compared to training simple text classifier e.g. Logistic Regression or an SVM

Cons:

- hard **limit on length of text**
  - usually to be less than 1000 tokens due all pairwise comparisons being performed
  - Often need to break the text into smaller chunks
- **need fast hardware** to train model
  - i.e. GPUs, typically not available on laptops
- takes **much longer** and requires **more effort** to train model
- model will be big
  - which means it will require more memory
  - and may be **slower** when making predictions
- predictions are **less interpretable**
  - although techniques exist to try to explain the predictions (e.g. LIME)

But also **don't need to do any feature engineering**!

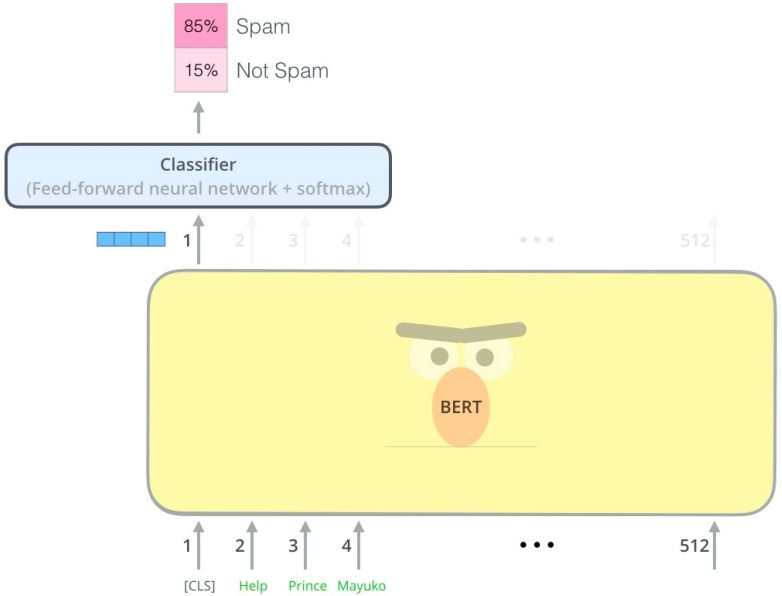
  - e.g. choose whether to run stemming, whether to use n-grams, etc.

How is BERT fine-tuned?

Mark Carman
20.10.2021

Add special **[CLS]** token to start of text

- instead of outputting a word in that position
- model is trained to produce the class label



Source: http://jalammar.github.io/illustrated-bert/

Supervised learning BERT

Mark Carman
20.10.2021

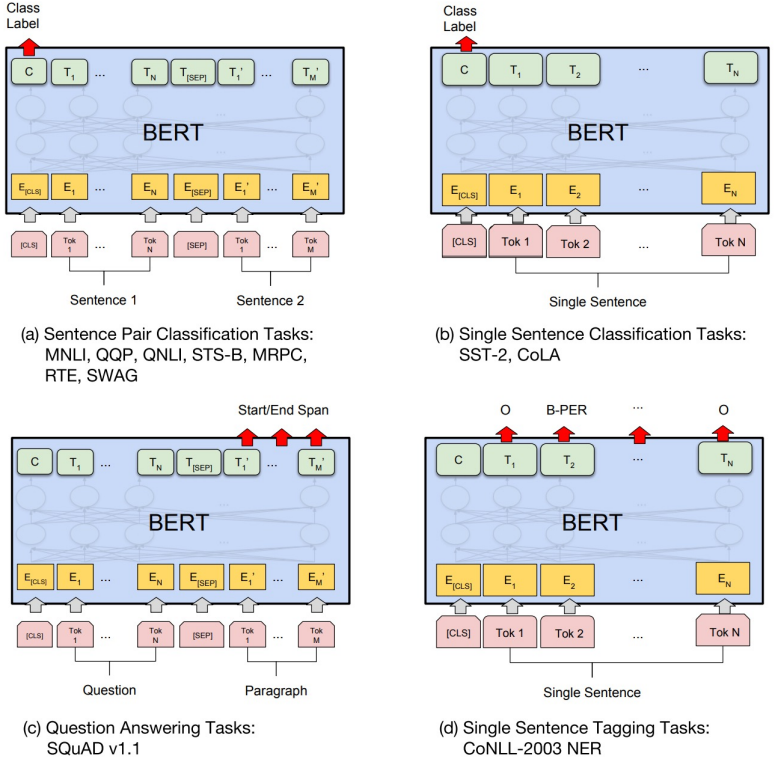
Language models are very flexible!

By simply adding two special tokens:

- [CLS] – the class
- [SEP] – separator

BERT can be fine-tuned for a large number of tasks:

- single text classification
- text pair classification
- question answering
- sequence labelling



Source: https://arxiv.org/pdf/1810.04805.pdf

What about sequence labeling?

BERT can also be fine-tuned on sequence labelling tasks

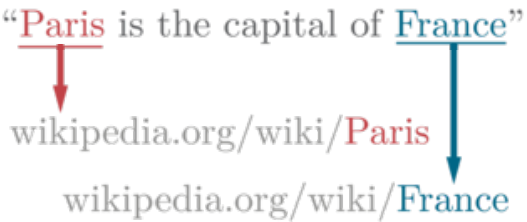
- E.g. **named-entity recognition**
  - task of identifying entities that are mentioned in a text
  - often a first step in extracting knowledge from text

"Have you heard of an associate professor from the Politecnico di Milano called Mark Carman?"

Institution Person

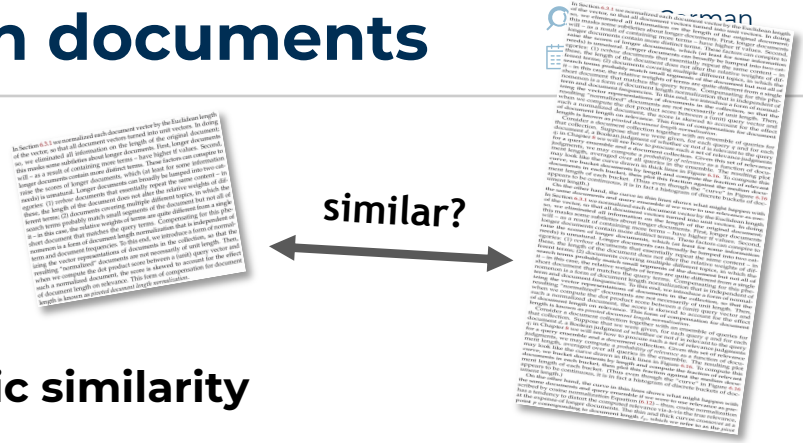
- and **entity-linkage**
  - more complicated problem of determining **which real-word entity** was referred to
  - often not as easy as it sounds …

I grew up in a small town just out of Paris.
Currently driving from Dallas to Paris.
Paris Hilton was photographed leaving the Paris Hilton.



Source: https://commons.wikimedia.org/wiki/File:Entity_Linking_-_Short_Example.png

Measuring similarity between documents



BERT can also be trained to estimate **semantic similarity** between documents

- Given the context length restriction (1000 tokens) usually people compare sentences or paragraphs and then aggregate to measures of document similarity

Why do we need a similarity measure between documents?

- clustering
- Web Search!

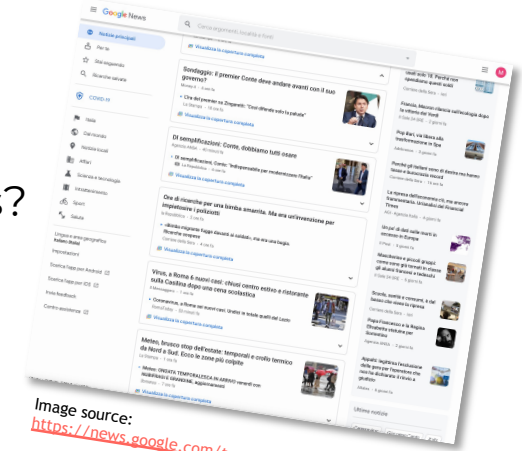


Image source: https://news.google.com/topstories?hl=it&gl=IT&ceid=IT:it

Fine tuning GPT-2 ...

Mark Carman
20.10.2021

GPT-2 can also be used as a text encoder for classification tasks, but the strength of GPT-2 is **text generation**

- so makes sense to use it for tasks such as **translation**, **summarisation**, **dialog**, etc.

During fine-tuning

- introduce special tokens to separate input from output
- and to indicate the type of output required

Training Dataset

I	am	a	student	<to-fr>	je	suis	étudiant
let	them	eat	cake	<to-fr>	Qu'ils	mangent	de
good	morning	<to-fr>	Bonjour				

Training Dataset

Article #1 tokens		<summarize>	Article #1 Summary
Article #2 tokens	<summarize>	Article #2 Summary	padding
Article #3 tokens		<summarize>	Article #3 Summary

Source: http://jalammar.github.io/illustrated-gpt2/

what can we do with deep models?

Text Generation
Fine Tuning
One-shot Learning

GPT-2 can be used even without fine-tuning

Mark Carman
20.10.2021

Language models are universal learners. Predicting text is flexible method for providing all sorts of functionality:

- **translation**:
  - in the context, give multiple strings of the form: english sentence = french sentence
  - then prompt with: english sentence = ?
- **question answering**:
  - prompt the model with the question
- **reading comprehension**:
  - give text and examples of questions with answers,
  - then prompt with unanswered question
- **summarization**:
  - Provide content to be summarised and prefix response with "**tl;dr:**"

Image source:
https://commons.wikimedia.org/wiki/File:Swiss_Army_Knife.svg

Examples: question answering

Mark Carman
20.10.2021

Language model can learn facts, and answer questions!

Most confident predictions from LM are quite impressive

- Not as reliable as a standard (IR based) question answering system (yet)
- But the system **has not been trained** to do this!

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%

Image source: “Language Models are Unsupervised Multitask Learners” by Radford et al.
https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

GPT-2 examples: translation

Mark Carman
20.10.2021

Not the best translator out there ;-)

- BUT the system was not trained to do translation!

Moreover, it was only trained on an ENGLISH corpus

- So how could it learn to “speak” French?

English reference	**GPT-2 French translation**
One man explained that the free hernia surgery he’d received will allow him to work again.	Un homme expliquait que le fonctionnement de la hernia fonctionnelle qu’il avait reconnaît avant de faire, le fonctionnement de la hernia fonctionnelle que j’ai réussi, j’ai réussi.
French reference	**GPT-2 English translation**
Un homme a expliqué que l’opération gratuite qu’il avait subie pour soigner une hernie lui permettrait de travailler à nouveau.	A man told me that the operation gratuity he had been promised would not allow him to travel.

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

GPT-2 examples: reading comprehension

Mark Carman
20.10.2021

Context (passage and previous question/answer pairs)

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

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

Q: What was the theme
A: "one world, one dream".

Q: What was the length of the race?
A: 137,000 km

Q: Was it larger than previous one? **Model answer**: Everest
A: No

Q: Where did the race begin?
A: Olympia, Greece

Q: Is there anything notable about that place?
A: birthplace of Olympic Games

Q: Where did they go after?
A: Athens

Q: How many days was the race?
A: seven

Q: Did they visit any notable landmarks?
A: Panathinaiko Stadium

Q: And did they climb any mountains?
A:

Research Applications

Want to try some **demos**?
Go to: http://131.175.120.138:61111/

Where do we use transformers?

Mark Carman
20.10.2021

Here are some research problems we're working on:

- **Text Analysis for Bioinformatics**
  - translating text into structured data to facilitate search
  - model protein sequences for drug repurposing
- **Text Analysis for Political Discourse**
  - identifying claims & finding evidence online
  - detecting fake and misleading news
- **Text meets Images/Video**
  - visual question answering
  - explaining AI with text explanations

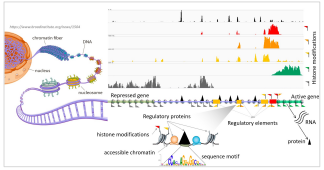


Image source: https://www.symanse.org/Portal.html#Symanse:con17083707/wiki/588600

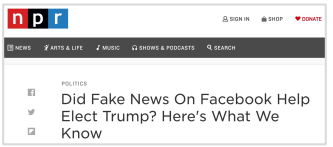
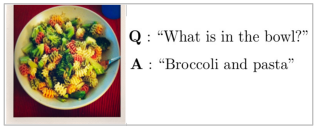


Image source
https://www.npr.org/2018/04/11/601323723/6-facts-we-know-about-fake-news-in-the-2016-election



applications: data integration

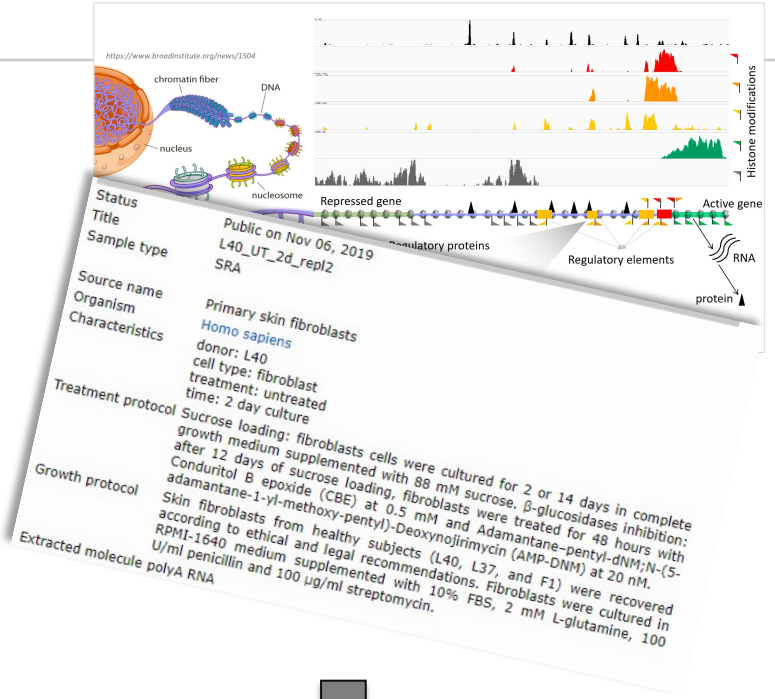
Mark Carman

LOTS of genomics data being generated by research groups around the world

- data is being aggregated in large repositories
- no agreed format for biologists describe their experiments
- much of **meta-data is free text**

Application:

- train translation models to **automatically extract database fields** from textual descriptions
- so bioinformaticians can easily find the genomic data they need for their analysis

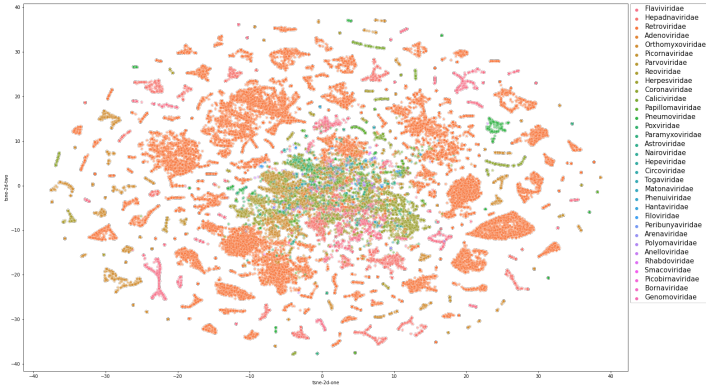


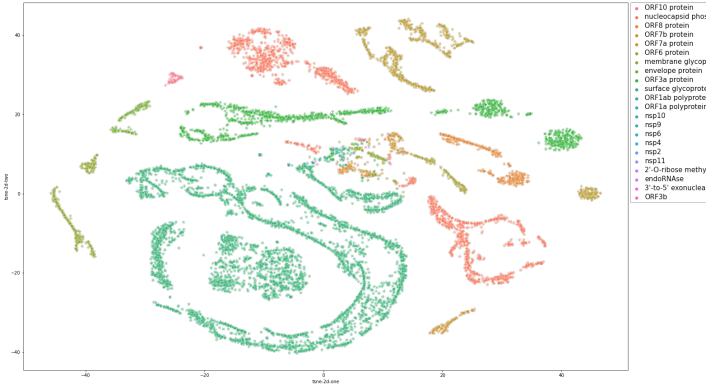
Species	*Cell Line*	*Disease*	*Factor*	*Tissue*	*...*
Homo sapiens	*KARPAS-422*	*B Cell Lymphoma*	*H327me3*	*-*	*...*

applications: protein sequence modeling

Mark Carman
20.10.2021

- Learning BERT embeddings of protein sequences across virus families to see if they are useful for drug-protein interaction prediction.





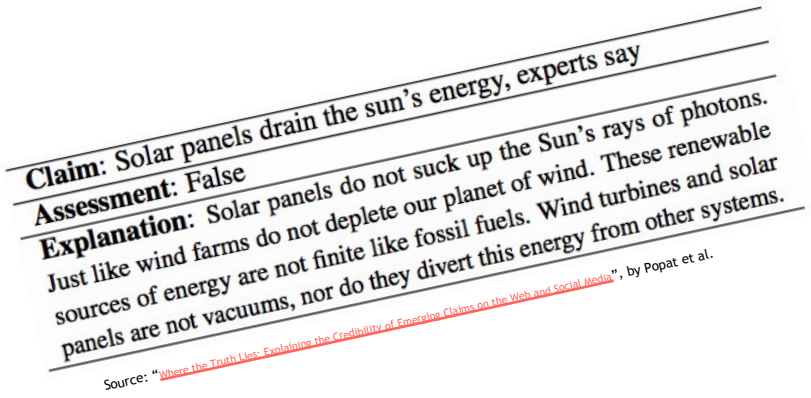
applications: claim detection & verification

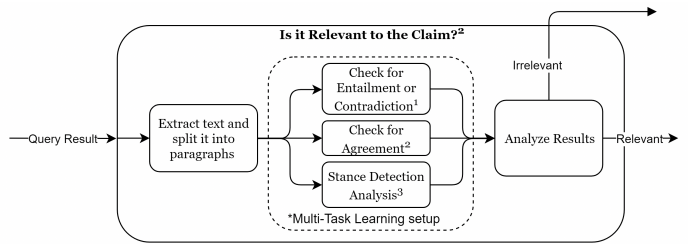
Mark Carman
20.10.2021

We make use of pretrained deep language models for
- detecting claims
- ranking evidence

Advantages:
- better performance
- multilinguality
  - train model on English text
  - apply to Arabic text

Applications:
- detecting fake news
- counteracting hate speech online
  - finding evidence for disputing





applications: image question answering

Mark Carman
20.10.2021

Deep learning is famous for build embeddings over images

- Why not combine the power of images and text embeddings together?
- To answer questions about images:

Conclusions

Conclusions

Language models becoming really powerful!
- able to model **long range dependencies**
- and scale learning to **billions of parameters**

Language models are generic learners
- plug-in component of larger systems
- pretrained on entire web, multi-lingual

Implications:
- **natural language interfaces** (e.g. Alexa, Siri) will get **better & better**
- **detecting fakes** (e.g. news, spam, assignments) will get **harder & harder**
- **personalisation & search** (e.g. movies, clothes) will get **better & more sophisticated**
- **mining knowledge** from text (e.g. patents, emails) will get **easier & easier**
- ...