Machine Learning: Self-Supervised Learning 2

Hao Tang

March 23, 2026

The task of language modeling is to

The task of language modeling is to predict

The task of language modeling is to predict the

The task of language modeling is to predict the next

The task of language modeling is to predict the next word

The task of language modeling is to predict the next word .

Language modeling is future prediction

- Language modeling is the task of predicting the next word given the past.
- Given a sentence of k words $t_1, \ldots, t_k$, the objective of language modeling is to maximize

$$p(t_1, \ldots, t_k) = \prod_{i=1}^{k} p(t_i | t_1, \ldots, t_k). \tag{1}$$

- Language modeling is a form of future prediction.

Deep contextualized word representations

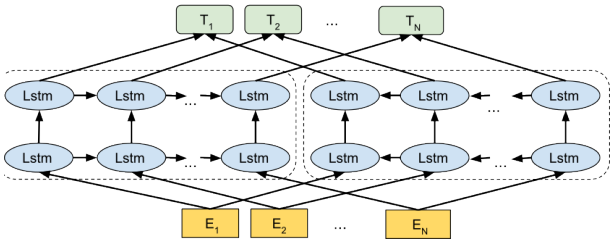
Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

- ELMo (embeddings from language models) use the same future prediction objective.

- In fact, ELMo actually includes past prediction as well.



(Devlin *et al.*, 2019)

Improving Language Understanding by Generative Pre-Training

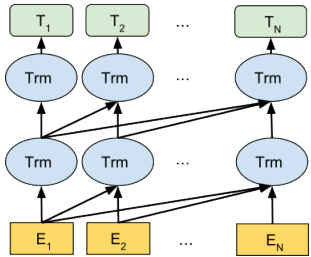
Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

- GPT uses the same future prediction objective but with Transformers.



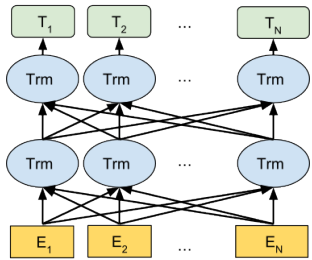
(Devlin *et al.*, 2019)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

- The loss function for training BERT is masked prediction.

Masked prediction is another form of language modeling

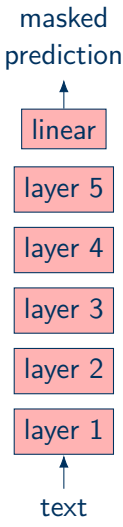


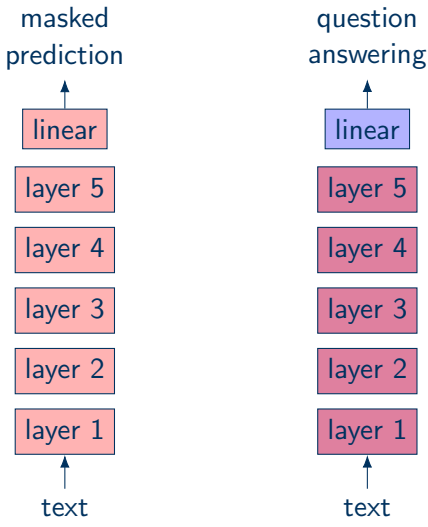
(Devlin *et al.*, 2019)

Masked [MASK] is another [MASK] of [MASK] modeling

- A pre-trained model can serve as an initialization of another model on a new task, i.e., for fine-tuning.

- However, if we want to understand what a model learns, we'd better not change the parameters.

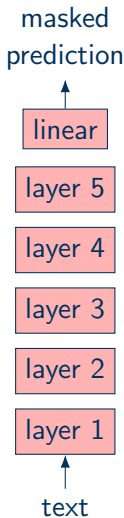

masked
prediction
linear
layer 5
layer 4
layer 3
layer 2
layer 1
text

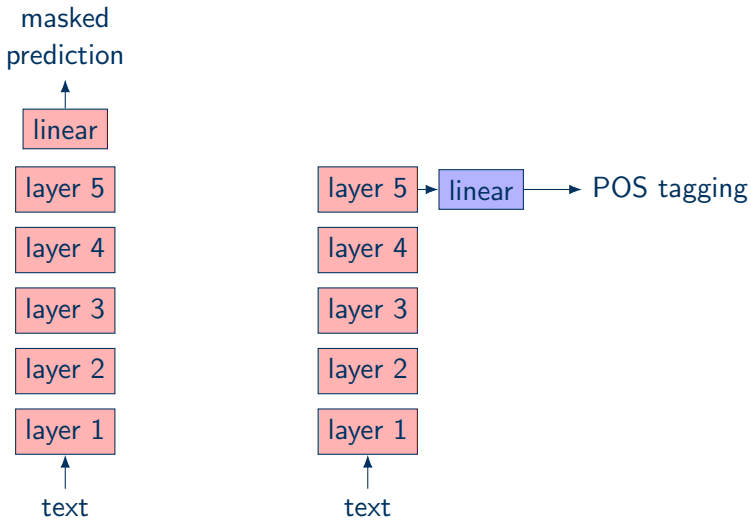

masked
prediction
linear
layer 5
layer 4
layer 3
layer 2
layer 1
text
question
answering
linear
layer 5
layer 4
layer 3
layer 2
layer 1
text

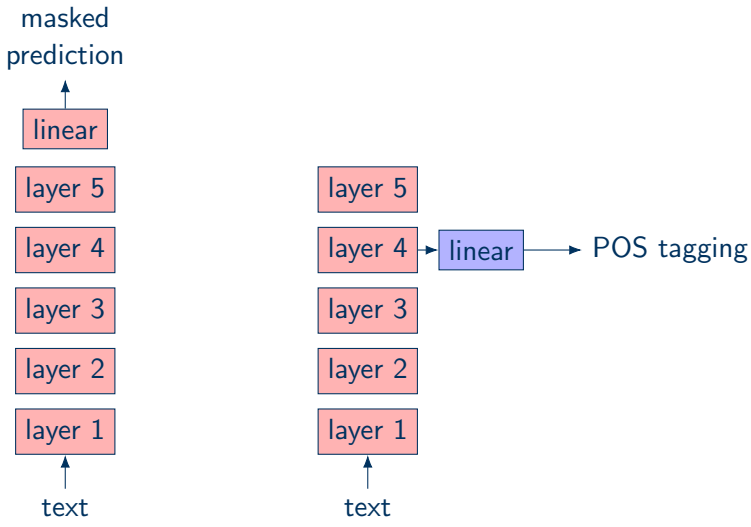
Probing as an approach to analyzing representations

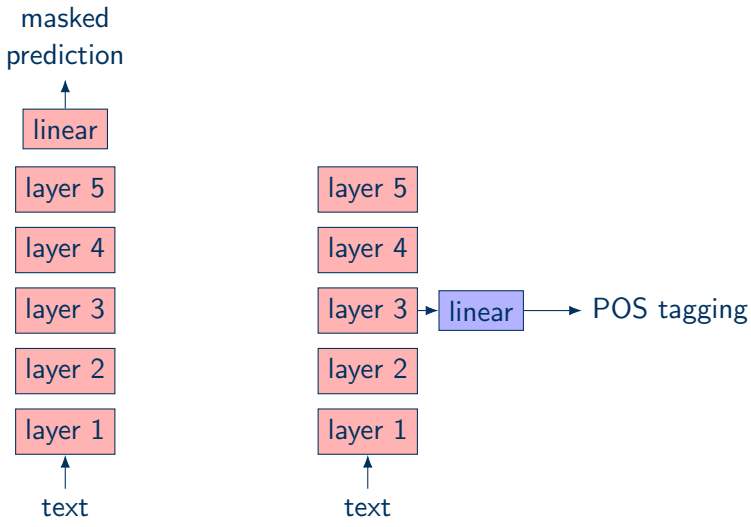
- The hidden vectors are often better representations than the input at solving tasks.
- Don't forget that input themselves are also representations.
- What aspects of the input are the hidden vectors "representing"?

- For example, can we decode part-of-speech (POS) from a given text representation?
- To answer this question, we train a linear classifier to predict POS tags.
- A non-trivial accuracy tells us to what extent we can decode POS tags from a given representation.
- The act is called **probing**, and the classifier is called a probing classifier.


masked
prediction
linear
layer 5
layer 4
layer 3
layer 2
layer 1
text


masked
prediction
linear
layer 5
layer 4
layer 3
layer 2
layer 1
text
layer 5
linear
POS tagging
layer 4
layer 3
layer 2
layer 1
text

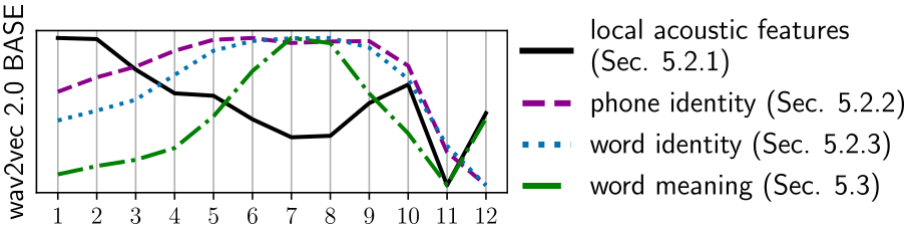

masked
prediction
linear
layer 5
layer 4
layer 3
layer 2
layer 1
text
layer 5
layer 4
linear
POS tagging
layer 3
layer 2
layer 1
text


masked
prediction
linear
layer 5
layer 4
layer 3
layer 2
layer 1
text
layer 5
layer 4
layer 3
linear
POS tagging
layer 2
layer 1
text

- Before we analyze what are represented, we need domain knowledge, i.e., properties of the input.
- Properties of a text sentence include syntax (how a sentence is structured), semantics (what a sentence means), and many other things.
- Properties of a speech utterance include linguistics (what's said), paralinguistics (how it's said), and extralinguistics (everything else).
- We cannot evaluate something without a hypothesis to begin with.

Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	Tense (Semantic)	SubjNum (Semantic)	ObjNum (Semantic)	SOMO (Semantic)	CoordInv (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	95.9 (3.4)	65.0 (64.8)	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	**96.2 (3.9)**	66.5 (66.0)	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	94.2 (2.3)	**69.8 (69.6)**	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	92.0 (0.5)	69.2 (69.0)	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	88.4 (-3.0)	63.5 (63.4)	**41.3 (13.0)**	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	**88.1 (21.9)**	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	83.7 (-7.7)	56.9 (56.7)	40.1 (12.0)	**84.1 (39.5)**	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	**82.2 (21.1)**	61.6 (11.7)	74.8 (24.9)
8	82.9 (-8.1)	51.1 (51.0)	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	80.1 (-11.1)	47.9 (47.8)	38.5 (10.8)	83.1 (39.8)	**87.0 (37.1)**	**90.0 (28.0)**	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	**78.7 (28.9)**
10	77.0 (-14.0)	43.4 (43.2)	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	73.9 (-17.0)	42.8 (42.7)	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	69.5 (-21.4)	49.1 (49.0)	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	**65.2 (15.3)**	74.9 (25.4)

(Jawahar *et al.*, 2019)



(Pasad *et al.*, 2021)

- Should we use a linear probing classifier or a deeper one?

- Should we use a linear probing classifier or a deeper one?
- A high accuracy with a linear probing classifier shows that the information is readily accessible.

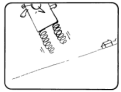
- Should we use a linear probing classifier or a deeper one?
- A high accuracy with a linear probing classifier shows that the information is readily accessible.
- A high accuracy with a deep probing classifier shows that the information is present.

- Should we use a linear probing classifier or a deeper one?
- A high accuracy with a linear probing classifier shows that the information is readily accessible.
- A high accuracy with a deep probing classifier shows that the information is present.
- A low probing accuracy is not an evidence that the information is absent.

- Should we use a linear probing classifier or a deeper one?
- A high accuracy with a linear probing classifier shows that the information is readily accessible.
- A high accuracy with a deep probing classifier shows that the information is present.
- A low probing accuracy is not an evidence that the information is absent.
- We don't know if there is exists a model with higher capacity that can achieve a high accuracy.

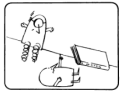
- Should we use a linear probing classifier or a deeper one?
- A high accuracy with a linear probing classifier shows that the information is readily accessible.
- A high accuracy with a deep probing classifier shows that the information is present.
- A low probing accuracy is not an evidence that the information is absent.
- We don't know if there is exists a model with higher capacity that can achieve a high accuracy.
- Are there universal representations?

The Amazing Code

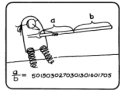


Dr. Zeta is a scientist from Helix, a galaxy in another space-time dimension. One day Dr. Zeta visited the earth to gather information about humans. His host was an American scientist named Herman.

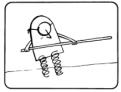
Using his powerful pocket computer, Dr. Zeta scanned the encyclopedia quickly, translating its entire content into one gigantic number. By putting a decimal point in front of the number, he made it a decimal fraction.



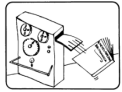
Herman: Why don't you take back a set of the Encyclopedia Britannica? It's a great summary of all our knowledge.
Dr. Zeta: Splendid idea, Herman. Unfortunately, I can't carry anything with that much mass.



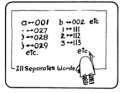
Dr. Zeta then placed a mark on his rod, dividing it accurately into lengths *a* and *b* so that the fraction *a*/*b* was equivalent to the decimal fraction of his code.



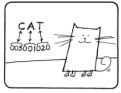
Dr. Zeta: However, I can encode the entire encyclopedia on this metal rod. One mark on the rod will do the trick.
Herman: Are you joking? How can one little mark carry so much information?



Dr. Zeta: When I get back to my planet, one of our computers will measure *a* and *b* exactly, then compute the fraction *a*/*b*. This decimal fraction will be decoded, and the computer will print your encyclopedia for us!



Dr. Zeta: Elementary, my dear Herman. There are less than a thousand different letters and symbols in your encyclopedia. I will assign a number from 1 through 999 to each letter or symbol, adding zeros on the left if needed so that each number used will have three digits.



Herman: I don't understand. How would you code the word *cat*?
Dr. Zeta: It's simple. We use the sort of code I just showed you. *Cat* might be coded 003001020.

Real numbers, data science and chaos: How to fit any dataset with a single parameter

Laurent Boué

SAP Labs

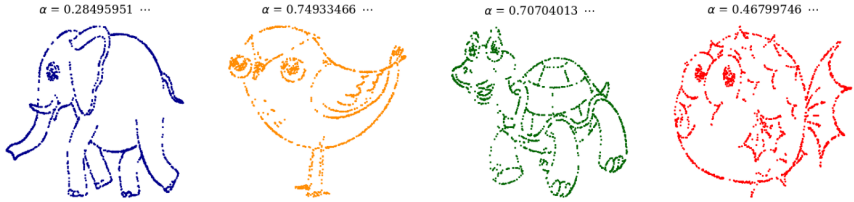
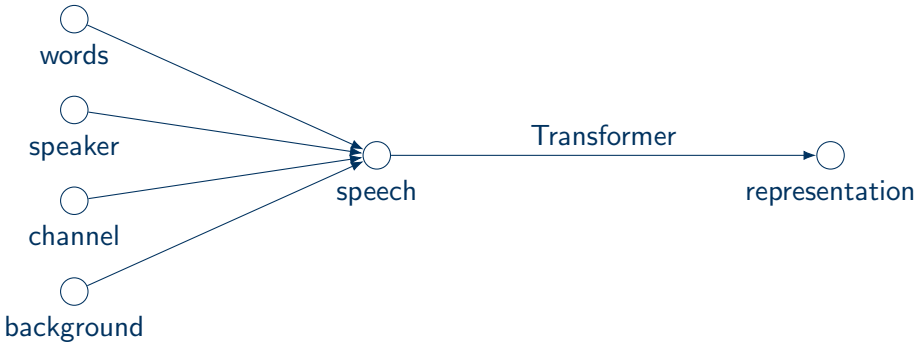
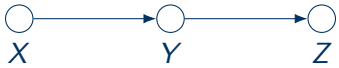


Figure 1: Animal shapes obtained with the different values of α defined on top of each image. One should consider the data as a scatter plot of pairs of values (x, y) where each $x \in \mathbb{N}$ is associated with a corresponding y value given by $y \equiv f_\alpha(x)$. One goal of the paper will be to show how to find the precise value of $\alpha \in \mathbb{R}$ required to fit any target dataset.

- Universal representations exist, but they might not be useful if the information we care about is not readily accessible.

- Can hidden representations be richer than the input?


words
speaker
channel
background
speech
Transformer
representation



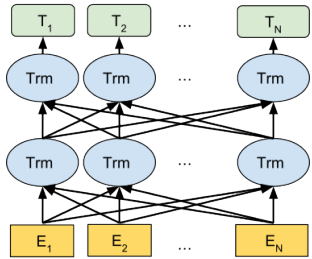
$$I(X, Y) \geq I(X, Z)$$

- The data processing inequality says that the more we process, the more we lose.
- The hidden representation cannot be richer than the input.

Masked prediction as denoising autoencoders

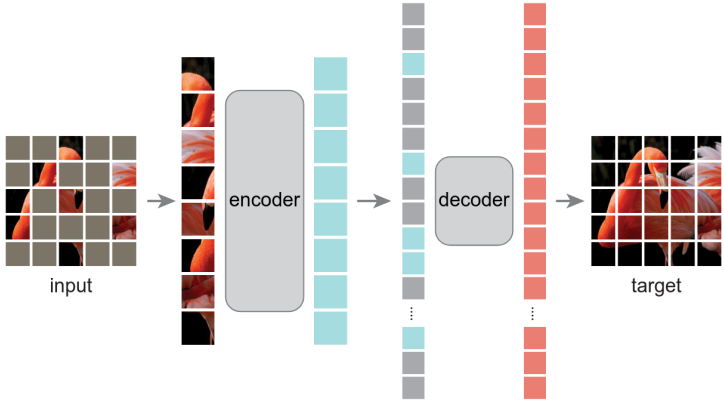
- The loss function for training BERT is masked prediction.

Masked prediction is another form of language modeling



(Devlin *et al.*, 2019)

Masked [MASK] is another [MASK] of [MASK] modeling



(He *et al.*, 2022)

- In general, an autoencoder consists of an encoder f and a decoder g.
- The goal of an autoencoder is to minimize

$$\|x - g(f(x))\|^2 \tag{2}$$

for a data point x.

- In general, an autoencoder consists of an encoder f and a decoder g.
- The goal of an autoencoder is to minimize

$$\|x - g(f(x))\|^2 \tag{2}$$

  for a data point x.
- This is likely to degenerate when f and g collaborate.

- What happends when $f(x) = W_1 x$ and $g(z) = W_2 z$?
- The reconstruction loss becomes

$$\|x - W_2 W_1 x\|^2 \tag{3}$$

  for a data point x.
- When $d_2 < d_1$ where $W_1 \in \mathbb{R}^{d_2 \times d_1}$, then the encoder is forced to compress.
- The training might not degenerate, depending on whether the compression is lossy or not.

- A denoising autoencoder instead minimizes

$$\mathbb{E}_{\epsilon}\|x - g(f(x + \epsilon))\|^2 \tag{4}$$

  for a data point x.

- Masked prediction is a form of denoising autoencoding.

Reference

- Peters et al., Deep contextualized word representations, 2018
- Radford et al., Improving language understanding by generative pre-training, 2019
- Devlin et al., BERT: Pre-training of deep bidirectional Transformers for language understanding, 2019
- Jawar et al., What does BERT learn about the structure of language?, 2019
- Boué, Real numbers, data science and chaos: How to fit any dataset with a single parameter, 2019
- Pasad et al., Layer-wise analysis of a self-supervised speech representation model, 2023
- He et al., Masked autoencoders are scalable vision learners, 2022