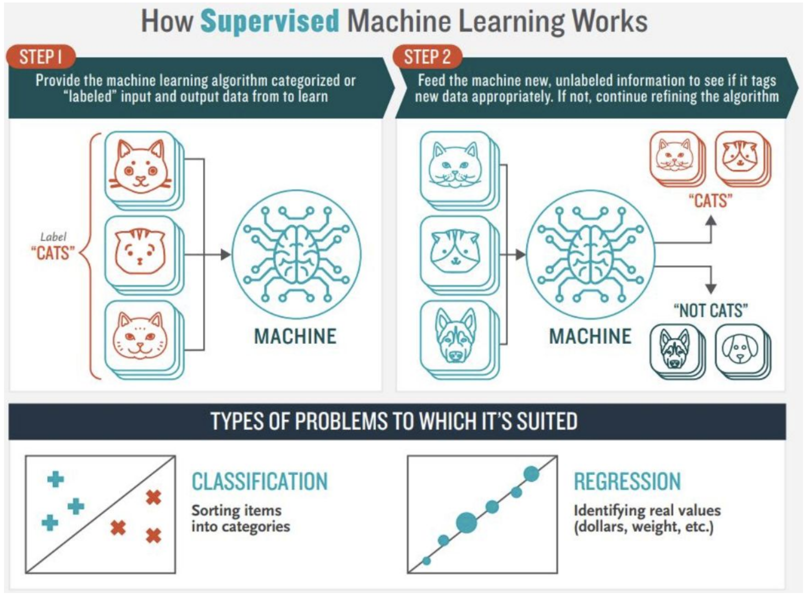
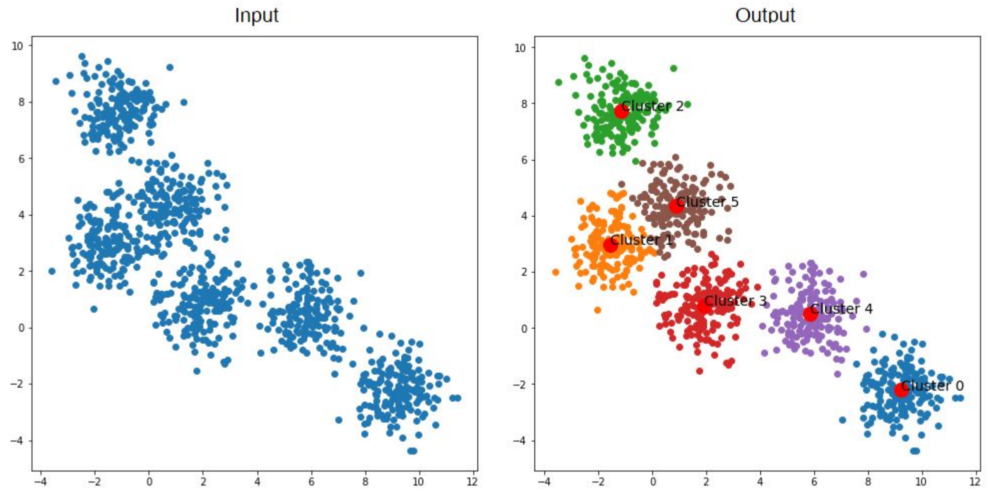
Emergent structures in transformer models trained by self-supervised methods

Supervised learning

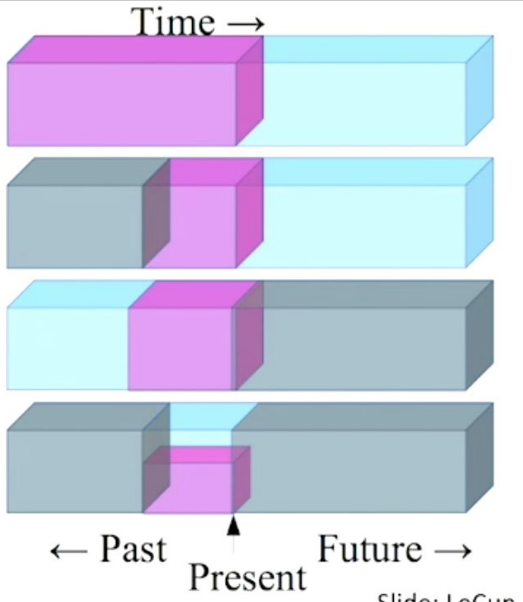

How Supervised Machine Learning Works
STEP 1
Provide the machine learning algorithm categorized or "labeled" input and output data from to learn
STEP 2
Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm
Label "CATS"
MACHINE
MACHINE
"CATS"
"NOT CATS"
TYPES OF PROBLEMS TO WHICH IT'S SUITED
CLASSIFICATION
Sorting items into categories
REGRESSION
Identifying real values (dollars, weight, etc.)

Unsupervised learning



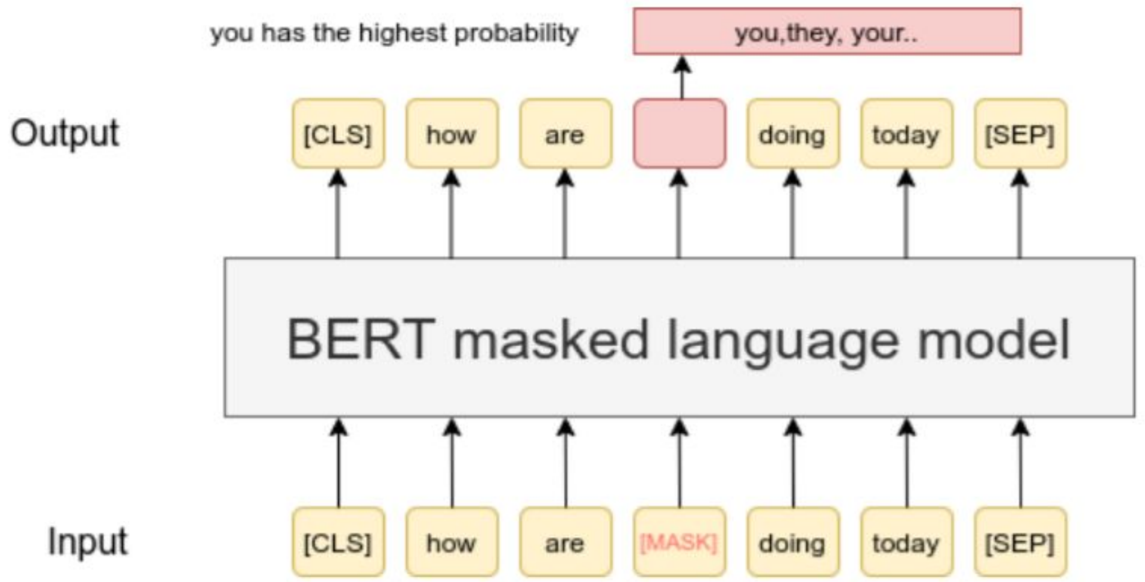
Self-supervised learning

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.

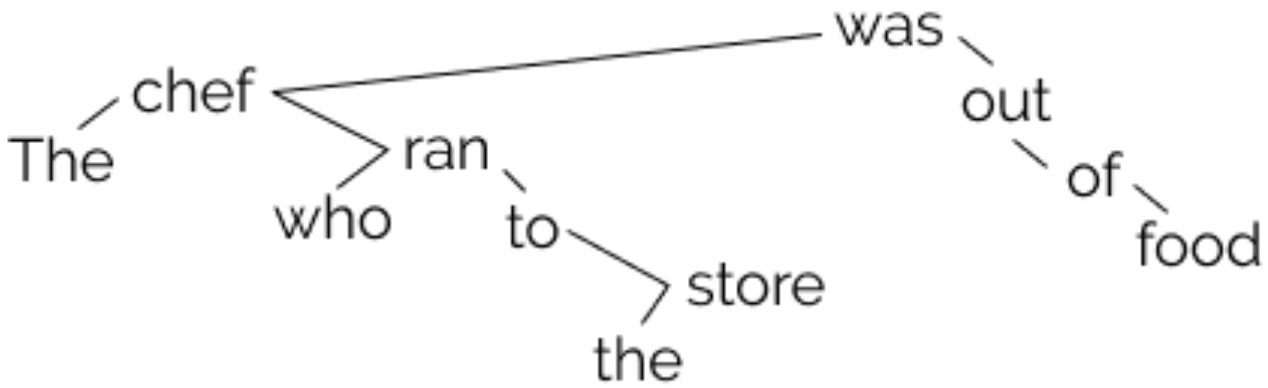


Slide: LeCun

Self-supervised learning



Syntactic dependency trees



https://github.com/john-hewitt/structural-probes/blob/master/doc-assets/tree.png

Emergent linguistic structure in artificial neural networks trained by self-supervision

Christopher D. Manning , Kevin Clark , John Hewitt , Urvashi Khandelwal, and Omer Levy

https://www.pnas.org/doi/10.1073/pnas.1907367117

BERT

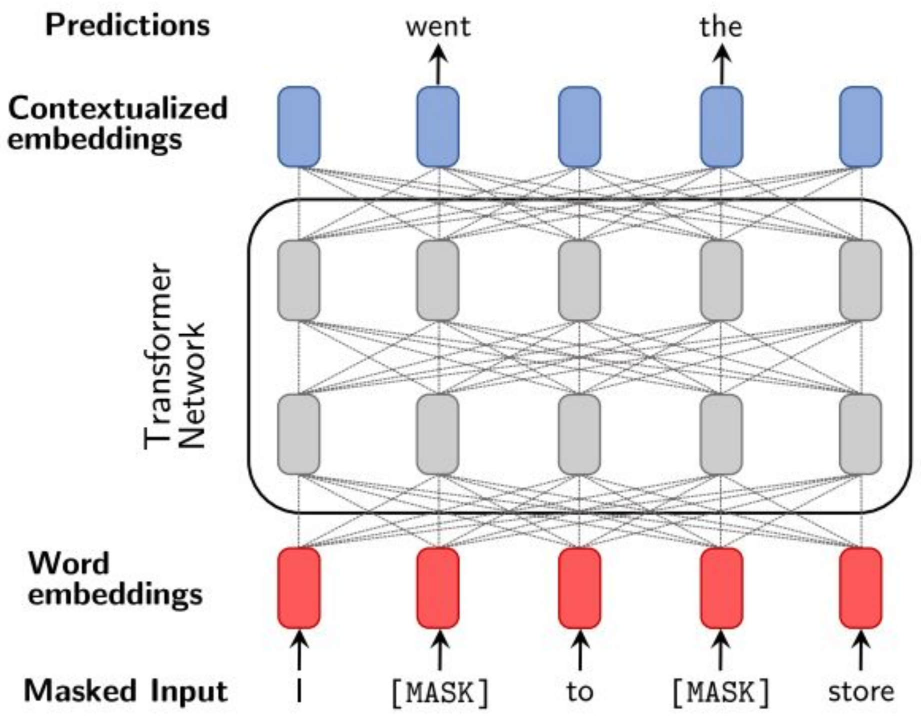


Fig. 3. A high-level illustration of BERT. Words in the input sequence are randomly masked out and then all words are embedded as vectors in $\mathbb{R}^d$. A Transformer network applies multiple layers of multiheaded attention to the representations. The final representations are used to predict the identities of the masked-out input words.

BERT attention heads act as simple classifiers

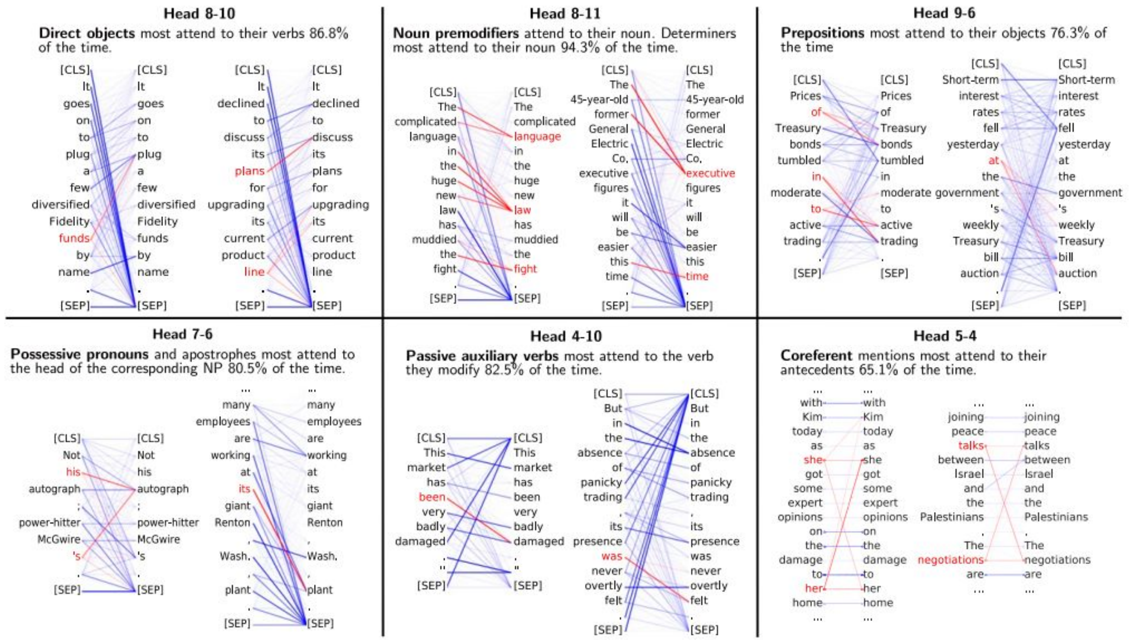
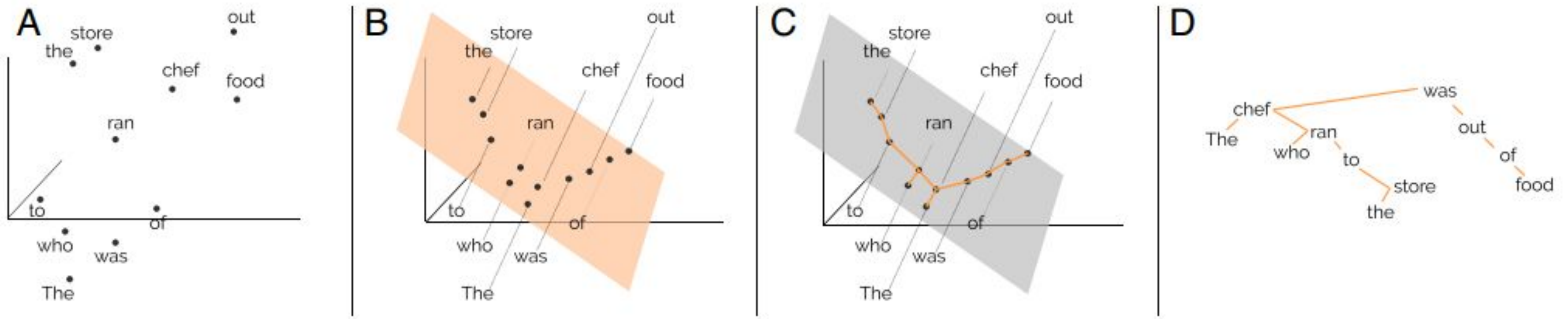

Head 8-10
Direct objects most attend to their verbs 86.8% of the time.
[CLS] It goes on to plug a few diversified Fidelity funds by name . [SEP]
[CLS] It declined to discuss its plans for upgrading its current product line . [SEP]
Head 8-11
Noun premodifiers attend to their noun. Determiners most attend to their noun 94.3% of the time.
[CLS] The complicated language in the huge new law has muddied the fight . [SEP]
[CLS] The 45-year-old former General Electric Co. executive figures it will be easier this time . [SEP]
Head 9-6
Prepositions most attend to their objects 76.3% of the time
[CLS] Prices of Treasury bonds tumbled in moderate to active trading . [SEP]
[CLS] Short-term interest rates fell yesterday at the government 's weekly Treasury bill auction . [SEP]
Head 7-6
Possessive pronouns and apostrophes most attend to the head of the corresponding NP 80.5% of the time.
[CLS] Not his autograph ; power-hitter McGwire 's . [SEP]
... many employees are working at its giant Renton , Wash. , plant . [SEP]
Head 4-10
Passive auxiliary verbs most attend to the verb they modify 82.5% of the time.
[CLS] This market has been very badly damaged . " [SEP]
[CLS] But in the absence of panicky trading , its presence was never overtly felt . [SEP]
Head 5-4
Coreferent mentions most attend to their antecedents 65.1% of the time.
... with Kim today as she got some expert opinions on the damage to her home ...
... joining peace talks between Israel and the Palestinians . The negotiations are ...

Fig. 6. Some BERT attention heads that appear sensitive to linguistic phenomena, despite not being explicitly trained on linguistic annotations. In the example attention maps, the darkness of a line indicates the size of the attention weight. All attention to/from red words is colored red; these words are chosen to highlight certain of the attention heads' behaviors. [CLS] (classification) and [SEP] (separator) are special tokens BERT adds to the input during preprocessing. Attention heads are numbered by their layer and index in BERT. Reprinted with permission from ref. 59, which is licensed under CC BY 4.0.

Syntax trees hidden within word-representations



Each of the words of the sentence *The chef who ran to the store was out of food* is internally represented in context as a vector.

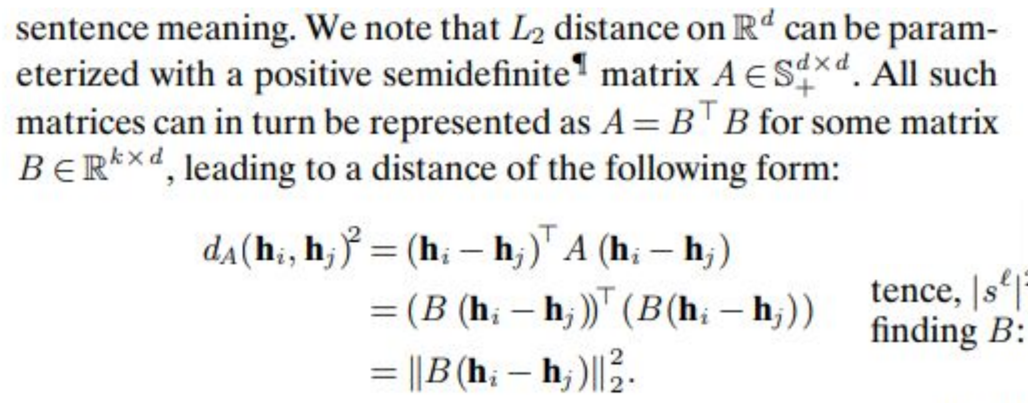
A structural probe finds a linear transform of that space under which squared L_2 distance between vectors best reconstructs tree path distance between words.

Once in this latent space, the structure of the tree is globally represented by the geometry of the vector space, meaning words that are close in the space are close in the tree.

In fact, the tree can be approximately recovered by taking a minimum spanning tree in the latent syntax space.

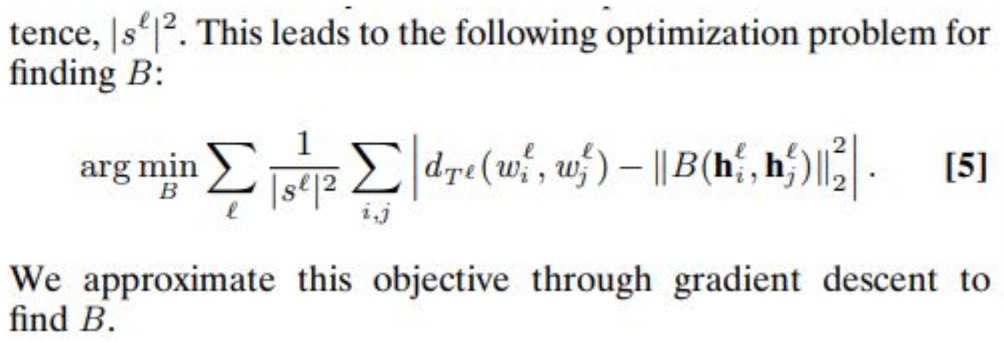
Fig. 7. (*A–D*) An overview of the structural probe method.

Learning this linear transform is a convex problem



sentence meaning. We note that L_2 distance on $\mathbb{R}^d$ can be parameterized with a positive semidefinite¶ matrix $A \in \mathbb{S}_{+}^{d\times d}$. All such matrices can in turn be represented as $A = B^\top B$ for some matrix $B \in \mathbb{R}^{k\times d}$, leading to a distance of the following form:

$$\begin{aligned} d_A(\mathbf{h}_i, \mathbf{h}_j)^2 &= (\mathbf{h}_i - \mathbf{h}_j)^\top A\, (\mathbf{h}_i - \mathbf{h}_j) \\ &= (B\,(\mathbf{h}_i - \mathbf{h}_j))^\top (B(\mathbf{h}_i - \mathbf{h}_j)) \\ &= \|B(\mathbf{h}_i - \mathbf{h}_j)\|_2^2. \end{aligned}$$



tence, $|s^\ell|^2$. This leads to the following optimization problem for finding B:

$$\arg\min_B \sum_\ell \frac{1}{|s^\ell|^2} \sum_{i,j} \left| d_{T^\ell}(w_i^\ell, w_j^\ell) - \|B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell)\|_2^2 \right|. \qquad [5]$$

We approximate this objective through gradient descent to find B.

Wait a minute... Isn’t this just transfer learning?

After all, the linear transformation is just a layer we’re adding to predict syntax trees right?

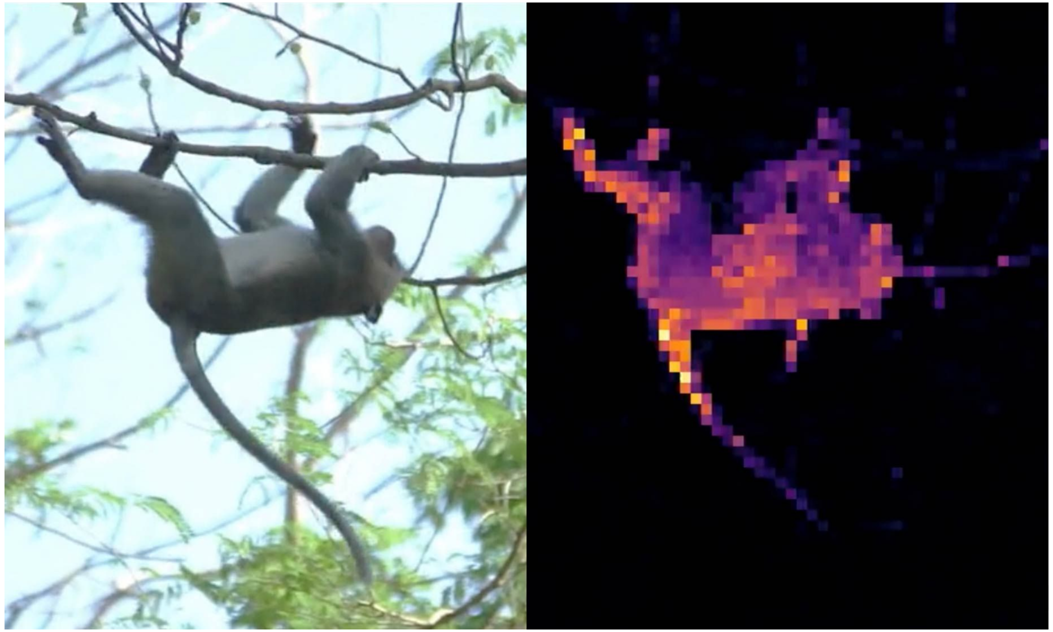
Well yes but actually no. The point was not to create a better parser, but to study the internal representations of BERT embeddings :)

Does this happen elsewhere?

Yes!

Transformer models trained with self supervision-like methodologies exhibit a similarity in high-level structures emerging out of just solving jigsaw puzzles.

DINO and PAWS



Implications

- One of the earliest motivations behind constructing syntax parse trees was to help solve more complex problems in NLP

- But in light of these discoveries, should we give up on training machines to solve auxiliary problems?

Thank you!