What Do Transformers Learn in NLP? Recent Insights from Model Analysis

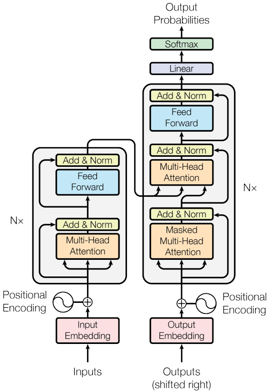
Rico Sennrich

University of Zurich UZH

Edinburgh
University of Edinburgh Natural Language Processing NLP

joint work with **Elena Voita**, Ivan Titov, David Talbot, Fedor Moiseev

Recent Developments in NLP Leaderboard Race



new neural architectures

pre-training becomes mainstream

Open Questions

how do neural architectures work?

Open Questions

(why) does pre-training objective matter?

- BERT-style masked language modeling better than causal language model
  [Lample and Conneau, 2019]
- “Language Modeling Teaches You More Syntax than Translation Does”
  [Zhang and Bowman, 2018]
- ...but multilingual NMT may allow better cross-lingual transfer than mBERT
  [Siddhant et al., 2019]

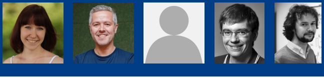
What Do Transformers Learn?

1. how do neural architectures work?

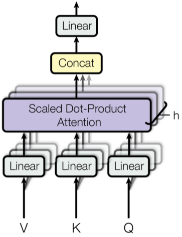
2. (why) does pre-training objective matter?

Analyzing Multi-Head Attention

[Voita, Talbot, Moiseev, Sennrich, Titov, ACL 2019]



- multi-head self-attention is key Transformer component
- questions:
  - how to identify important attention heads?
  - can we prune unimportant ones?
  - which functions do attention heads have?
- spoiler (paper title): “Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned”



Determining Importance of Self-Attention Heads

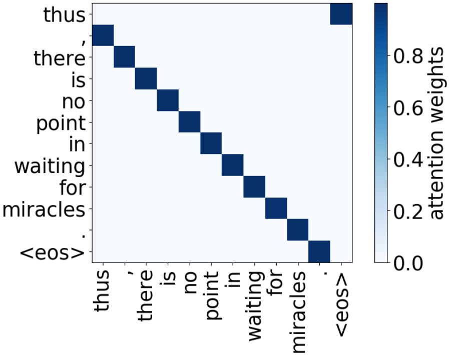
- method 1: assume that “confident” heads (low entropy weight distribution) are important
- method 2: layerwise relevance propagation (LRP) [Bach et al., 2015, Ding et al., 2017]
- method 3: pruning: add regularization term to training objective which deactivates unimportant heads (L_0 norm on scalar gates drawn from Hard Concrete Distribution)

Determining Function of Self-Attention Heads

function of attention head is determined via simple rules

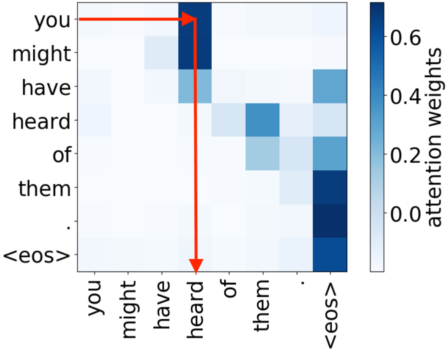
Determining Function of Self-Attention Heads

positional: maximum attention weight is given to specific relative position



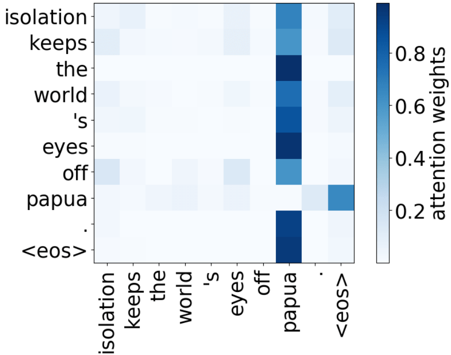
Determining Function of Self-Attention Heads

syntactic: maximum attention weight is given to token in specific dependency relation



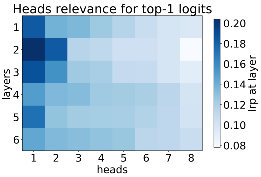
Determining Function of Self-Attention Heads

rare tokens: maximum attention weight is given to least frequent token

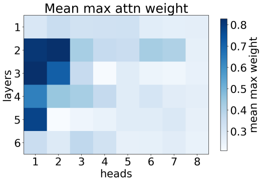


Important Self-Attention Heads Are Specialized

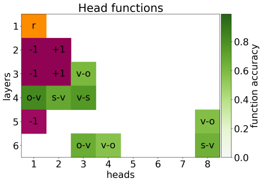
LRP



confidence



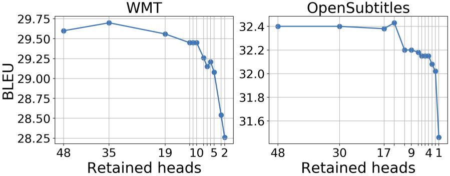
function



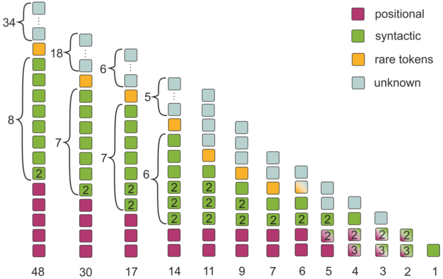
important heads tend to be positional, syntactic, or focus on rare tokens.

Pruning Self-Attention Heads

most heads (in encoder) can be pruned with little quality loss



most heads that survive pruning have one of the functions we identified



What Do Transformers Learn?

1. how do neural architectures work?
2. (why) does pre-training objective matter?

The Evolution of Representations in the Transformer

[Voita, Sennrich, Titov, EMNLP 2019]

compare representations of models only differing in objective function:

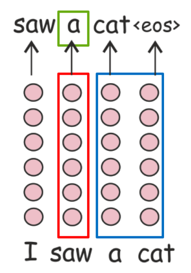
- same architecture (Transformer encoder)
- same (source-side) training data (WMT EN→{DE,FR})

background: information bottleneck principle

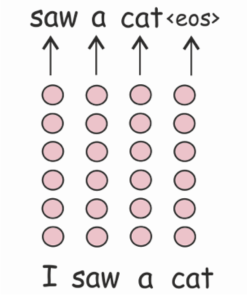
[Tishby et al., 1999, Tishby and Zaslavsky, 2015]

hypothesis: deep neural model learns to compress input representation, retaining information necessary to:

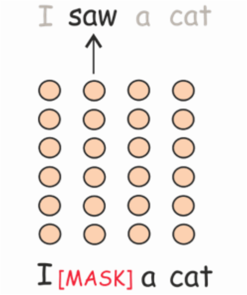
- predict the output label
- build representations of other tokens



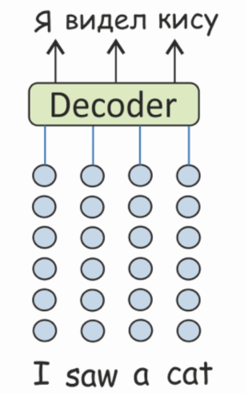
Same Architecture, Different Objective Functions



language model
(causal, LM)

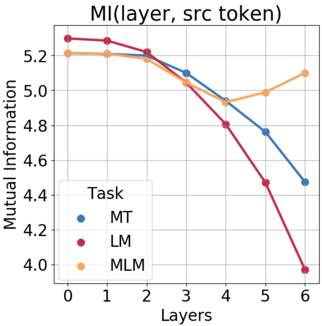


masked language model
(MLM, aka BERT)

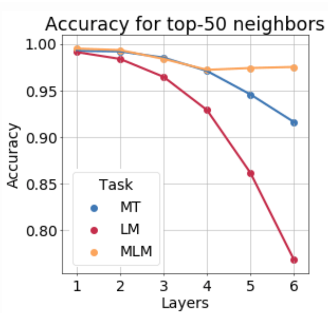


machine translation
(MT)

Is (Input) Token Identity Preserved?



mutual information estimator



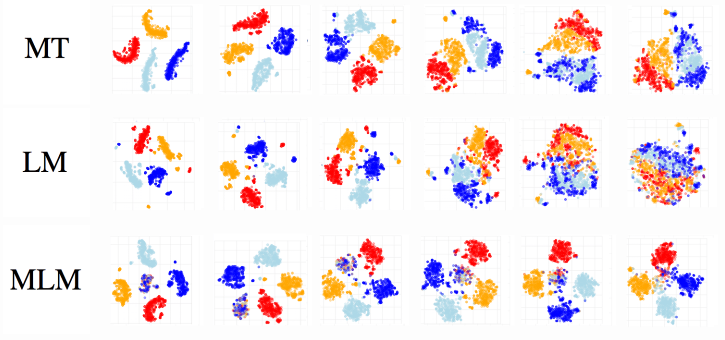
clustering k-nearest neighbor accuracy

how can this be non-monotonic?

- we measure MI/acc per position
- in MLM, information about token is distributed across sentence

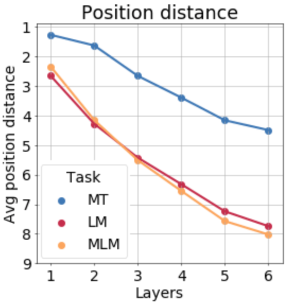
Is Token Identity Preserved?

representations of is, are, were, was
(t-sne projection)

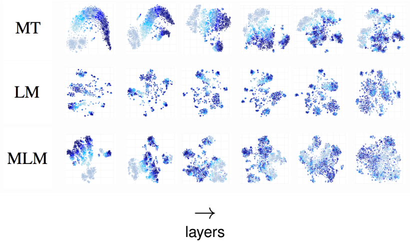


→
layers

MT Preserves Token Position the Most

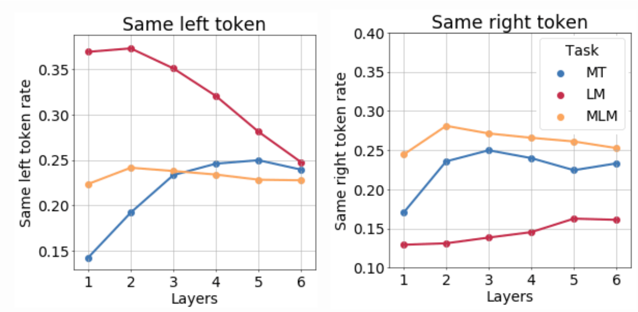


distance between position of a representation and its k-nearest neighbors



visualisation via t-sne projection

Causal LM: Past and Future



- lower layers in (causal) LM represent input (left token)
- higher layers form representations predictive of output (right token)

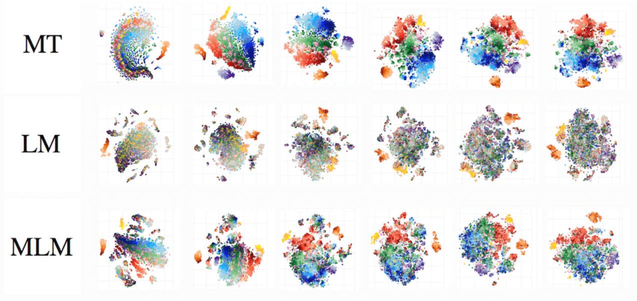
Conclusions

- analysis of pruning of self-attention heads could lead to:
  - model interpretability
  - efficiency
- learning objective affects information flow in Transformer
- analysis of representations complements probing experiments
  → can be used to explain why:
  - some pre-training objectives are more successful
  - lower layers may perform better in some probing tasks than higher ones

Thank you for your attention

more content in blog posts and papers!

- `https://lena-voita.github.io/posts/acl19_heads.html`
- `https://lena-voita.github.io/posts/emnlp19_evolution.html`



t-sne clustering of CCG tags

Bibliography I

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015).
On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.
PloS one, 10(7):e0130140.

Ding, Y., Liu, Y., Luan, H., and Sun, M. (2017).
Visualizing and understanding neural machine translation.
In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1150–1159, Vancouver, Canada. Association for Computational Linguistics.

Lample, G. and Conneau, A. (2019).
Cross-lingual language model pretraining.

Siddhant, A., Johnson, M., Tsai, H., Arivazhagan, N., Riesa, J., Bapna, A., Firat, O., and Raman, K. (2019).
Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation.

Tishby, N., Pereira, F. C., and Bialek, W. (1999).
The information bottleneck method.
In Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing, pages 368–377.

Tishby, N. and Zaslavsky, N. (2015).
Deep learning and the information bottleneck principle.
2015 IEEE Information Theory Workshop (ITW), pages 1–5.

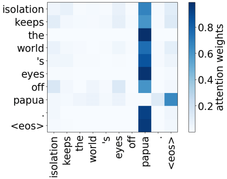
Zhang, K. and Bowman, S. (2018).
Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis.
In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

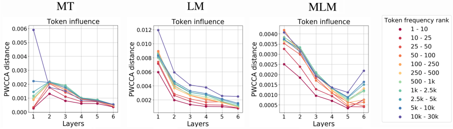
High Importance of Rare Tokens: Overfitting?

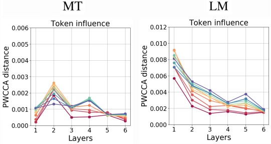
[Voita et al. ACL 2019]: some heads specialize on rare tokens

[Voita et al. EMNLP 2019]: rare tokens highly influential

but effect goes away after randomly swapping 10% of tokens


isolation
keeps
the
world
's
eyes
off
papua
<eos>
attention weights


MT
LM
MLM
Token influence
PWCCA distance
Layers
Token frequency rank
500 - 1k
1k - 2.5k
2.5k - 5k
5k - 10k
10k - 30k


MT
LM
Token influence
PWCCA distance
Layers