Analysis methods in NLP: Probing

Christopher Potts

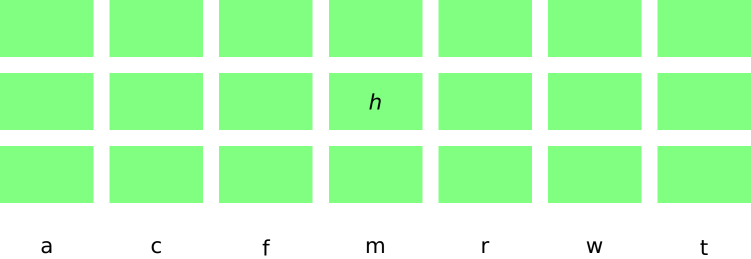
Stanford Linguistics

CS224u: Natural language understanding

Overview

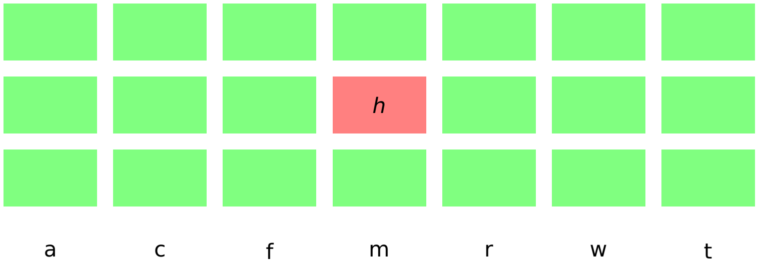
1. Core idea: use supervised models (the probes) to determine what is latently encoded in the hidden representations of our target models.
2. Often applied in the context of BERTology – see especially Tenney et al. 2019.
3. A source of valuable insights, but we need to proceed with caution:
   - A very powerful probe might lead you to see things that aren't in the target model (but rather in your probe).
   - Probes cannot tell us about whether the information that we identify has any *causal* relationship with the target model's behavior.
4. Final section: unsupervised probes.

Core method



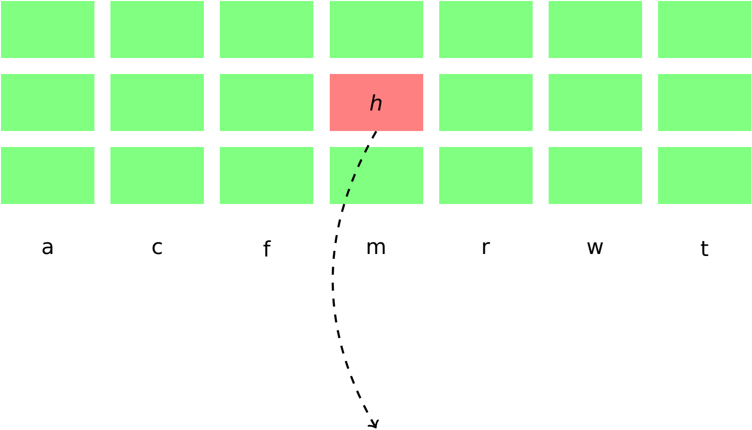
Conneau et al. 2018; Tenney et al. 2019

Core method



Conneau et al. 2018; Tenney et al. 2019

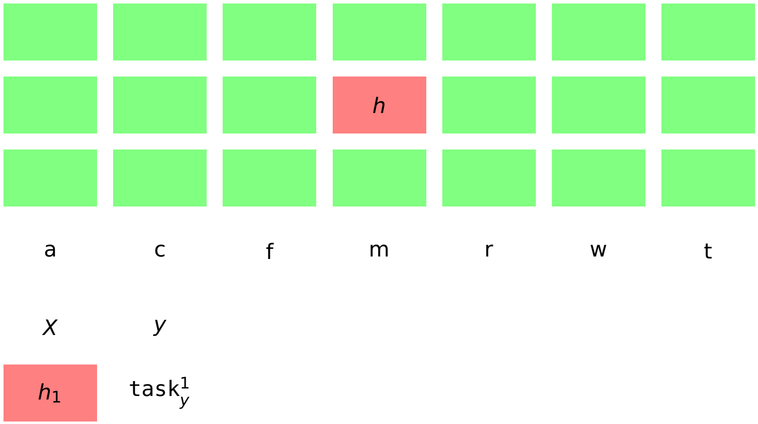
Core method



$$\text{SmallLinearModel}(h) = \textit{task}$$

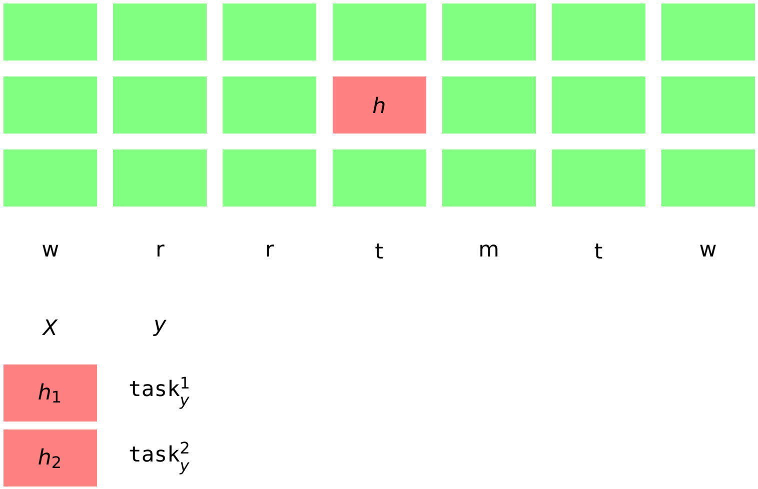
Conneau et al. 2018; Tenney et al. 2019

Core method



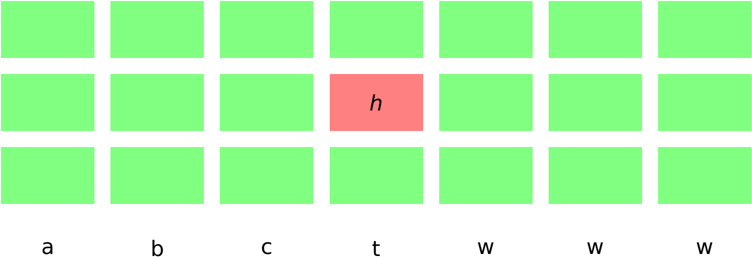
Conneau et al. 2018; Tenney et al. 2019

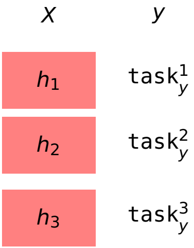
Core method



Conneau et al. 2018; Tenney et al. 2019

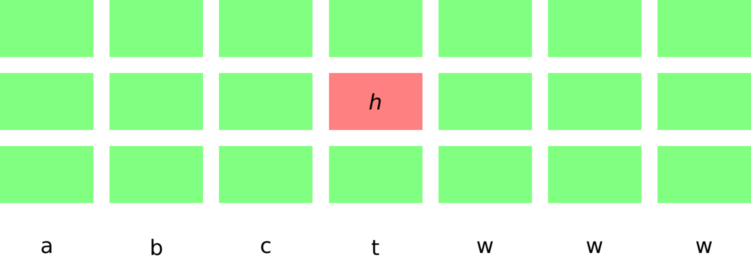
Core method





Conneau et al. 2018; Tenney et al. 2019

Core method



X	y
h_1	$\mathtt{task}_y^1$
h_2	$\mathtt{task}_y^2$
h_3	$\mathtt{task}_y^3$

SmallLinearModel(X, y)

Conneau et al. 2018; Tenney et al. 2019

Core method

Conneau et al. 2018; Tenney et al. 2019

Probing or learning a new model?

1. Probes in the above sense are supervised models whose inputs are frozen parameters of the model we are probing.
2. This is hard to distinguish from simply fitting a supervised model as usual, with a particular choice for featurization.
3. At least some of the information that we identify is likely to be stored in the probe model.
4. More powerful probes might “find” more information – by storing more information in the probe parameters.

Control tasks and probe selectivity

Control task

A random task with the same input/output structure as the target task.

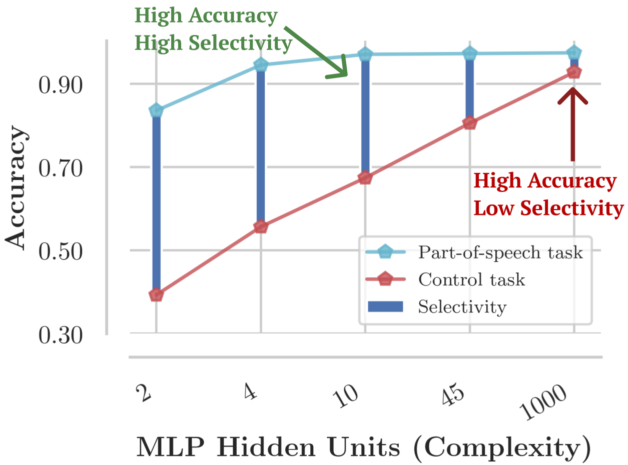
- Word-sense classification: words assigned random fixed senses.
- POS tagging task: words assigned random fixed tags.
- Parsing: assigned edges randomly using simple strategies.

Selectivity

The difference between probe performance on the task and probe performance on the control task.

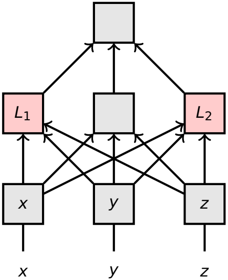
Hewitt and Liang 2019

Control tasks and probe selectivity



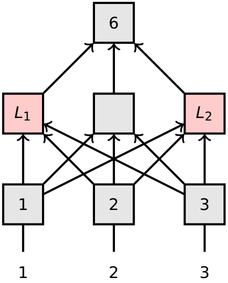
Hewitt and Liang 2019

A fundamental limitation: No causal inference



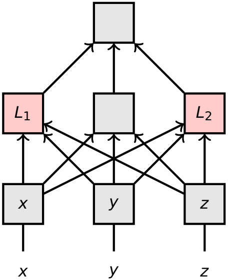
Belinkov and Glass 2019; Vig et al. 2020

A fundamental limitation: No causal inference



Belinkov and Glass 2019; Vig et al. 2020

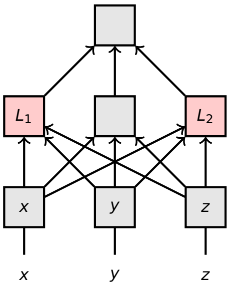
A fundamental limitation: No causal inference



1. Probe L_1: it computes $x + y$

Belinkov and Glass 2019; Vig et al. 2020

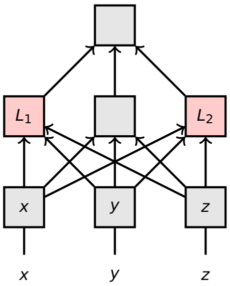
A fundamental limitation: No causal inference



1. Probe L_1: it computes $x + y$
2. Probe L_2: it computes z

Belinkov and Glass 2019; Vig et al. 2020

A fundamental limitation: No causal inference



1. Probe L_1: it computes $x + y$
2. Probe L_2: it computes z
3. But neither has any impact on the output!

$$W_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad W_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad W_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\mathbf{w} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Model:

$$(\mathbf{x}W_1; \mathbf{x}W_2; \mathbf{x}W_3)\,\mathbf{w}$$

Belinkov and Glass 2019; Vig et al. 2020

Unsupervised probes

1. Saphra and Lopez (2019): Singular Vector Canonical Correlation Analysis as a probing technique
2. Clark et al. (2019) and Manning et al. (2020): Inspecting attention weights.
3. Hewitt and Manning (2019) nd Chi et al. (2020): Linear transformations of hidden states to identify latent syntactic structures in BERT.
4. Rogers et al. (2020): extensive discussion of probing and related efforts and what they have revealed about BERT representations.

References I

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. ArXiv:2002.12327.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias.