Not All Models Localize Linguistic Knowledge in the Same Place: A Layer-wise Probing on BERToids' Representations

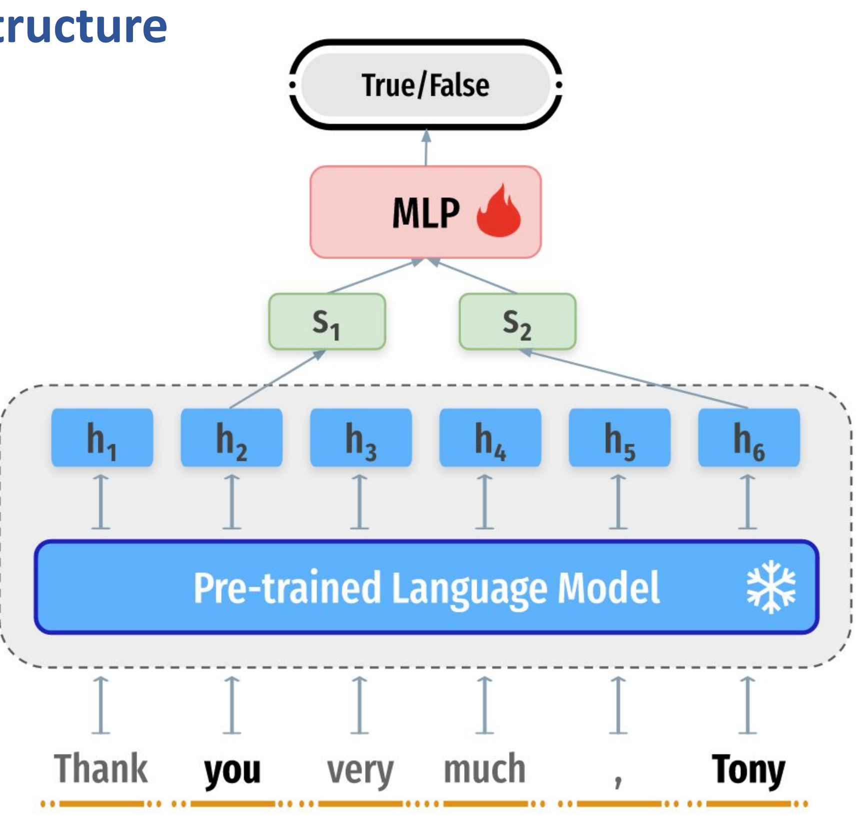
Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, Mohammad Taher Pilehvar

University of Tehran, Iran University of Science and Technology,

Tehran Institute for Advanced Studies, Khatam University

Introduction

Probing Structure

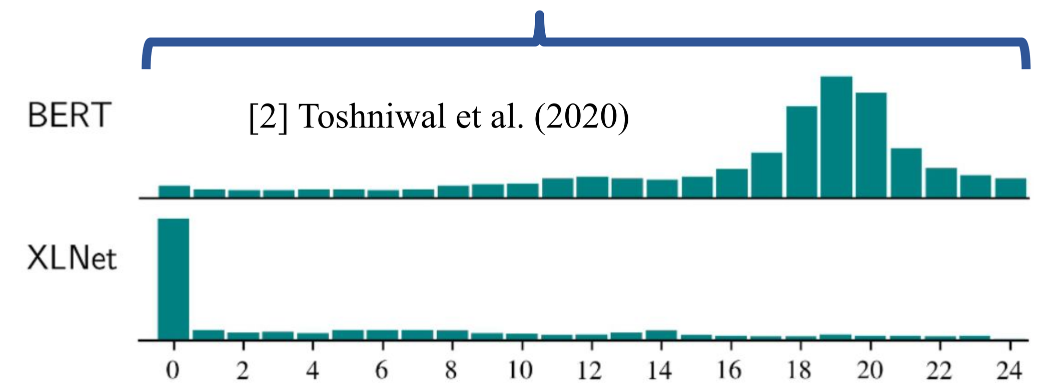


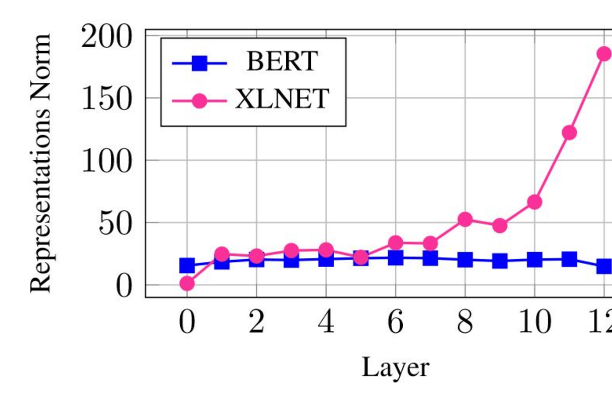
Dependencies	I think it will **[help]**$_2$ **[me]**$_1$ very much in my role . → obj (object)
NER	thirty eight years ago founded **[the special Olympics]** . → EVENT
SRL	Their father **[called]**$_1$ later **[to see if they were fine]**$_2$. → ARGM-PRP (Purpose)
Coreference	Thank **[you]**$_1$ very much , **[Tony]**$_2$. → True
Rel. (SemEval)	NASA Kepler mission sends **[names]**$_1$ into **[space]**$_2$. → Entity-Destination(e_1,e_2)

Edge Probing "Scalar Mixing Weights" Reliability Issue

$$\mathbf{h}_{i,\tau} = \gamma_\tau \sum_{\ell=0}^{L} s_\tau^{(\ell)} \mathbf{h}_i^{(\ell)} \quad [1]$$

$$\mathbf{s}_\tau = \mathrm{softmax}(\mathbf{a}_\tau)$$





→ The model tries to compensate for relatively small representation norms in XLNet's first layer

MDL Probing

- An information-theoretic probing which measures minimum description length (MDL) of labels given representations.
- MDL characterizes both probe quality and the amount of effort needed to achieve it.
- Results of MDL probes are more informative and stable than those of standard probes.
- As the number of targets *N* will affect the final codelength (MDL), we preferred to use the compression evaluation metric, which is defined as:

$$\mathbf{c} = \frac{N \cdot \log_2(K)}{\mathrm{MDL}}$$

C: Compression
N: Number of targets
K: Each label has K classes
MDL: Minimum Description Length

[3] Voita and Titov (2020)

Probing Pre-trained Representations

MDL Probing Compression (Best Among Layers) / Edge Probing F1

	BERT		XLNet		ELECTRA	
Task	F1 Score	Compression	F1 Score	Compression	F1 Score	Compression
Deps.	94.18	15.25	93.93	14.13	**94.77**	**16.15**
NER	95.61	16.87	95.51	15.46	**96.07**	**16.88**
SRL	90.91	13.94	90.56	13.32	**91.69**	**14.44**
Coref.	91.17	4.58	91.34	3.97	**92.94**	**5.88**
Rel.	80.63	3.04	82.07	2.97	**82.41**	**3.37**

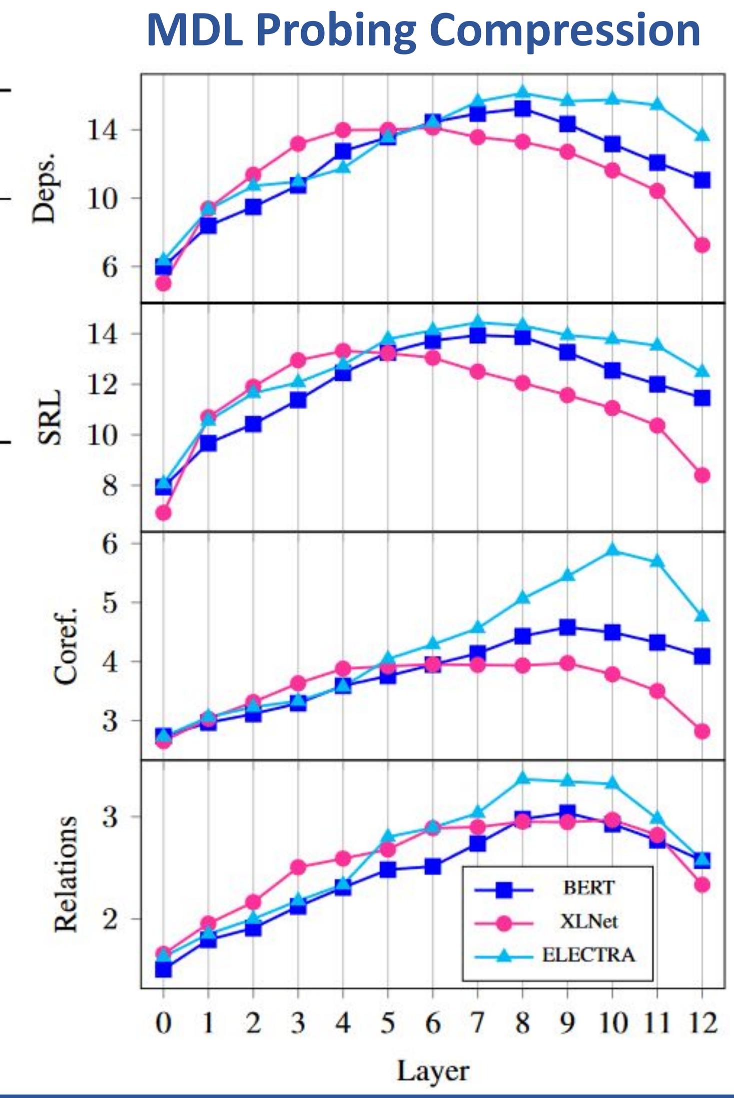
→ ELECTRA seems to have the best pre-training objective for incorporating linguistic knowledge among the three models.

→ XLNet displays comparable results to BERT, which is interesting given the relatively better fine-tuned performance of the former in a variety of downstream tasks.

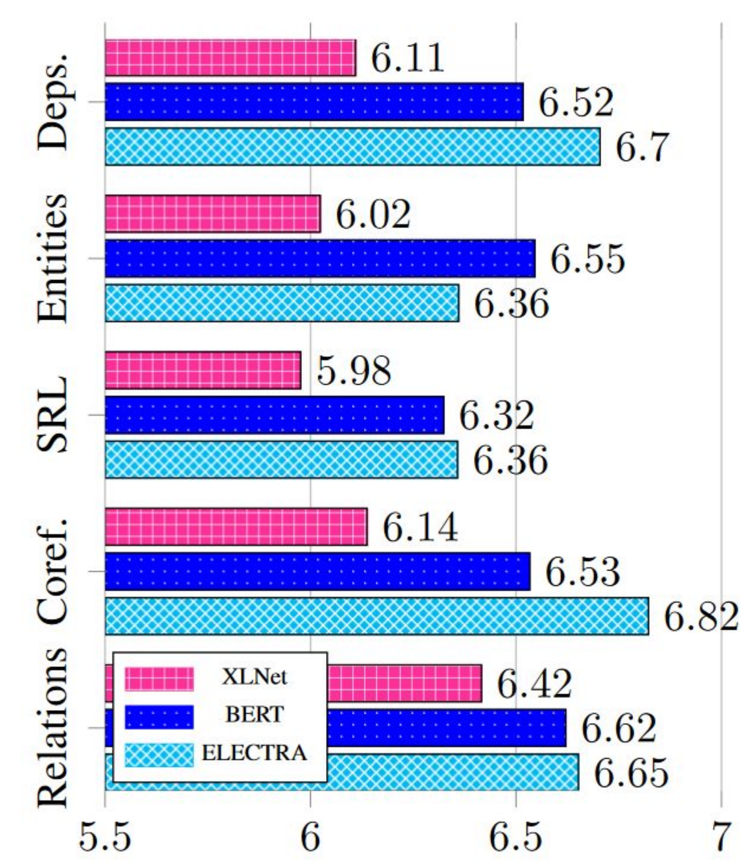
ELECTRA attains the highest compression in different layers across most tasks, especially in the deeper layers. ←

All models start with relatively low compressions and reach higher values in their middle layers and decrease towards the final layer. ←

MDL Probing Compression



MDL Probing Compression Center of Gravity



$$\bar{E}_{\mathbf{c}}[\ell] = \frac{\sum_{\ell=0}^{L} \ell \cdot \mathbf{c}^{(\ell)}}{\sum_{\ell=0}^{L} \mathbf{c}^{(\ell)}}$$

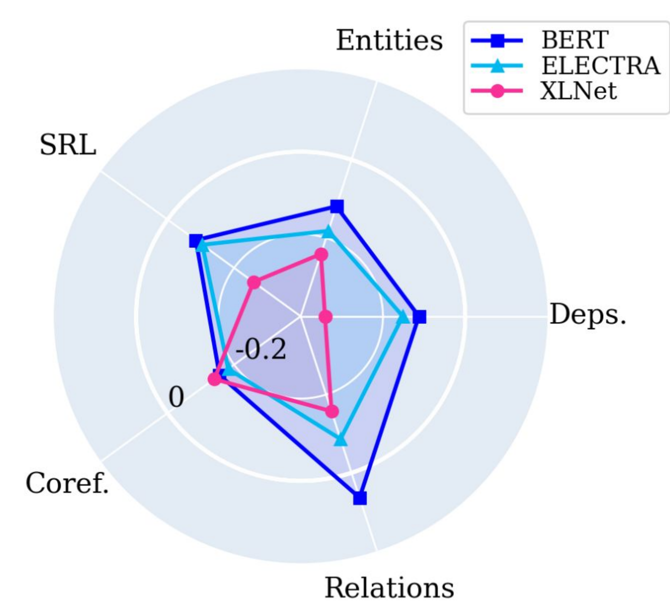
→ XLNet's linguistic knowledge is concentrated in earlier layers than BERT, while ELECTRA's knowledge is mostly accumulated in deeper layers.

→ Recovering input tokens in the final layers of the model in the pre-training objective of BERT and XLNet is a surface task.

→ Whereas the pre-training objective in ELECTRA might be considered as a more semantic task, in which detecting replaced tokens requires more context-aware representations.

Probing Fine-tuned Representations

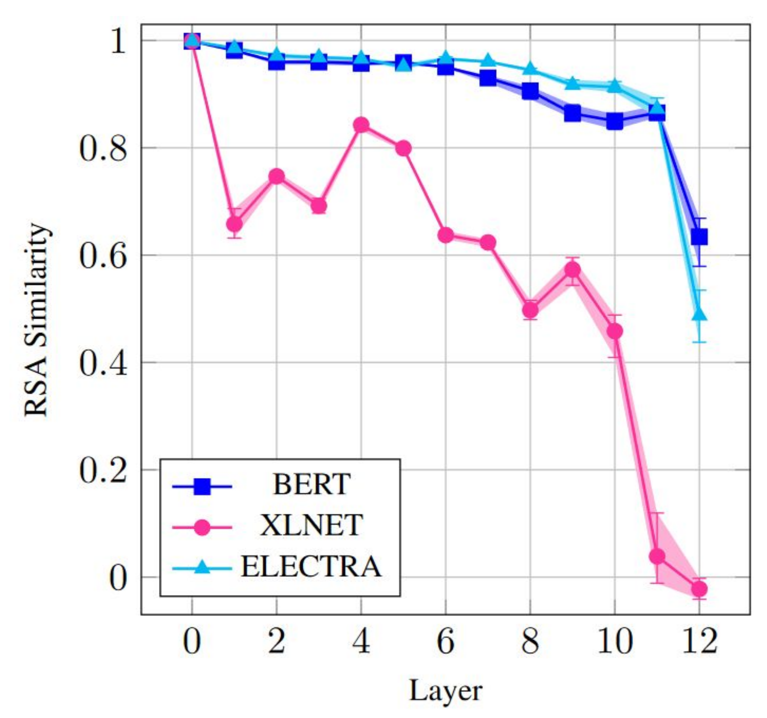
The Change in Centers of Gravity After Fine-tuning



$$\Delta \bar{E}_{\mathbf{c}} = \bar{E}_{\mathbf{c}}^{finetuned} - \bar{E}_{\mathbf{c}}^{pretrained}$$

→ XLNet in most tasks falls back to earlier layers than the two other models because it forgets the most linguistic knowledge in the final layers.

Similarity of The Representations Before and After Fine-tuning



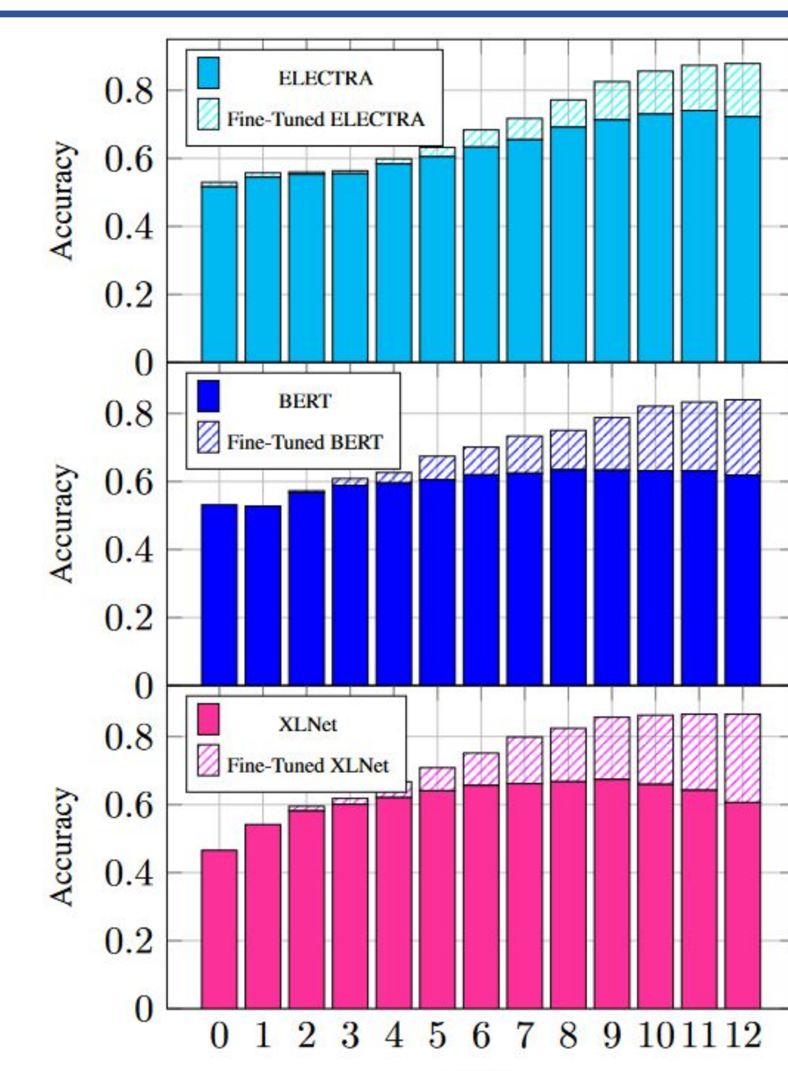
→ XLNet changes drastically during fine-tuning, while in BERT and ELECTRA, only the top layers are primarily affected.

Quality of The Representations for Downstream Tasks

XLNet encodes most essential information for the downstream task in the shallower layers, BERT in the middle ones, and ELECTRA in the deeper layers. ←

XLNet significantly improves performance in its second half of layers, while ELECTRA undergoes smaller adjustments. ←

The changes in layers and their extent are similar to what we saw in the RSA results. ←



Conclusions

- Weight mixing results in edge probing does not lead to reliable conclusions in layer-wise cross model analysis studies and MDL probing is more informative in this setup.
- Compared to BERT, XLNet accumulates linguistic knowledge in its earlier layers, whereas ELECTRA does in its final layers
- ELECTRA undergoes slight changes during fine-tuning, whereas XLNet experiences significant adjustments.

References

- [1] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy.
- [2] Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. 2020. A cross-task analysis of text span representations. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 166–176, Online.
- [3] Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online.