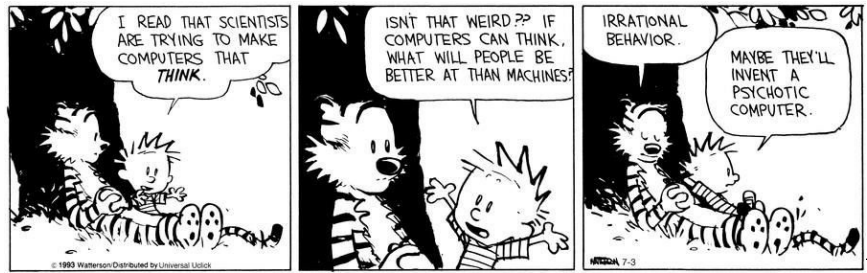
BERTology review: What Does BERT Look At? An Analysis of BERTs attention

Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher Manning (2019)

Sualeh Asif

MIT Computer Science Artificial Intelligence Laboratory
Massachusetts Institute of Technology



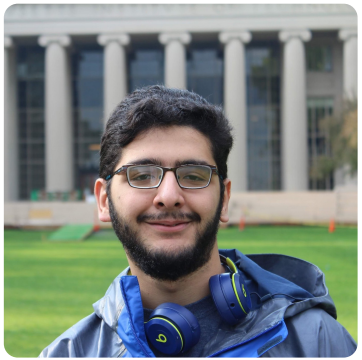
About Me:

Sualeh Asif

MIT Computer Science
Math
Theatre

Interests:

Machine Learning
- Representations
- Emergence
- Visual
- NLP

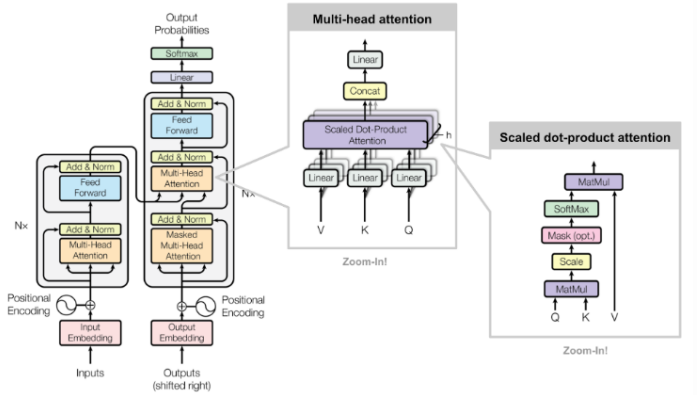


Sualeh@mit.edu
617) 599 5080

Performance Engineering
- CPUs + GPUs

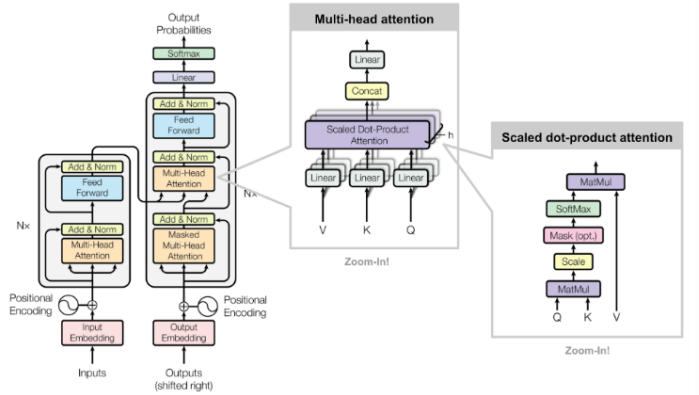
Computational Number Theory

BERTology ....


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Multi-head attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
Zoom-In!
Scaled dot-product attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Zoom-In!

BERTology ....

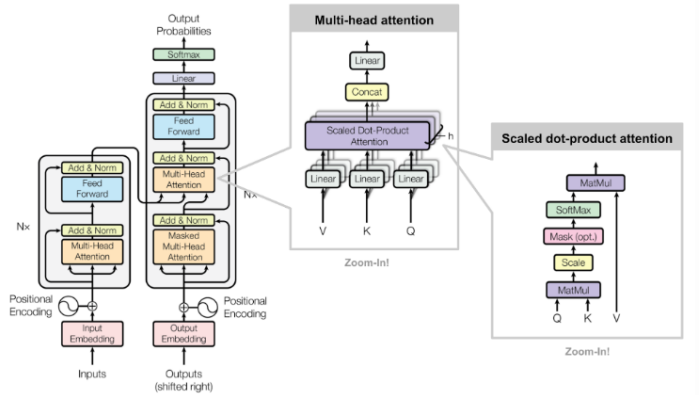
But ... why?



BERTology ....

But ... why?

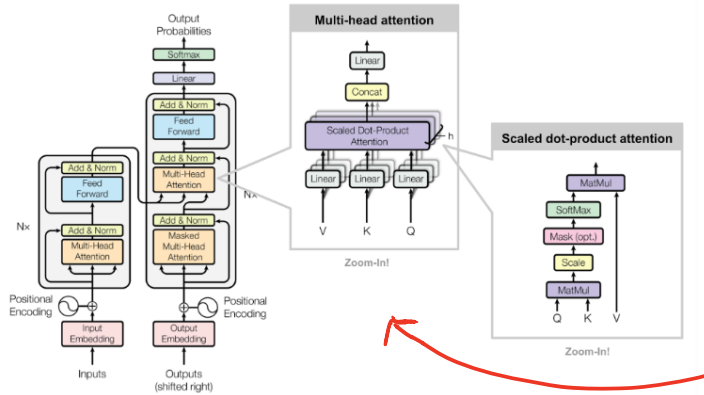
* what is BERT and how did we get there? (short history)



BERTology ....

But ... why?

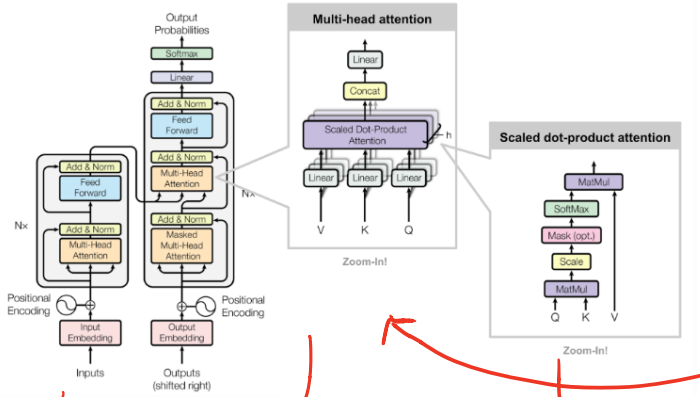
* what is BERT and how did we get there?
* what does BERT teach us?



BERTology ....

But ... why?

* what is BERT and how did we get there?
* what does BERT teach us?



Architecture → Attention → Linguistics

tinkering ...

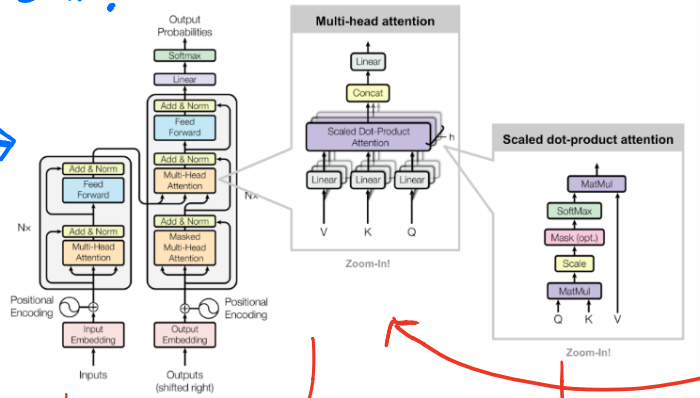
BERTology ....

But ... why?

Framework:

What we bake in?

* what is BERT and how did we get there?
* what does BERT teach us?



What emerges?

Architecture → Attention → Linguistics

tinkering ...

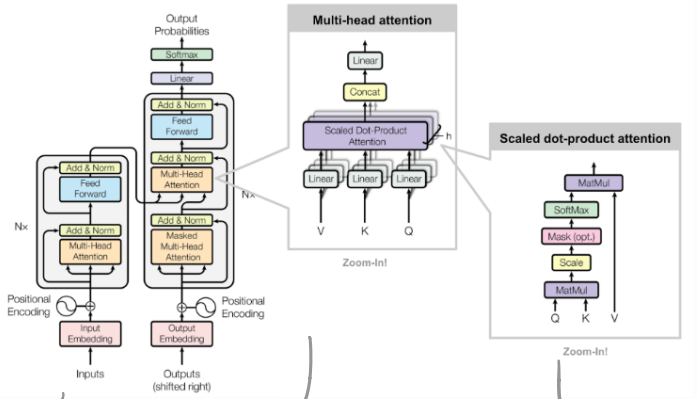
BERTology ....

But ... why?

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?

... do our lessons generalise?

- NLP
- vision
- learning systems



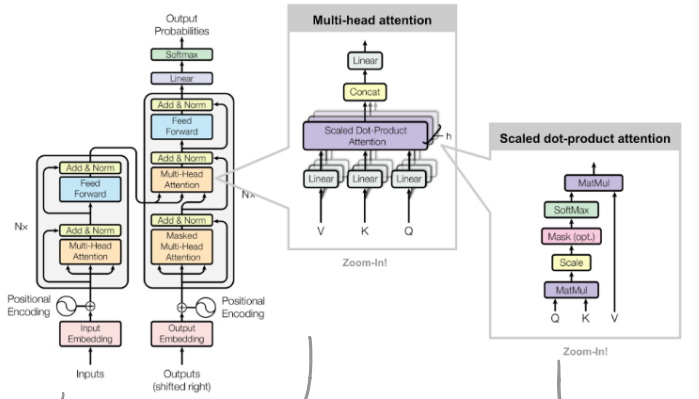
Architecture → Attention → Linguistics

tinkering ...

BERTology ....

But ... why?

* What is BERT and how did we get there?
* What does BERT teach us?
* What could come next?
  ... do our lessons generalise?



Architecture → Attention → Linguistics

tinkering ...

* BECAUSE WE ARE CURIOUS HUMANS!

Running Example

(and my programming task)

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

Restraunt_reviews.txt

```
The O           Food    B-ASP
bread   B-ASP   is  O
is  O           always  O
top B-OP        fresh   B-OP
notch   I-OP    and O
as  O           hot B-OP
well    O       -   O
.   O           ready   O
                to  O
                eat O
                !   O
```

Running Example

(and my programming task)

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

Restraunt_reviews.txt

The	O	Food	B-ASP
bread	B-ASP	is	O
is	O	always	O
top	B-OP	fresh	B-OP
notch	I-OP	and	O
as	O	hot	B-OP
well	O	-	O
.	O	ready	O
		to	O
		eat	O
		!	O

1485 sentences

tokenized, embedded with bert

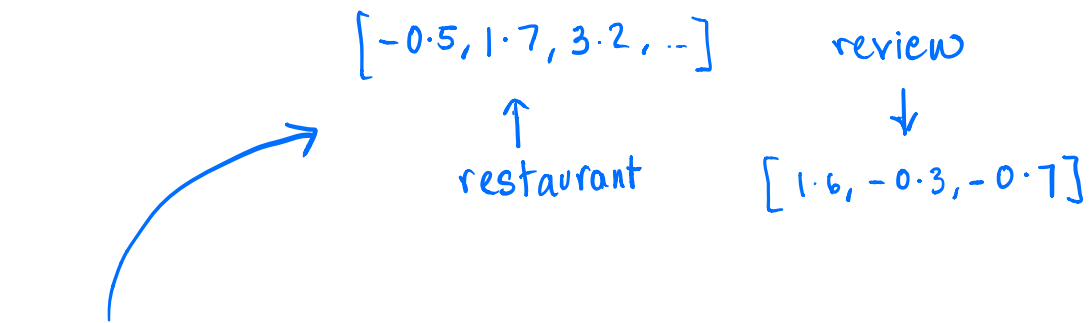
(almost) all data in the presentation comes from this dataset!

A BERT RECAP


* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
... do our lessons generalise?

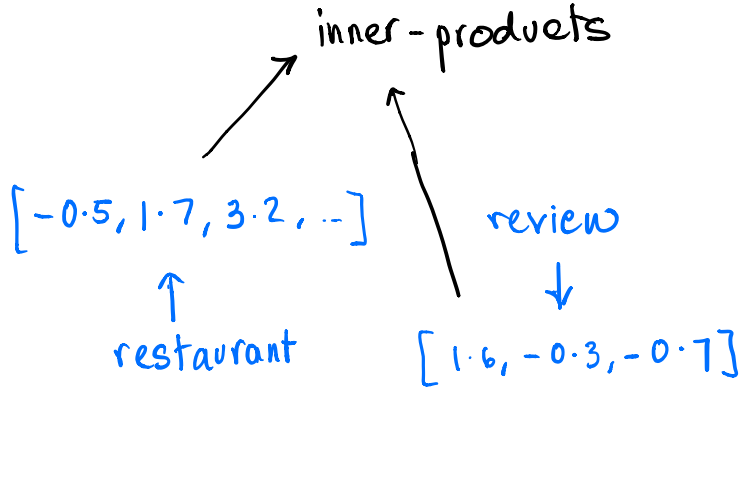
A BERT RECAP

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?



Word Embeddings (word2vec, 2013; GLOVE, 2014)

A BERT RECAP



Word Embeddings (word2vec, 2013 / GLOVE, 2014) + Pre-training

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

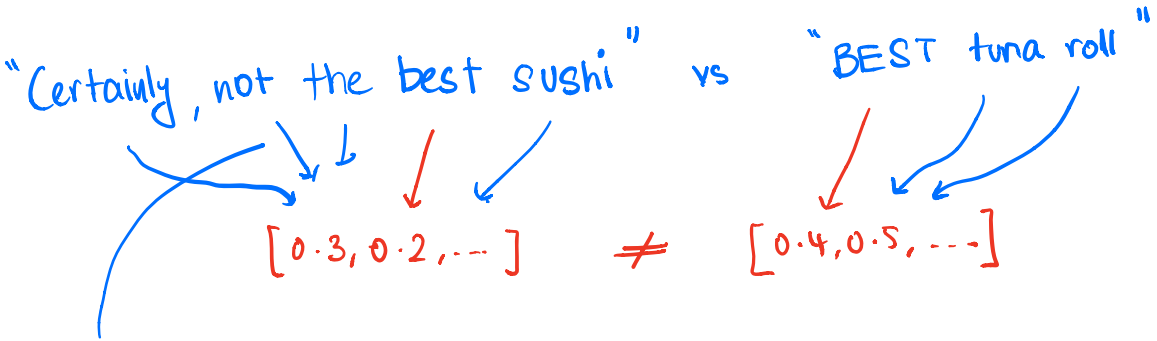
A BERT RECAP

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

Context & Representations (2010-2014)

Word Embeddings (word2vec, 2013 / GLOVE, 2014) + Pre-training

A BERT RECAP



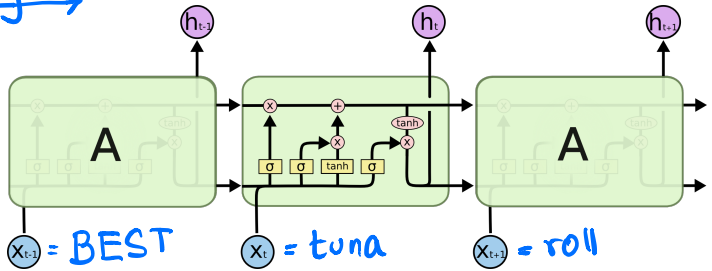
Context & Representations (2010-2014)

Word Embeddings (word2vec, 2013 / GLOVE, 2014) + Pre-training

- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

A BERT RECAP

directionality →



" Large-scale" Supervision & LSTMs (Google, 2015)

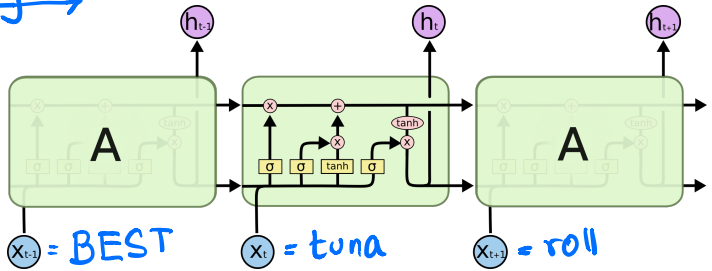
Context & Representations (2010-2014)

Word Embeddings (word2vec, 2013 GLOVE, 2014) + Pre-training

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

A BERT RECAP

directionality →



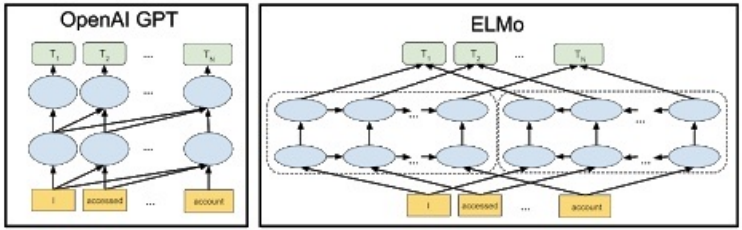
" Large-scale " Supervision & LSTMs (Google, 2015)

Context & Representations (2010-2014)

Word Embeddings (word2vec, 2013 / GLOVE, 2014) + Pre-training

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

A BERT RECAP



Bidirection " L ⟷ R " (ELMo) + Generative Pre-Training

" Large-scale" Supervision & LSTMs (Google, 2015)

Context & Representations (2010-2014)

Word Embeddings (word2vec, 2013 / GLOVE, 2014) + Pre-training

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

A BERT RECAP

"Transformers"

Multi-headed "Attention" mechanisms (Vaswani et.al 2017)

Bidirection "L ⟷ R" (ELMo) + Generative Pre-Training

"Large-scale" Supervision & LSTMs (Google, 2015)

Context & Representations (2010-2014)

Word Embeddings (word2vec, 2013 / GLOVE, 2014) + Pre-training

---

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

A BERT RECAP

BERT = Bidirectional Encoder Representations from Transformers

- Multi-headed "Attention" mechanisms (Vaswani et.al 2017)
- Bidirection" L ⟷ R "(ELMo) + Generative Pre-Training
- "Large-scale" Supervision & LSTMs (Google, 2015)
- Context & Representations (2010-2014)
- Word Embeddings (word2vec, 2013 / GLOVE, 2014) + Pre-training

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

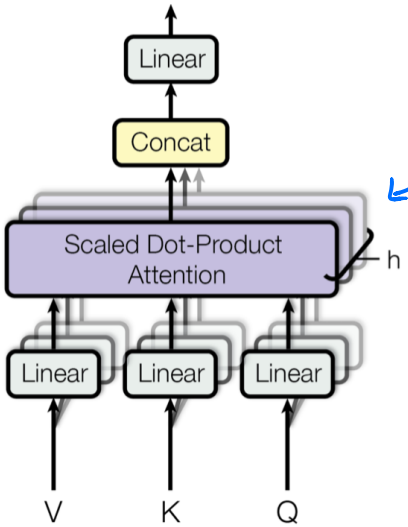
A BERT RECAP

Multi-headed "Attention" mechanisms (Vaswani et.al 2017)

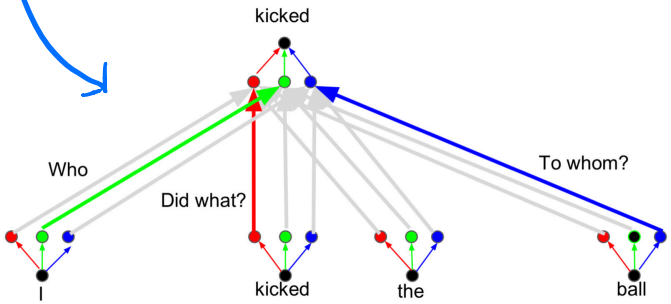
* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

A BERT RECAP

Multi-headed "Attention" mechanisms (Vaswani et.al 2017)



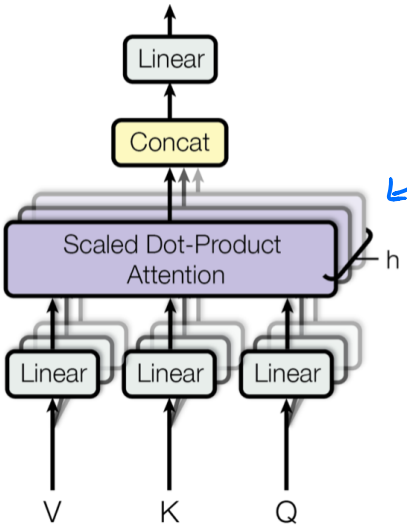
Canonical Views



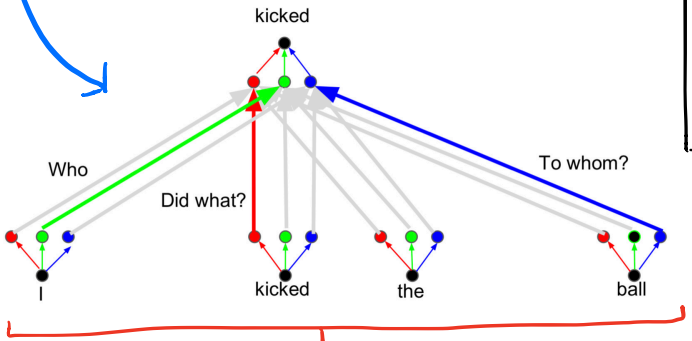
* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

A BERT RECAP

Multi-headed "Attention" mechanisms (Vaswani et.al 2017)



Canonical Views



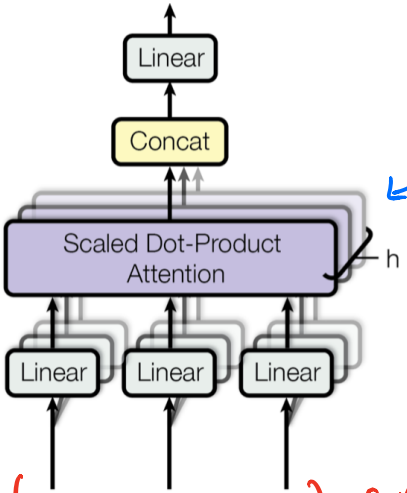
attention is a weighted average

"fancy"

- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

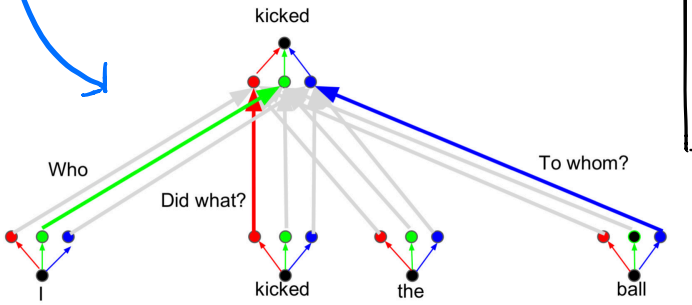
A BERT RECAP

Multi-headed "Attention" mechanisms (Vaswani et.al 2017)



$$A(V, K, Q) = s.m\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Canonical Views



attention is a weighted average

↓

"fancy"

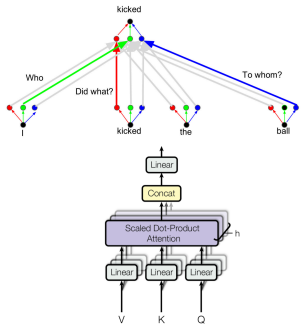
- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

A BERT RECAP

Multi-headed "Attention" mechanisms (Vaswani et.al 2017)

But what do we bake in?

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?



A BERT RECAP

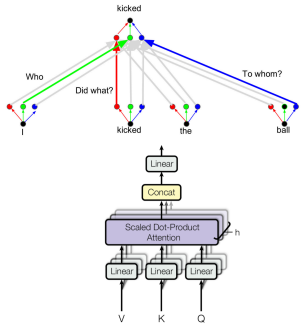
Multi-headed "Attention" mechanisms (Vaswani et.al 2017)

But what do we bake in?

* well-constructed "context"
* separate learning capacity
  .... allows composition!

but ... warning!!

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?



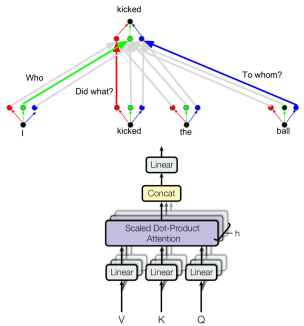
A BERT RECAP

Multi-headed "Attention" mechanisms (Vaswani et.al 2017)

But what do we bake in?

* well-constructed "context"
* separate learning capacity
  .... allows composition!
* curious to hear what you think ??

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?



* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Clark, Khandelwal, Levy, Manning

An Analysis Of Berts Attention

But Why?

* Language modeling is the "ultimate" task
  ... so we want to understand how BERT models language.

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

But Why?

* Language modeling is the "ultimate" task
  ... so we want to understand how BERT models language.
* Our understanding of attention could better guide what we bake in future models.

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

But Why?

* Language modeling is the "ultimate" task
  ... so we want to understand how BERT models language.
* Our understanding of attention could better guide what we bake in future models.
* Curiosity: we don't fully understand if BERT understands the "structure" of language? What about features?

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

But How?

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

But How?

* Syntactic trees. (Lin et.al.)
* Parts of speech/syntax (Tenney/Liu et.al.)
* Subject - Predicate Agreement. (Goldberg et.al.)
* Numerology (Wallace et.al.)
* Knowledge bases/Reasoning (Petroni et.al.)

spoiler: BERT is not great at this!

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

* Forward atth. vs Backward atth.

  ⟹ ~50% next token, ~50% prev. token

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?

  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

* Forward attn. vs Backward attn?
  ⟹ ~50% next token, ~50% prev. token
* What about specific tokens?
  ⟹ substantial amount of attn. to [SEP], [CLS]

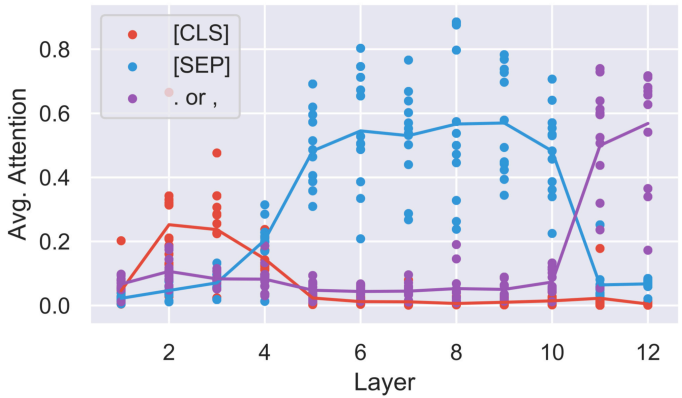
Reminder:
[CLS] <sentence> [SEP]

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

* Forward attn. vs Backward attn?
  ⟹ ~50% next token, ~50% prev. token
* What about specific tokens?
  ⟹ substantial amount of attn. to [SEP], [CLS]

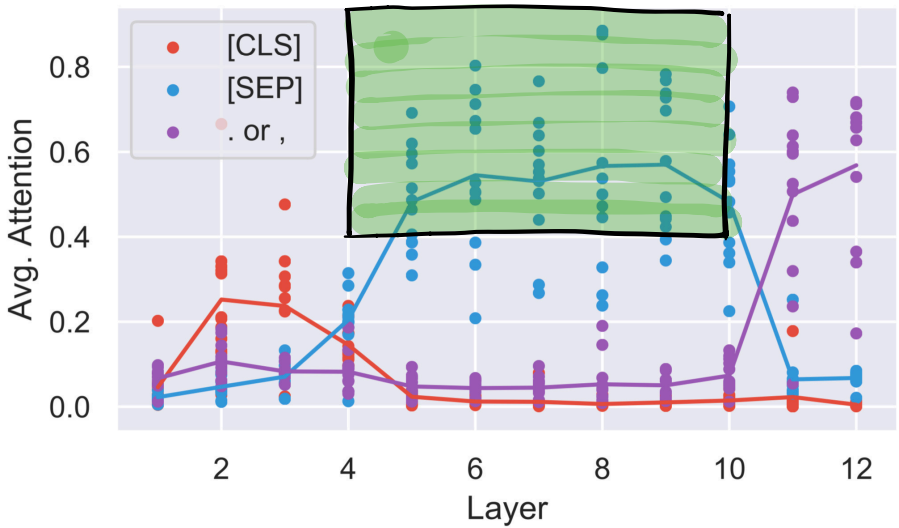


Reminder:
[CLS] <sentence> [SEP]

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?



- What is BERT and how did we get there?
- What does BERT teach us?
- What could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

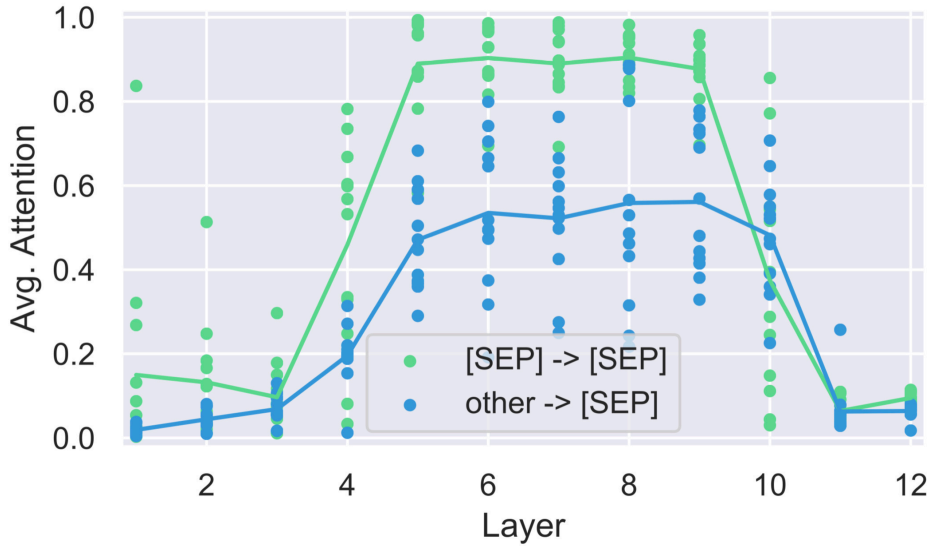
What is on the surface?

* Forward attn. vs Backward attn?
  ⟹ ~50% next token, ~50% prev. token
* What about specific tokens?
  ⟹ substantial amount of attn. to [SEP], [CLS]
  hmm ... why?
  * Maybe, "aggregate surface-level information"?

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?



- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

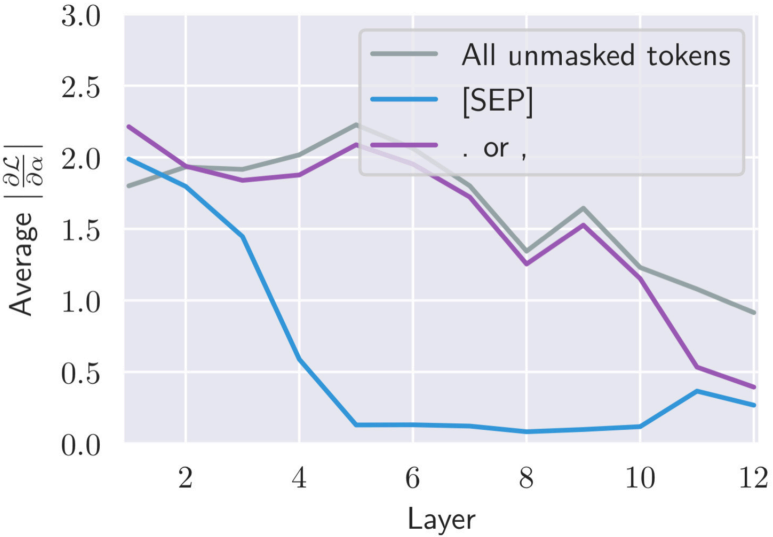
What is on the surface?

* Forward attn. vs Backward attn?
  ⟹ ~50% next token, ~50% prev. token
* What about specific tokens?
  ⟹ substantial amount of attn. to [SEP], [CLS]
  hmm ... why?
  ~~* Maybe, "aggregate surface-level information"?~~
  * Maybe, its just a "NO-OP"

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?



- What is Bert and how did we get there?
- What does Bert teach us?
- What could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

* Forward attn. vs Backward attn?
  ⟹ ~50% next token, ~50% prev. token
* What about specific tokens?
  ⟹ substantial amount of attn. to [SEP], [CLS]
  hmm ... why?
  ~~* Maybe, "aggregate surface-level information"?~~
  * Maybe, its just a "NO-OP"

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

* Forward attn. vs Backward attn?
  ⟹ ~50% next token, ~50% prev. token
* What about specific tokens?
  ⟹ substantial amount of attn. to [SEP], [CLS]
* Focused vs. Broad Attn.

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

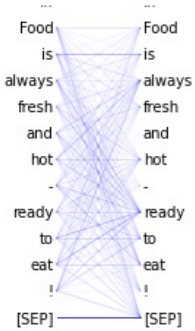
* Forward attn. vs Backward attn?
  ⟹ ~50% next token, ~50% prev. token
* What about specific tokens?
  ⟹ substantial amount of attn. to [SEP], [CLS]
* Focused vs. Broad Attn.
  ⟹ lower layers: "broad" attention ← high entropy.
  higher layers: "focused" attention

* what is Bert and how did we get there?
* what does Bert teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

(0,0)

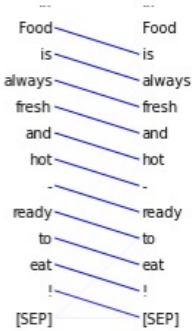


"broad" attention

(2,1)

"prev" word

(2,3)



"next" word

(10,5)

exclusive [SEP] focus.

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?

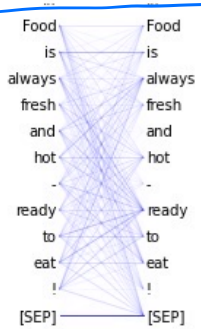
... do our lessons generalise?

An Analysis Of Berts Attention

What is on the surface?

specificity increases with depth!

(0,0)



"broad" attention

(2,1)

"next" word

"prev" word

(2,3)

(10,5)

exclusive [SEP] focus.

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

Idea: A word "predicts" another word if it gets most of its attention.

Do "predections" learn language features?

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?

... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

Idea: A word "predicts" another word if it gets most of its attention.

Do "predections" learn language features?

Evaluation: Check accuracy with Penn Tree-bank annotated with Stanford Dependencies!

⇒ no single head does well overall. but they do very well with "composition"

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

Relation	Head	Accuracy	Baseline
All	7-6	34.5	26.3 (1)
prep	7-4	66.7	61.8 (-1)
pobj	9-6	**76.3**	34.6 (-2)
det	8-11	**94.3**	51.7 (1)
nn	4-10	70.4	70.2 (1)
nsubj	8-2	58.5	45.5 (1)
amod	4-10	75.6	68.3 (1)
dobj	8-10	**86.8**	40.0 (-2)
advmod	7-6	48.8	40.2 (1)
aux	4-10	81.1	71.5 (1)
poss	7-6	**80.5**	47.7 (1)
auxpass	4-10	**82.5**	40.5 (1)
ccomp	8-1	**48.8**	12.4 (-2)
mark	8-2	**50.7**	14.5 (2)
prt	6-7	**99.1**	91.4 (-1)

common relations:
- prepositions
- noun subjects
- adjective modifiers

Best head!

surprising difference.

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

But ... what exactly do these look like ?

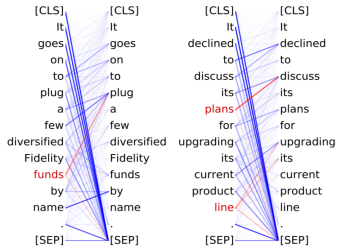
* what is BERT and how did we get there ?
* what does BERT teach us ?
* what could come next ?
  ... do our lessons generalise ?

An Analysis Of Berts Attention

Learning Features Of Language

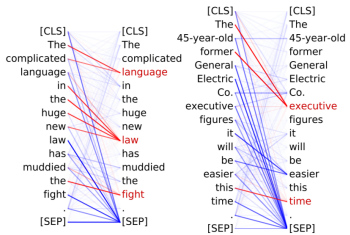
Head 8-10

- **Direct objects** attend to their verbs
- 86.8% accuracy at the dobj relation



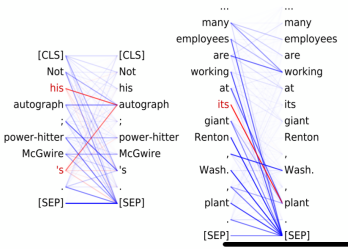
Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the det relation



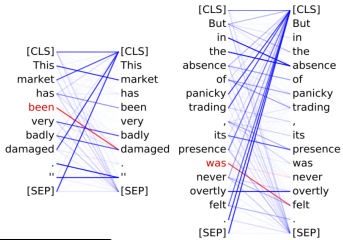
Head 7-6

- **Possessive pronouns** and apostrophes attend to the head of the corresponding NP
- 80.5% accuracy at the poss relation



Head 4-10

- **Passive auxiliary verbs** attend to the verb they modify
- 82.5% accuracy at the auxpass relation



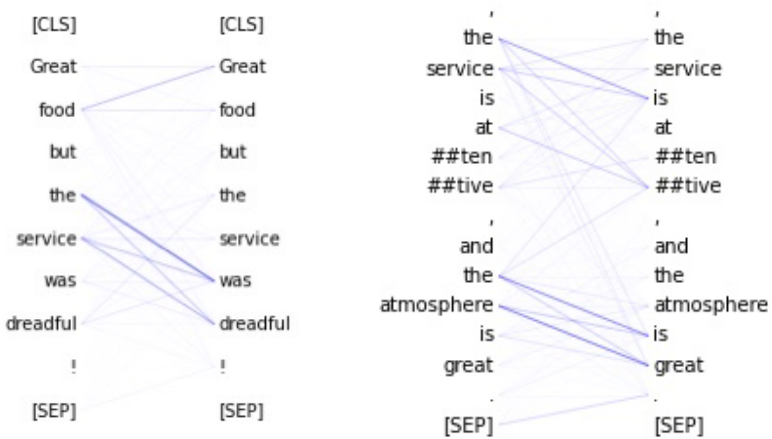
* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?

... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

amodj: adjectival modifiers



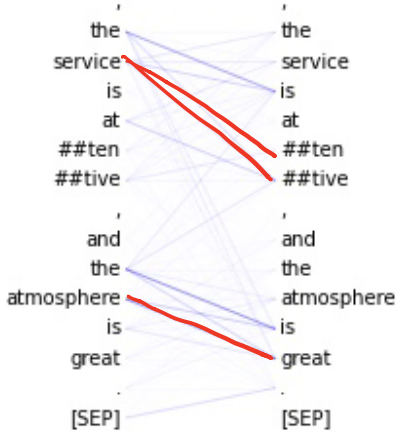
- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

amodj: adjectival modifiers

(8,4)



- What is Bert and how did we get there?
- What does Bert teach us?
- What could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

nmodj: nominal modifiers

- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

nmodj: nominal modifiers

(9,7)

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

early "broad" focused heads.

(0,0)

- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Learning Features Of Language

How far can we push extracting dependencies?

Idea: probing heads

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Similarity and Composition

What kind of similarities do heads have?

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Similarity and Composition

What kind of similarities do heads have?

Idea: Compute distances ...

$$d(H_i, H_j)$$

plot this on 2d.

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Similarity and Composition

What kind of similarities do heads have?

Idea: Compute

$$d(H_i, H_j) = JS(H_i, H_j)$$

plot this on 2d.

$$JS(p\|q) = \frac{1}{2} D_{KL}\left(p \,\|\, \frac{p+q}{2}\right) + \frac{1}{2} D_{KL}\left(q \,\|\, \frac{p+q}{2}\right)$$

$$D(p\|q) = \int p \log \frac{p}{q} \, dx$$

- what is Bert and how did we get there?
- what does Bert teach us?
- what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Similarity and Composition

What kind of similarities do heads have?

Idea: Compute

$$d(H_i, H_j) = \sum_{\text{tokedata}} JS(H_i(tok), H_j(tok))$$

plot this on 2d.

$$JS(p\|q) = \frac{1}{2} D_{KL}\left(p \,\|\, \frac{p+q}{2}\right) + \frac{1}{2} D_{KL}\left(q \,\|\, \frac{p+q}{2}\right)$$

- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

An Analysis Of Berts Attention

Similarity and Composition

What kind of similarities do heads have?

Idea: Compute $d(H_i, H_j) = \sum_{\text{tokedata}} JS(H_i(tok), H_j(tok))$.

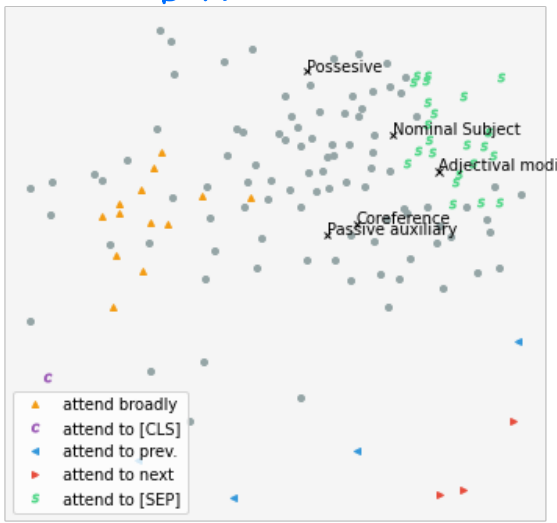
... plot!! (hammer: multi-dimensional scaling)

- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

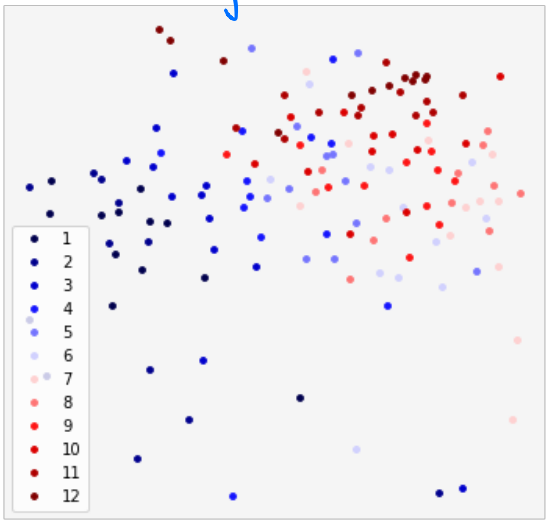
An Analysis Of Berts Attention

Similarity and Composition

behaviour:



layers:

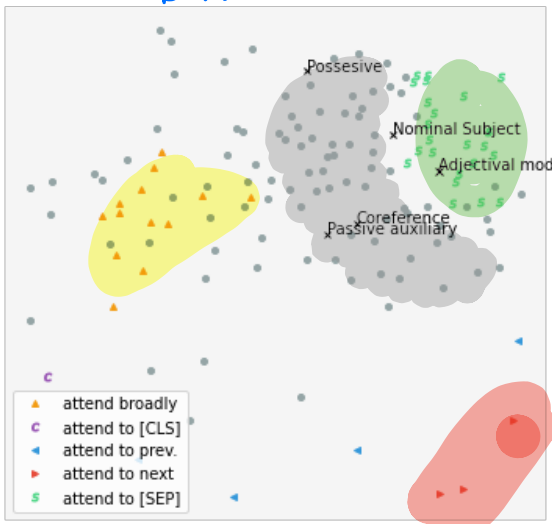


* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

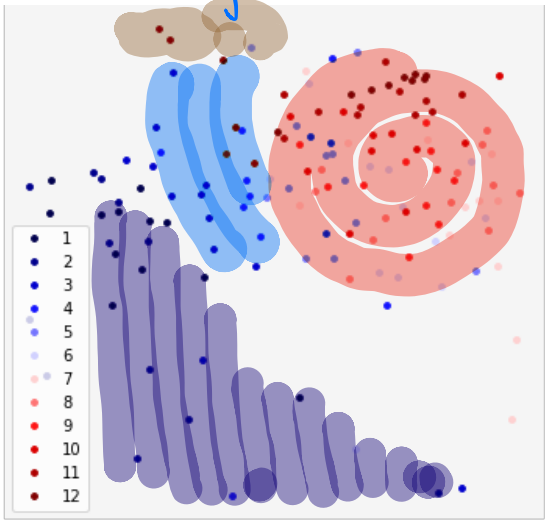
An Analysis Of Berts Attention

Similarity and Composition

behaviour:



layers:



- what is BERT and how did we get there?
- what does BERT teach us?
- what could come next?
  ... do our lessons generalise?

What Comes Next

* what is BERT and how did we get there?
* what does BERT teach us?
* what could come next?
  ... do our lessons generalise?

Yeah ..... so ????

What Comes Next

Yeah ..... so ????

* We have a Toolbox !! to study language nets.

* what is BERT and how did we get there ?
* what does BERT teach us ?
* what could come next ?
  ... do our lessons generalise ?

What Comes Next

Yeah ..... so ????

* We have a Toolbox !! to study language nets.
* "Attention" is more interesting (boring?) than vectors/neurons.
  i.e. turn "our" attention to higher level abstractions.
  think multi-modal, many agents ....

* what is BERT and how did we get there ?
* what does BERT teach us ?
* what could come next ?
  ... do our lessons generalise ?

What Comes Next

Yeah ..... so ????

* We have a Toolbox !! to study language nets.
* "Attention" is more interesting (boring?) than vectors/neurons.
  i.e. turn "our" attention to higher level abstractions.
  think multi-modal, many agents ....
* We shouldn't be afraid of using more computation!
  i.e. there is hope and we should keep trying ...

* what is BERT and how did we get there ?
* what does BERT teach us ?
* what could come next ?
  ... do our lessons generalise ?

What Comes Next

Yeah ..... so ???? ....

* We have a Toolbox !! to study language nets.
* "Attention" is more interesting (boring?) than vectors/neurons.
  i.e. turn "our" attention to higher level abstractions.
  think multi-modal, many agents ....
* We shouldn't be afraid of using more computation!
  i.e. there is hope and we should keep trying ...

What do all of you think ??

* what is BERT and how did we get there ?
* what does BERT teach us ?
* what could come next ?
  ... do our lessons generalise ?

BERTology review: What Does BERT Look At? An Analysis of BERTs attention

Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher Manning (2019)

Sualeh Asif

MIT Computer Science Artificial Intelligence Laboratory
Massachusetts Institute of Technology