Neural Language Models

The New Frontier of Natural Language Understanding

Gabriele Sarti

University of Trieste, SISSA & ItaliaNLP Lab

StaTalk 2019

Table of Contents

A Problem of Representations

Representation Learning is central for AI, neuroscience and semantics.

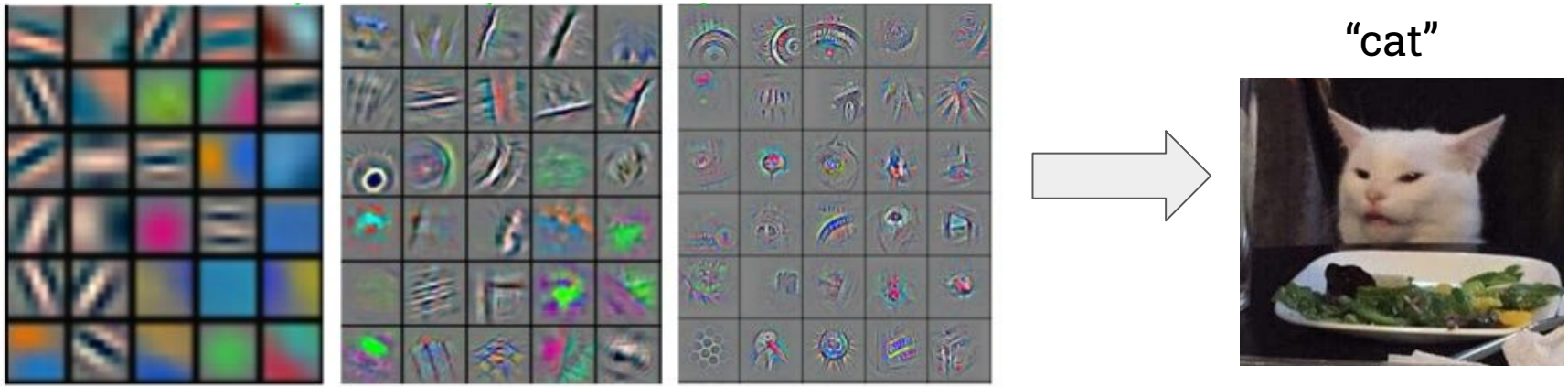
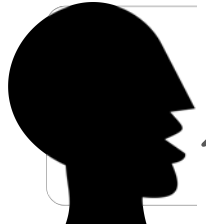


Figure 1: Hierarchy of features visualized for a CNN trained on ImageNet (Wan et al. 2013).

- For images, hierarchical representations exploiting locality of features.
- What about language? **Not so easy!**
- **Distributional Hypothesis:** Semantically related words are distributed in a similar way and occur in similar contexts.



"You shall know a word by the company it keeps" (J.R. Firth, 1957)

Introduced in linguistics by Harris 1954, currently explored in cognitive science.

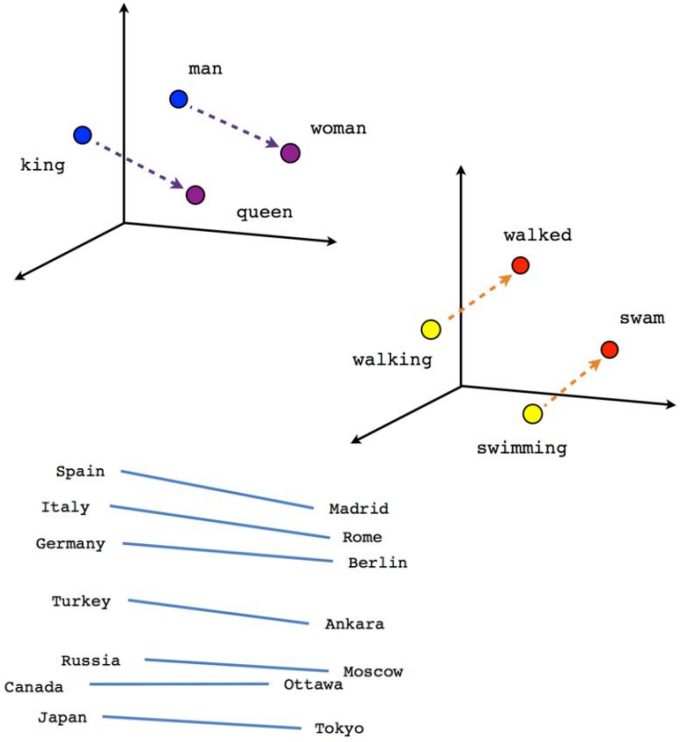
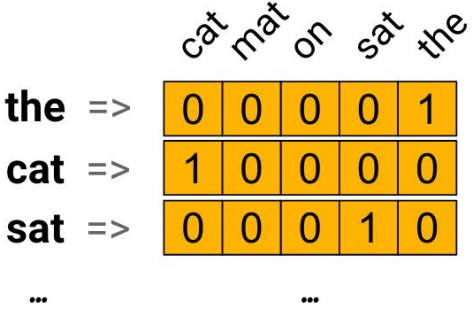


Figure 2: Linear word relations from Tensorflow Tutorials.

Early Years: Statistical Representations for NLP



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 3: Some examples of statistical and machine learning approaches to learn text representations. ***From left to right:*** one-hot encoding of vocabulary terms, sentence-level lexical, syntactic and morpho-syntactic features and term frequency-inverse document frequency (tf-idf) formula.

Recent Times: Unsupervised Representations

- **Word embeddings**: Dense vector representations of words learned by optimizing a loss function.
- **Main problems**: biases and disambiguating polysemy.

Well-known examples of pretrained static embeddings are **Word2Vec** (Mikolov et al. 2013), **GloVe** (Pennington et al. 2014) and **FastText** (Bojanowsky et al. 2017).

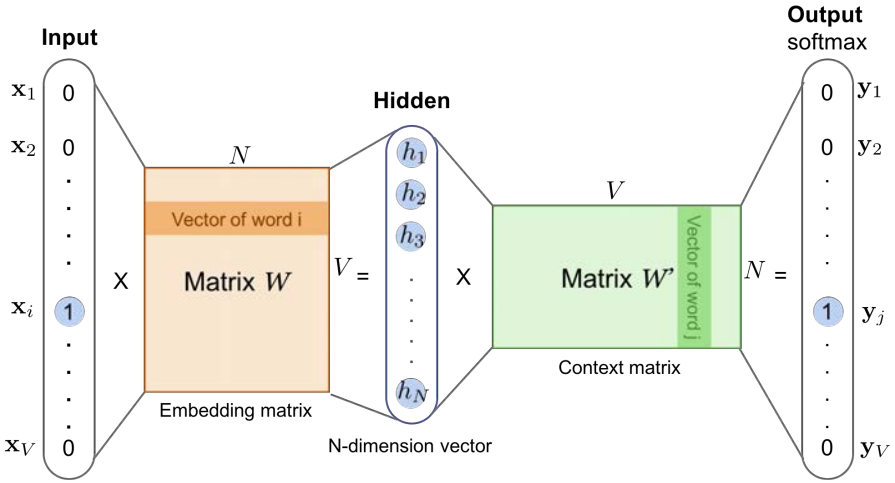


Figure 4: A visual representation of the Skip-gram method used to train Word2Vec embeddings. Input is a one-hot of each pair of target-context word in the sliding window. W and W' are target and context representations.

Context is Key for Meaning

- **Contextual Embedding:** Embeddings as functions of the entire input sentence.
- **Idea**: Use a task to induce contextual representations inside a neural network exploiting sentence information.

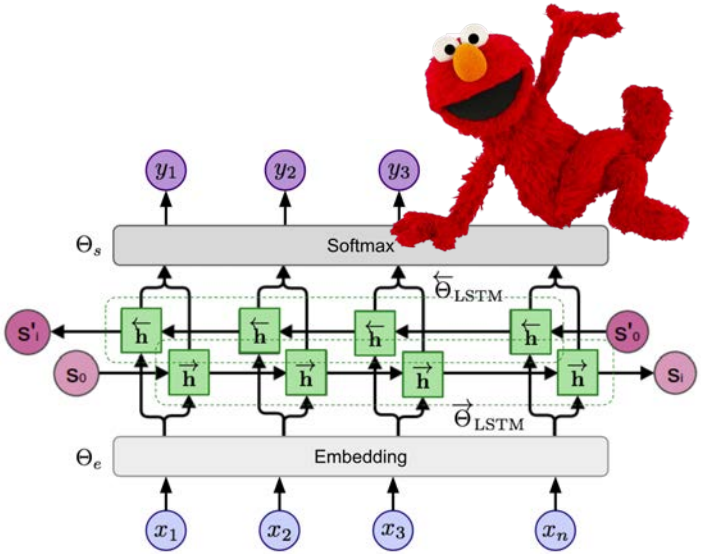


Figure 5: A bidirectional LSTM (Hochreiter and Schmidthuber, 1997) representing the base model used for ELMo contextual word embeddings. ELMo is a task-specific combination of the internal representations in the biLSTM and uses regular and backward LM.

Introduced by CoVe (McCann et al. 2017) for the machine translation task, popularized by ELMo (Peters et al. 2018) for language modeling.

The Language Modeling Task

- **Language Modeling (LM):** Predict future token given history.

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1}) \qquad \mathcal{L} = -\sum_{i=1}^{n} \log p(x_i | x_1, \ldots, x_{i-1}; \Theta)$$

Figure 6: From left to right, the joint probability of a sentence defined as single word probabilities given previous context and the loss function used for LMs. Minimizing the log likelihood corresponds to maximizing the probability of correctly guessing words.

- **Why LM?** Unsupervised, requires knowledge and improve generalization. Alternatives: Masked Language Modeling and MT.

LMs are Unsupervised Multitask Learners

- **Problem:** ELMo still require task-specific models to leverage contextual embeddings.
- **Solutions:**
  - Task-specific fine-tuning in ULMFiT (Howard & Ruder 2018) inspired by ImageNet pre-training for CV tasks.
  - Generative pre-training of a transformer LM (Radford et al. 2018) with possible supervised fine-tuning.
- **Results:** SOTA on most language-related tasks, from sentiment analysis to NER.

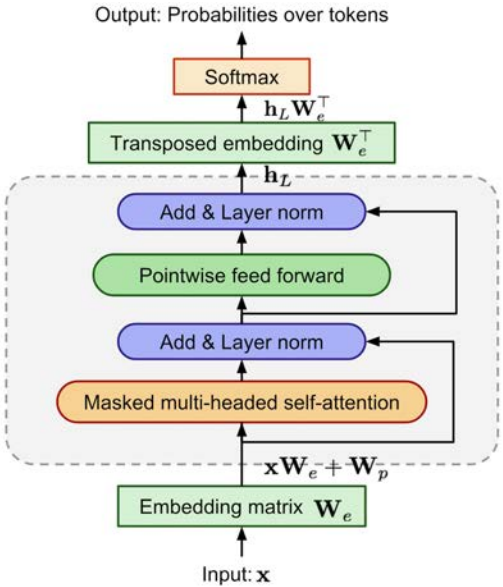


Figure 7: The transformer decoder in OpenAI GPT. The grey block represent the transformer block described in Vaswani et al. 2017 and can be stacked.

Attention Is All You Need

- RNNs are problematic since hidden states must be computed sequentially.
- **Attention mechanisms** were used in conjunction with RNN to capture long-range relations, inspired by MT.
- Transformers (Vaswani et al. 2017) use only attention and fully connected layers to create highly scalable networks capturing distant patterns.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Figure 8: Scaled dot-product self-attention introduced by Vaswani et al. **Queries** Q and **keys** K have dimension d_k This type of attention is efficient thanks to matrix multiplications and can be augmented with multiple heads capture information from different representation subspaces.

More Data and Parameters Are All You Need (?)

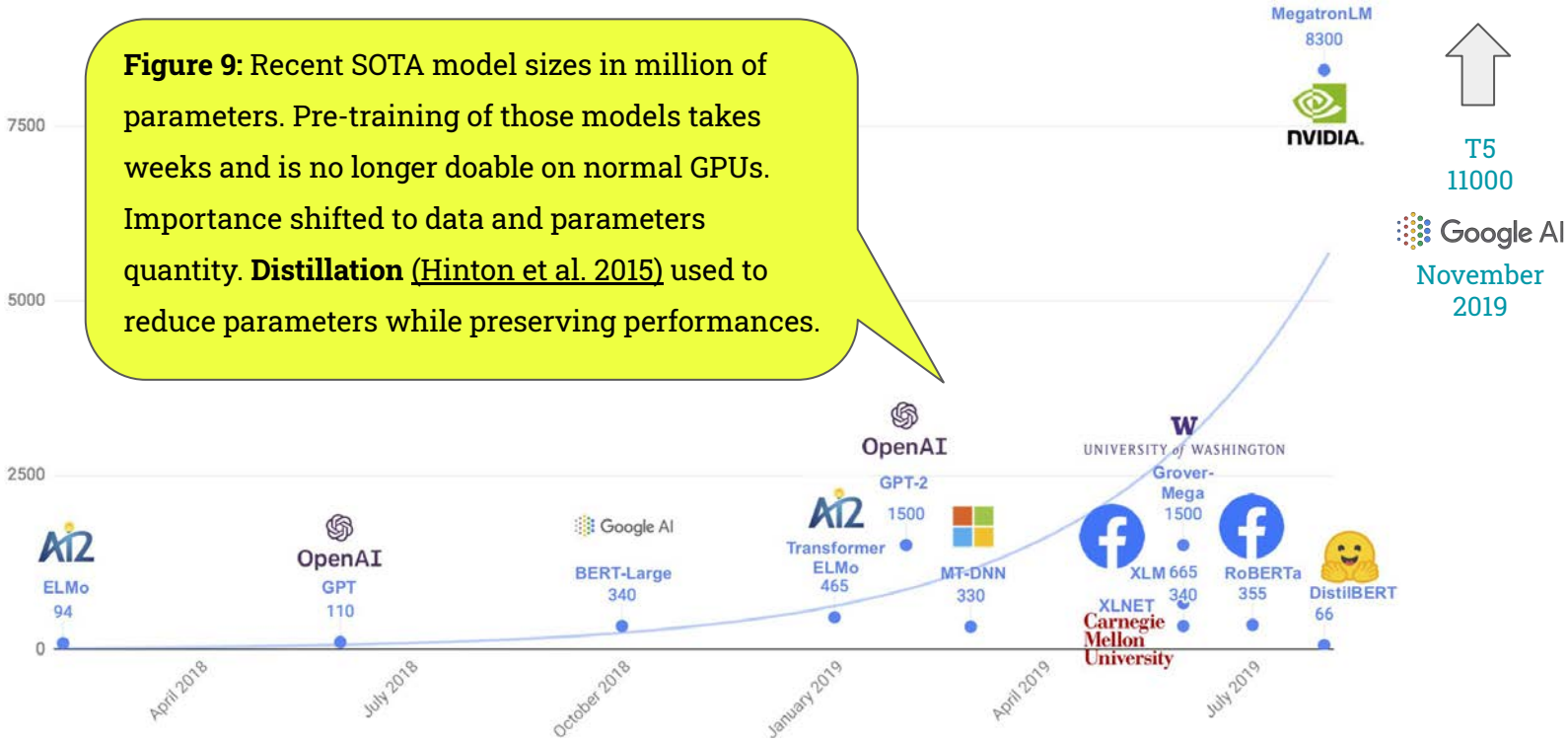
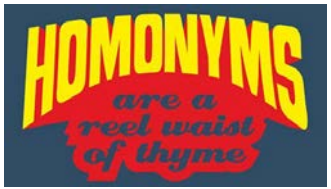


Figure 9: Recent SOTA model sizes in million of parameters. Pre-training of those models takes weeks and is no longer doable on normal GPUs. Importance shifted to data and parameters quantity. **Distillation** (Hinton et al. 2015) used to reduce parameters while preserving performances.

What Does it Mean to Understand Language?



Disambiguation

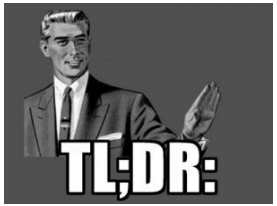
Answering

Well-formedness

UNDERSTANDING

Abstract Reasonings

Inference



Summarizing

Modeling is not Understanding

- **Perplexity:** Exponentiation of the entropy of a discrete probability distribution. How “unsure” the model is in predicting next event.
- Used as a quality measure for language models, the lower the better.

$$2^{H(s)} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 p(w_i)} = (2^{\sum_{i=1}^{N}\log_2 p(w_i)})^{-\frac{1}{N}} = (p(w_1)\dots p(w_N))^{-\frac{1}{N}}$$

Figure 10: Perplexity of a sentence s assuming each word has the same frequency -1/N.

- Lower perplexity doesn’t imply better understanding. Performance on NLU tasks can be improved by statistical cues (Niven & Kao 2019)
- Need to evaluate understanding and generalization in other ways.

Current Directions: NLU & NLI Benchmarks

Figure 11: Some of the most popular benchmarks used to evaluate LM generalization capabilities. GLUE and SuperGLUE (Wang et al.) focus on language understanding tasks, decaNLP (McCann et al. 2018) is a set of 10 general NLP tasks and SWAG/HellaSWAG (Zellers et al.) focus on inference with adversarial filtering and grounded dialogue.

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-g	AX-b
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	99.3/99.7	76.6
2	T5 Team - Google	T5		88.9	91.0	93.0/96.4	94.8	88.2/62.3	93.3/92.5	92.5	76.1	93.8	92.7/91.9	65.6
3	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	91.0/78.1	57.9
4	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	97.8/57.3	29.6

Figure 12: SuperGLUE leaderboard as of November 18, 2019. Despite having been created to be tricky for transformer LMs, current models are approaching human performances, suggesting a new direction will soon be needed.

Interpreting and Probing Language Models

- **Explainability** is a common trend in black-box deep learning approaches.
- For NLP models, its declinations are:
  - Probing language models for linguistic information (Hewitt et al. 2019, Jawahar et al. 2019, Lin et al. 2019, Tenney et al. 2019).
  - Study attention activations and the evolution of representations (Voita et al. 2019, Vig et al. 2019, Michel et al. 2019, Clark et al. 19)

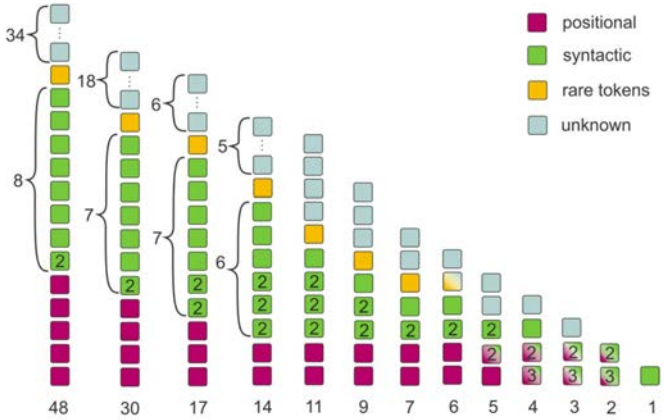


Figure 13: An analysis of attention heads specialized behavior for different linguistic information. Specialized heads do the heavy lifting, the rest can be pruned (Voita et al. 2019).

Perspectives: NLP Rediscovers the Human Brain

- Availability of cerebral data from different sources (EEG, eye-tracking, fMRI).
  - Using neuroscientific techniques (e.g. RSA by Kriegeskorte et al. 2008) to compare brain and LM activations (Abnar et al. 2019, Abdou et al. 2019, Gauthier and Levy 2019)
  - Use human signals to improve model behavior (Hollenstein et al. 2019, Barrett et al. 2018)
- **Target:** More parsimonious models that achieve human-like, interpretable behavior.

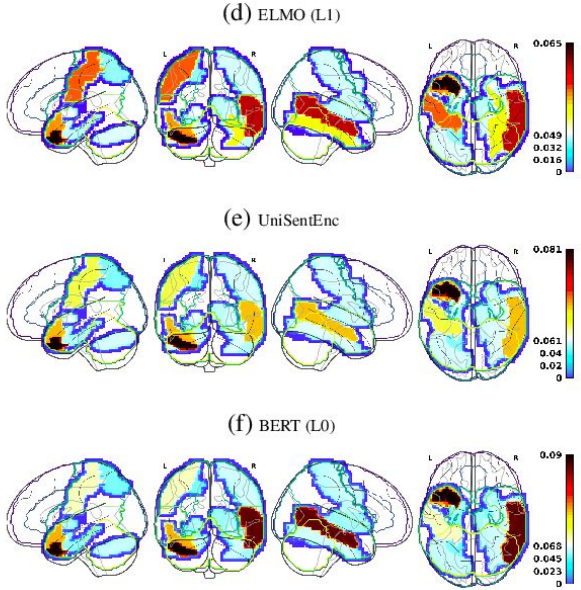


Figure 14: RSA between activations in different model layers and in a human subject brain. LSTM seems to have a more similar behavior with respect to transformers. (Abnar et al. 2019)

Thanks for your $\text{softmax}(\frac{\mathbf{QK}^\top}{\sqrt{n}})\mathbf{V}$!

Gabriele Sarti

gabriele.sarti996@gmail.com

gsarti.com

@gsarti_

References

A Problem of Representations

- Harris Z., "Distributional structure", Words, 1954
- Firth J.R., "**A synopsis of linguistic theory 1930-1955**", Studies in Linguistic Analysis, 1957
- Wan et al., "Regularization of Neural Networks using DropConnect", PMLR, 2013

Recent Times: Unsupervised Representations

- Mikolov et al., "Efficient Estimation of Word Representations in Vector Space", ICLR 2013
- Pennington et al., "GloVe: Global Vectors for Word Representation", EMNLP 2014
- Bojanowsky et al., "Enriching Word Vectors with Subword Information", TACL 2017

Context is Key for Meaning

- Hochreiter & Schmidhuber, "Long Short-Term Memory", Neural Computation 1997

- McCann et al., "Learned in Translation: Contextualized Word Vectors", arXiv 2017
- Peters et al., "Deep Contextualized Word Representations", NAACL 2018

LMs Are Unsupervised Multitask Learners

- Vaswani et al., "Attention is All You Need", NeurIPS 2017
- Howard & Ruder, "Universal Language Model Fine-tuning for Text Classification", ACL 2018
- Radford et al., "Improving Language Understanding by Generative Pre-Training", Published 2018

More Data and Parameters Are All You Need (?)

- Hinton et al. "Distilling the Knowledge in a Neural Network", arXiv 2015
- Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL 19
- Radford et al., "Language Models are Unsupervised Multitask Learners", Published 2019 **(GPT-2)**
- Liu et al., "Multi-Task Deep Neural Networks for Natural Language Understanding", ACL 2019 **(MT-DNN)**
- Yang et al., "XLNet: Generalized Autoregressive Pretraining for Language Understanding", NeurIPS 2019

- Lample & Conneau, "Cross-lingual Language Model Pretraining", NeurIPS 2019 **(XLM)**
- Zellers et al., "Defending Against Neural Fake News", NeurIPS 2019 **(Grover)**
- Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv 2019
- Sahn et al., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", arXiv 2019
- NVidia, "MegatronLM: Training Billion+ Parameter Language Models Using GPU Model Parallelism", 2019
- Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", arXiv 19 **(T5)**

Modeling is Not Understanding

- Niven & Kao, "Probing Neural Network Comprehension of Natural Language Arguments", ACL 2019

Current Directions: NLU & NLI Benchmarks

- Wang et al., "GLUE: A Multitask Benchmark and Analysis Platform for NLU", ICLR 2019
  - "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems", arXiv 2019
- Zellers et al., "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference", EMNLP 2018

- ➢ "HellaSwag: Can a Machine Really Finish Your Sentence?", ACL 2019
- McCann et al., "The Natural Language Decathlon: Multitask Learning as Question Answering", arXiv 2018

Interpreting and Probing Language Models

- Hewitt et al., "A Structural Probe for Finding Syntax in Word Representations", NAACL 2019
- Jawahar et al., "What Does BERT Learn about the Structure of Language?", ACL 2019
- Lin et al., "Open Sesame: Getting Inside BERT's Linguistic Knowledge", ACL 2019
- Tenney et al., "BERT Rediscovers the Classical NLP Pipeline", ACL 2019
- Voita et al., "Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned", ACL 2019
- Voita et al., "The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives", IJCNLP 2019
- Michel et al., "Are Sixteen Heads Really Better than One?", NeurIPS 2019
- Vig et al., "Analyzing the Structure of Attention in a Transformer Language Model", ACL 2019
- Clark et al. "What Does BERT Look at? An Analysis of BERT's Attention", BlackBoxNLP 2019

Perspectives: NLP Rediscovers the Human Brain

- Kriegeskorte et al., "Representational Similarity Analysis", Frontiers 2008
- Abnar et al., "Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains", BlackBoxNLP 2019
- Abdou et al., "Higher-order Comparisons of Sentence Encoder Representations", EMNLP 2019
- Gauthier & Levy, "Linking artificial and human neural representations of language", IJCNLP 2019
- Hollenstein et al., "Advancing NLP with Cognitive Language Processing Signals", arXiv 2019
- Barrett et al., "Sequence Classification with Human Attention", CoNLL 2018

General Inspiration and Structure of the Talk

- Weng L., "Learning Word Embeddings", Blog Post, 2017
- Weng, L., "Attention? Attention!", Blog post, 2018
- Weng L., "Generalized Language Models", Blog Post, 2019