The Convexity of BERT: From Cause to Solution

Final Report - Deep Learning Project

S. M. Ali Modarressi   Hosein Mohebbi

August 16, 2020

Introduction: SentEval Probing Tasks

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)

Table 2: Probing task performance for each BERT layer. The value within the parentheses corresponds to the difference in performance of trained vs. untrained BERT.

Table borrowed from (Jawahar et al., ACL 2019)

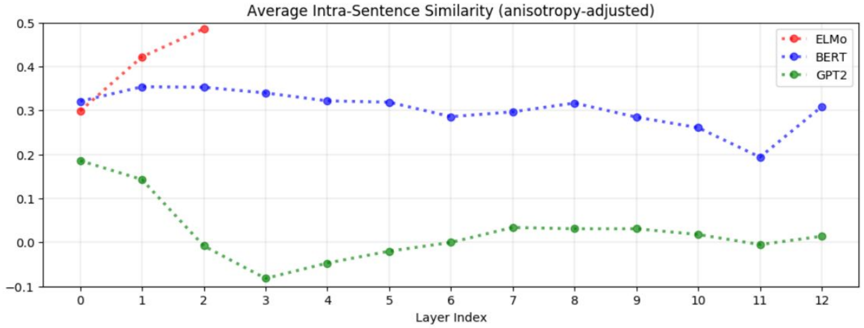
Introduction: SentEval Probing Tasks

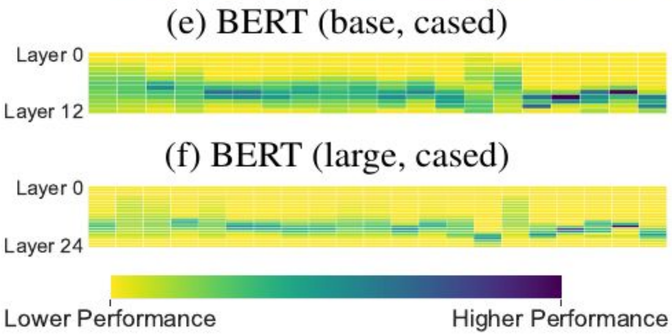
Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	**Tense** (Semantic)	**SubjNum** (Semantic)	**ObjNum** (Semantic)	**SOMO** (Semantic)	**CoordInv** (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	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)

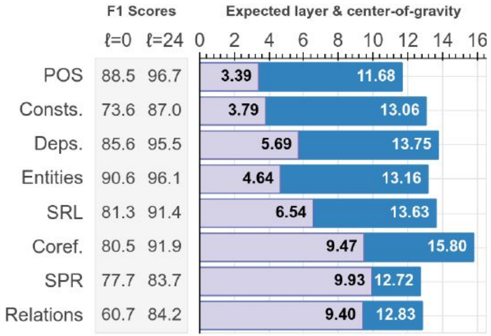
Table 2: Probing task performance for each BERT layer. The value within the parentheses corresponds to the difference in performance of trained vs. untrained BERT.

Table borrowed from [(Jawahar et al., ACL 2019)]

Related Work

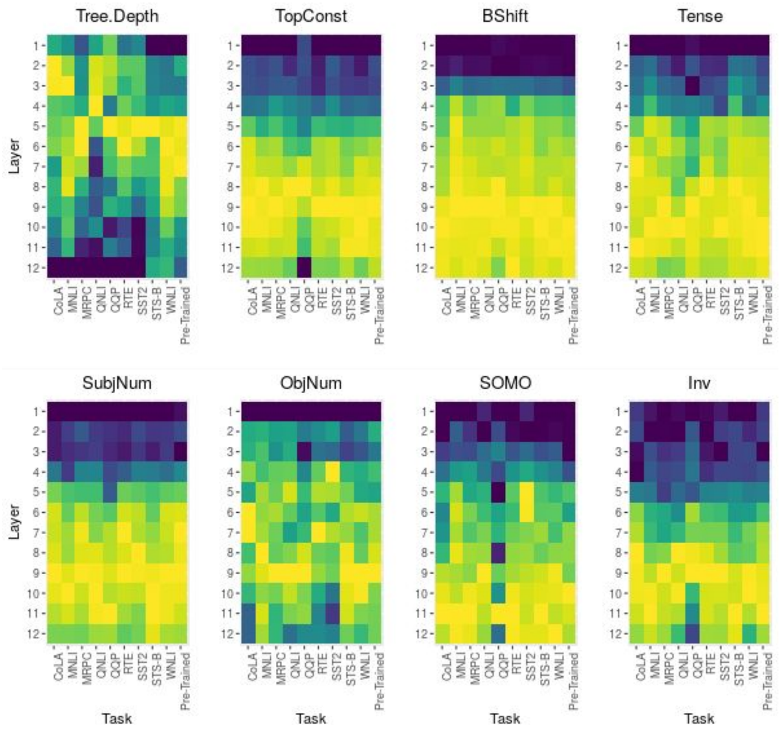






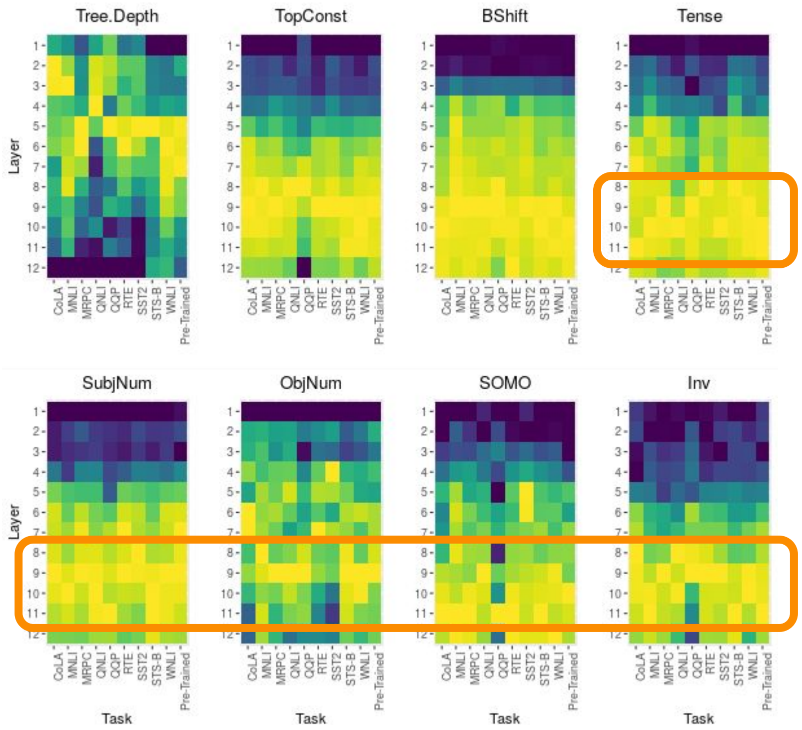
Figures borrowed from (Ethayarajh et al., ACL 2019), (Liu et al., ACL 2019), (Tenney et al., ACL 2019)

On Fine-tuned BERT Probing Results



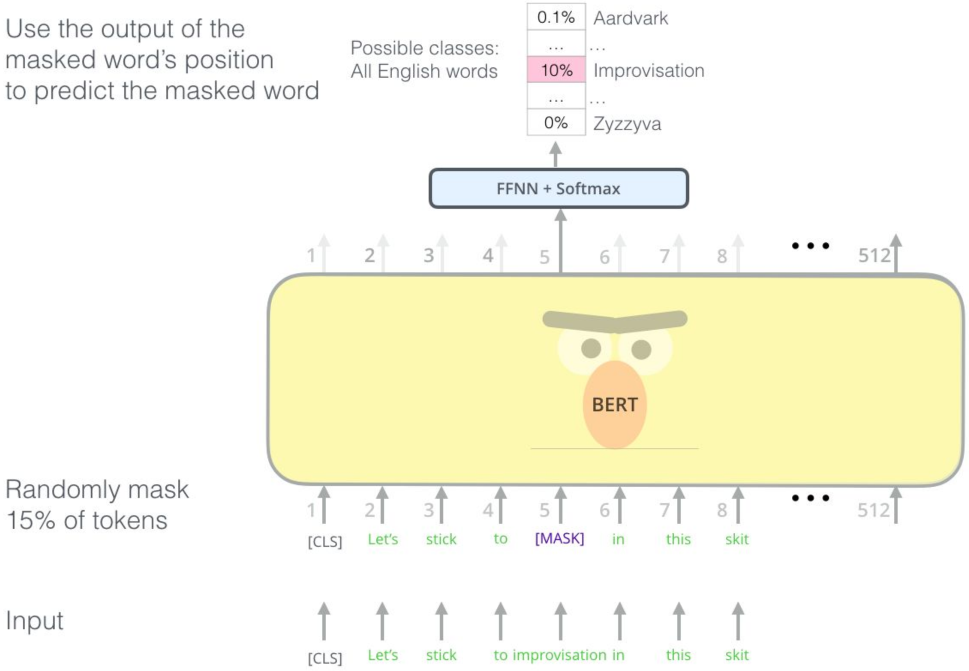
- Fine-tuned BERT using GLUE tasks
- Probing [CLS] Representations in each layer using diagnostic classifiers

On Fine-tuned BERT Probing Results



- Fine-tuned BERT using GLUE tasks
- Probing [CLS] Representations in each layer using diagnostic classifiers

BERT MLM Pre-Training



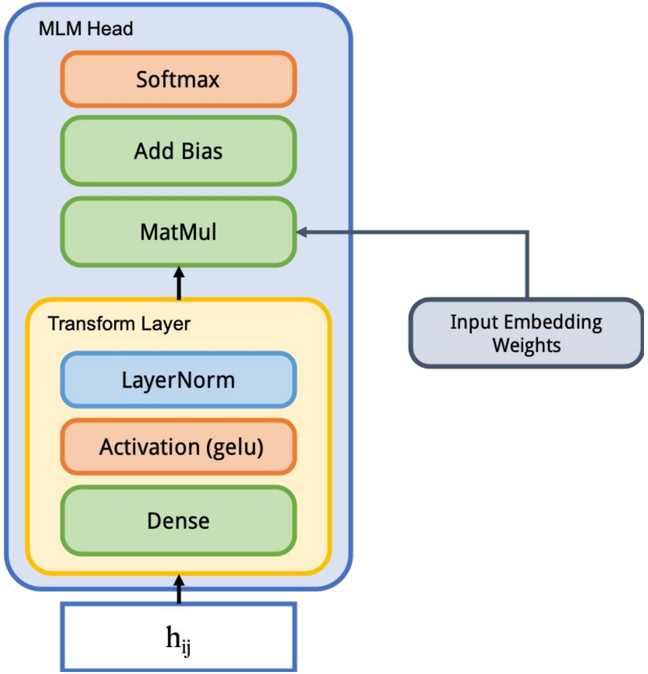
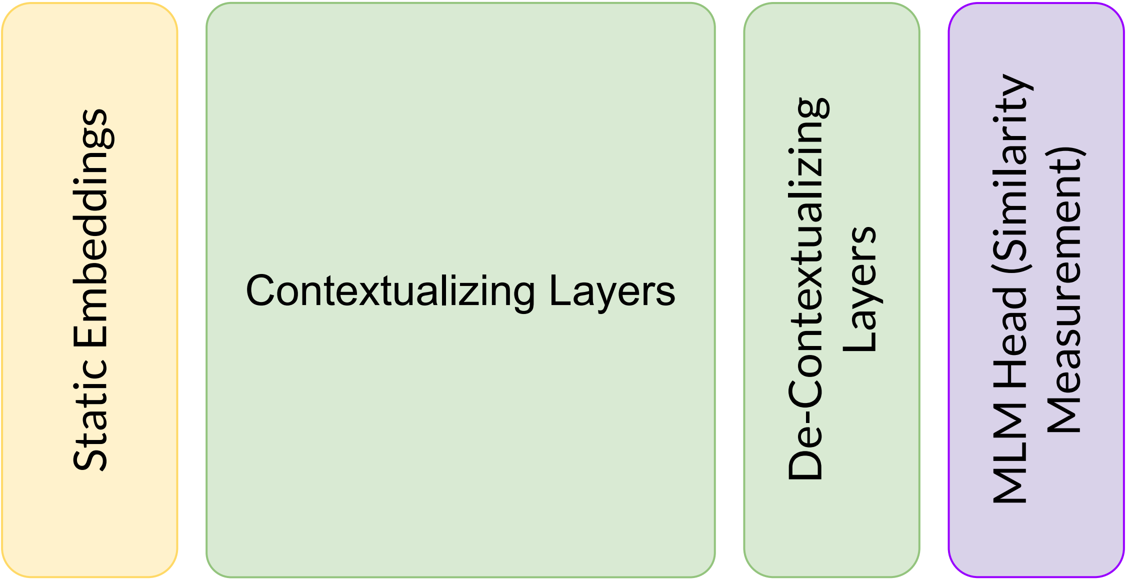
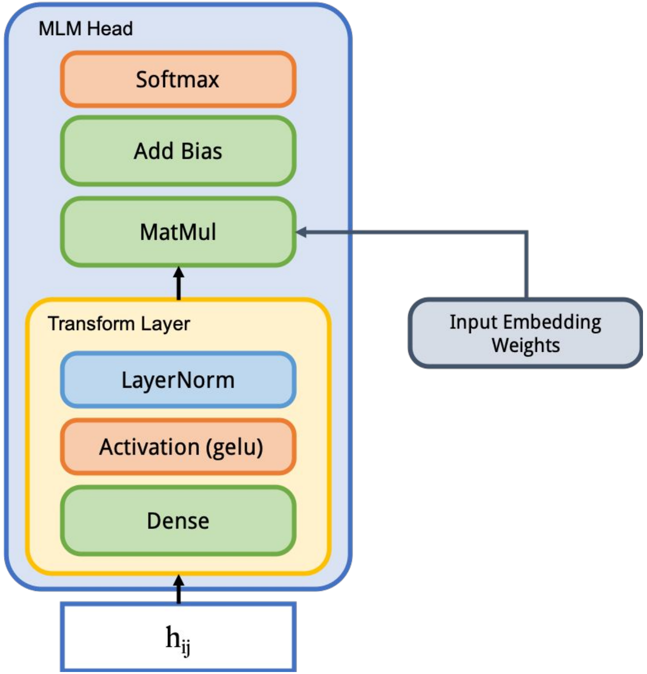


Figure left borrowed from (https://jalammar.github.io/illustrated-bert/)

BERT MLM Pre-Training

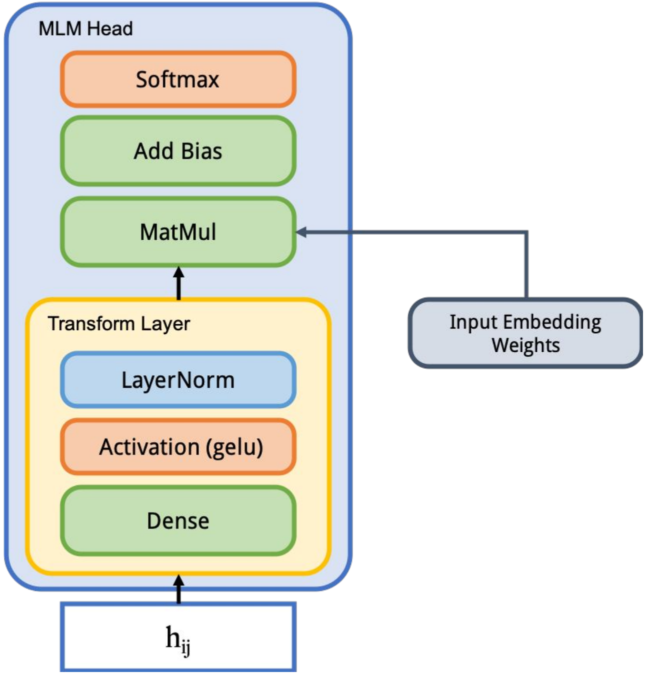


BERT MLM Probing Method



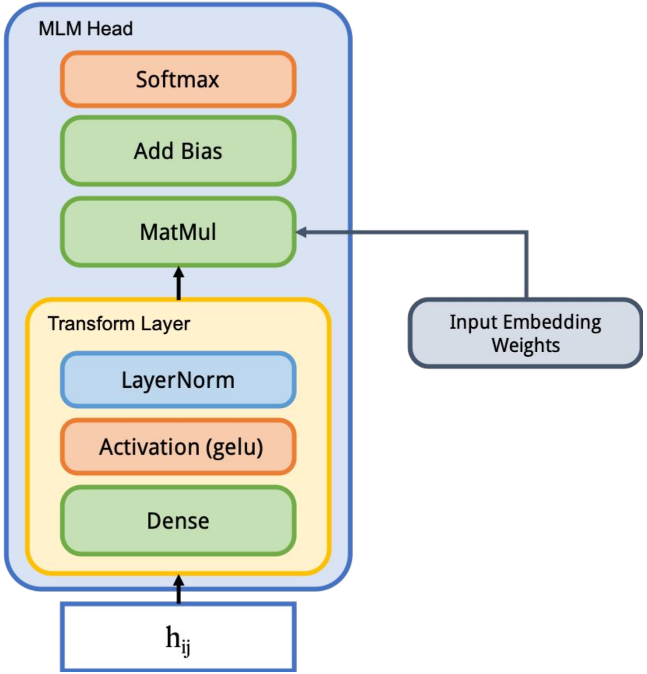
- Using MLM Head
  - Pre-Trained?

BERT MLM Probing Method



- Using MLM Head
  - ~~Pre-Trained?~~
  - Train for each layer's representation
    - Using [MASK]?

BERT MLM Probing Method



- Using MLM Head
    - ~~Pre-Trained?~~
    - Train for each layer's representation
        - ~~Using [MASK]?~~
        - Without Masking?

BERT MLM Probing Method

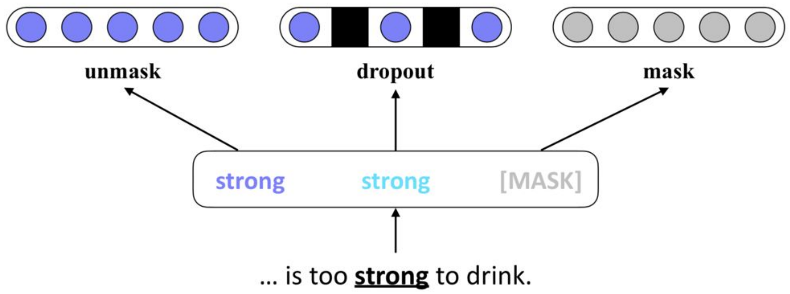
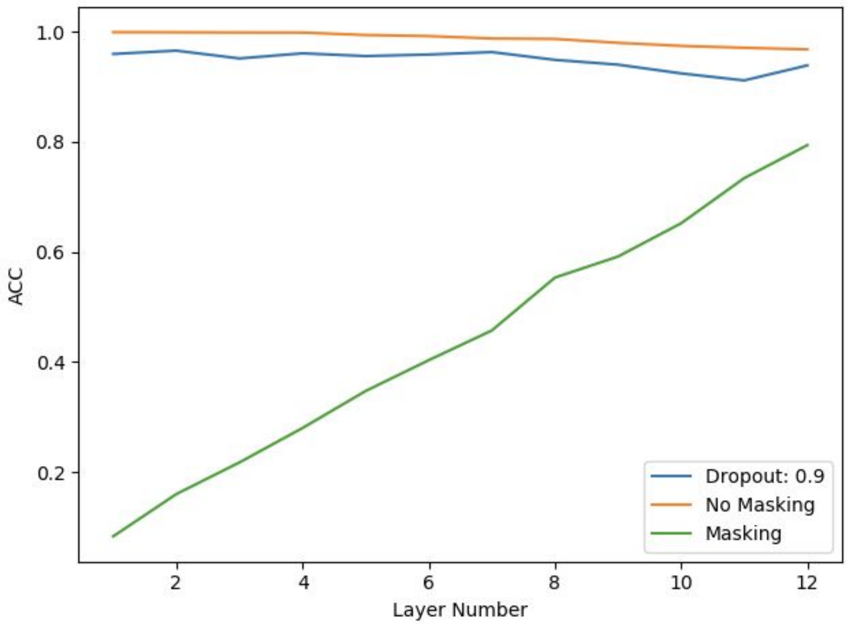


Figure borrowed from (Zhou et al., ACL 2019)

- Using MLM Head
  - ~~Pre-Trained?~~
  - Train for each layer's representation
    - ~~Using [MASK]?~~
    - ~~Without Masking?~~
    - Highly Dropouted Embeddings

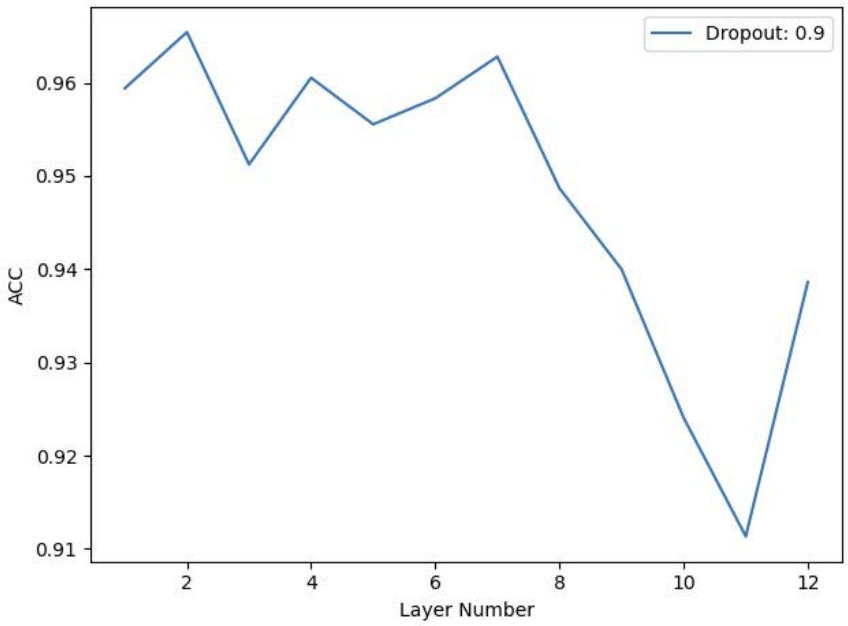
BERT MLM Probing Method

Results



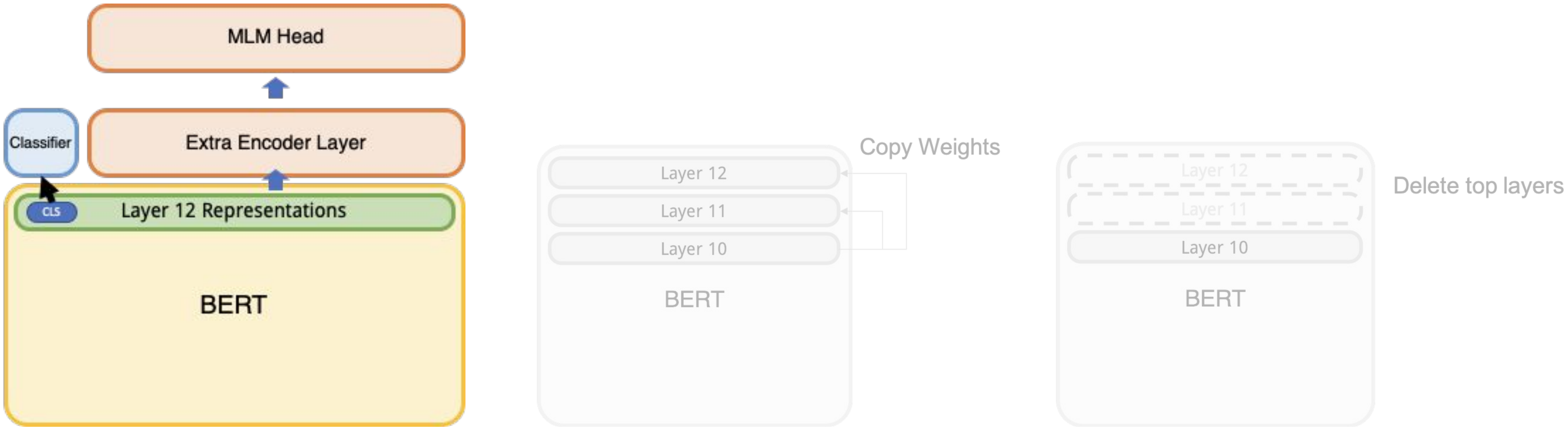
- Using MLM Head
  - ~~Pre-Trained?~~
  - Train for each layer's representation
    - ~~Using [MASK]?~~
    - ~~Without Masking?~~
    - Highly Dropouted Embeddings

BERT MLM Probing Method
Results

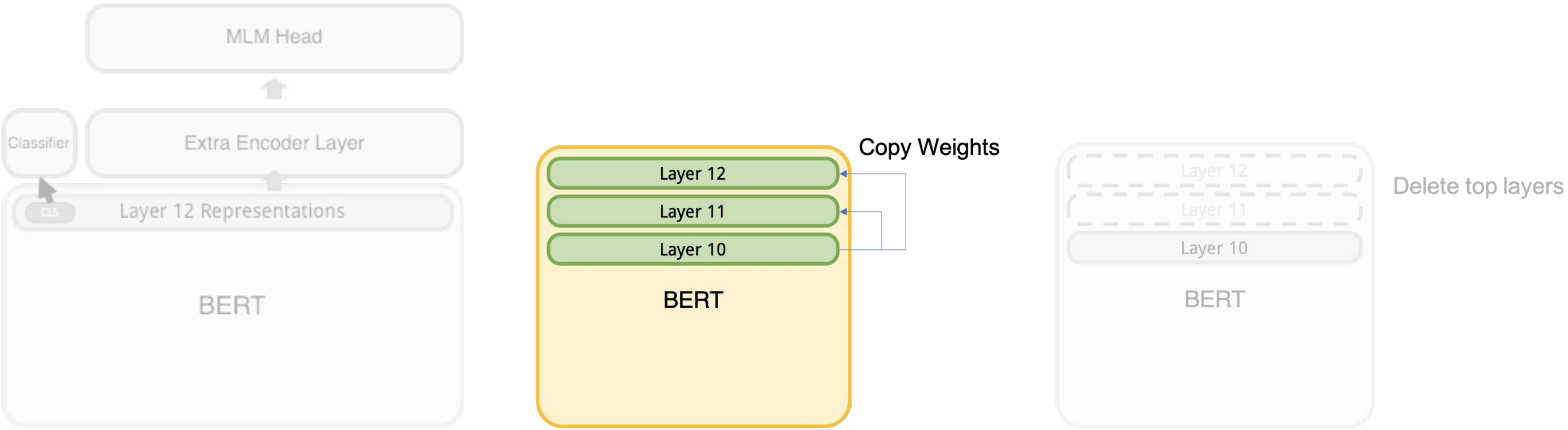


- Initial layers are more similar to the static embeddings
- Similarity decreases by advancing through the network
- Last layer regains similarity (lower contextuality)

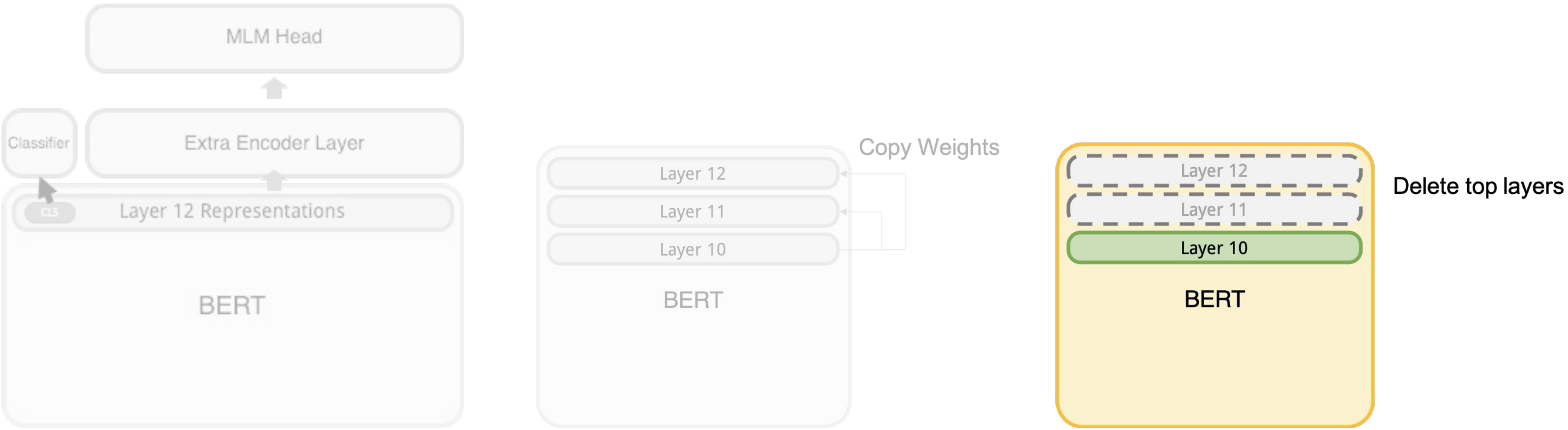
Our Proposed Methods to Utilize Maximum Contextualism



Our Proposed Methods to Utilize Maximum Contextualism



Our Proposed Methods to Utilize Maximum Contextualism



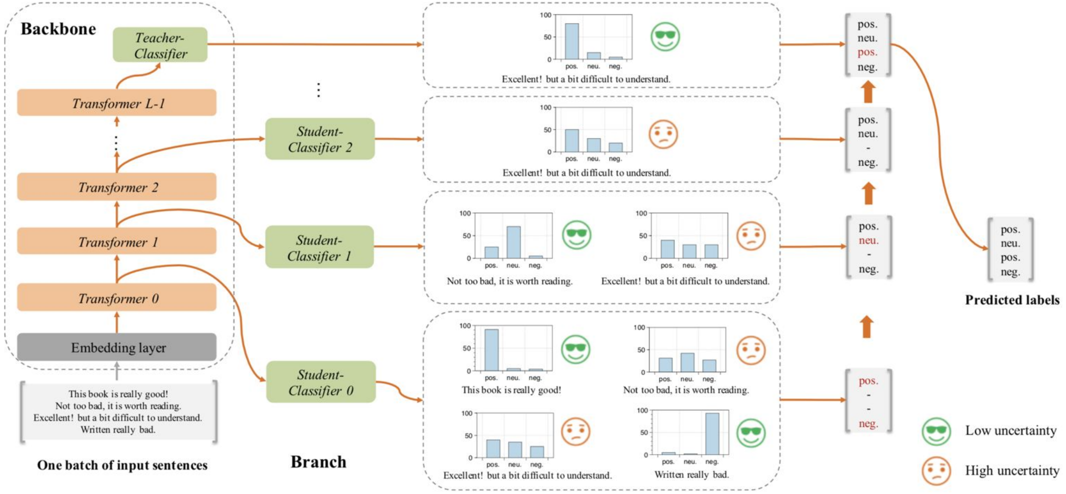
Our Proposed Methods to Utilize Maximum Contextualism

Model	CoLA	MRPC	STS-B	RTE	Average
BERT(base, uncased)	58.73	85.29	87.37	**72**	**75.84**
Multi-objective fine-tuning	59.83	83.33	**89.86**	70.31	75.83
Multi-objective + extra Layer fine-tuning	**60.6**	84.8	89.59	66.43	75.35
Coping last layers fine-tuning	58.8	**86.76**	88.33	63.18	74.26

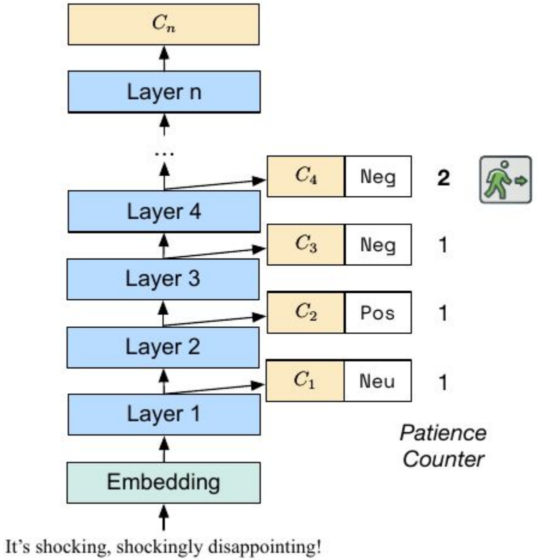
Model	CoLA	MRPC	STS-B	RTE	Average
BERT(base, cased)	**58.71**	84.64	**87.9**	**71.48**	**75.68**
Dropping last two layers fine-tuning	55.27	**85.05**	85.84	67.15	73.32

Table 1: Results on GLUE dev sets. CoLA is evaluated using Matthew's Correlation. STS-B is evaluated using Pearson's correlation coefficient. MRPC and RTE are evaluated using accuracy.

Other Recent Methods...



[FastBERT (Liu et al., ACL 2020)]



[BERT Loses Patience (Zhou et al., 2020)]

Thank You for your Attention