# Analysis of Model’s Inner Workings

## Natural Language Processing

(based on revision of John Hewitt Lectures)

# Announcement

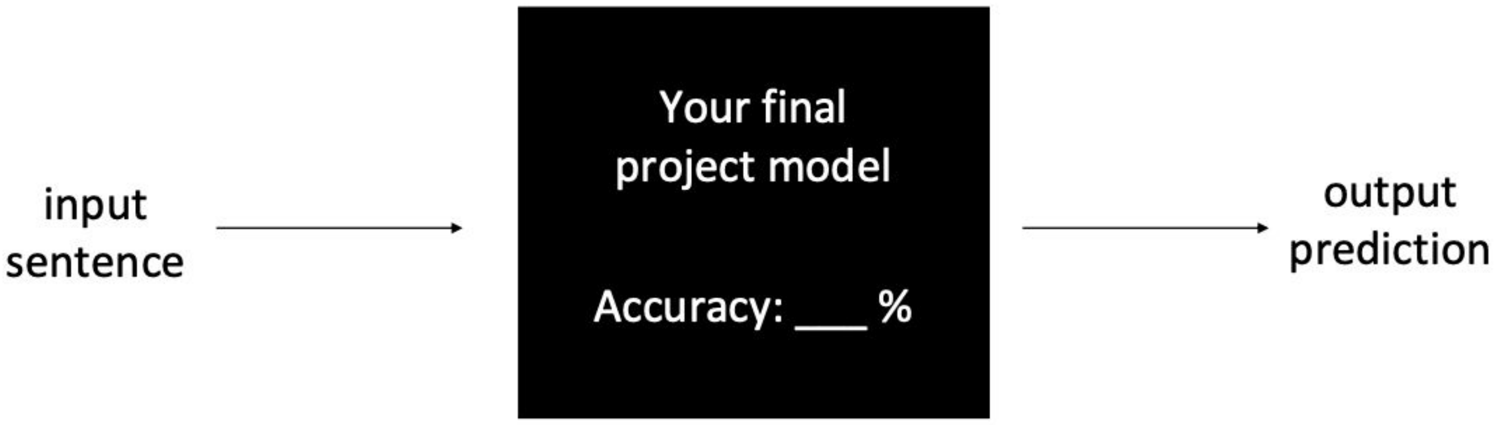
- TA announcements (if any)...

# Researchers wear many hats (type of contributions)

- **Proof**
  - Mathematicians: those who are very good in math and wants to prove/disprove that something really/does not really work
- **New / Improved Approaches**
  - Engineers: those who always ponder how we can do things differently
- **Novel / Creative Applications**
  - Practitioners:  applying existing models in an interesting problem / unique applications
- **Analysis**
  - Scientists: ← trying to understand why the model works and when will it not work; sometimes try to break stuff!

today!

# Using one single metric can be misleading



input sentence

Your final project model

Accuracy: ___ %

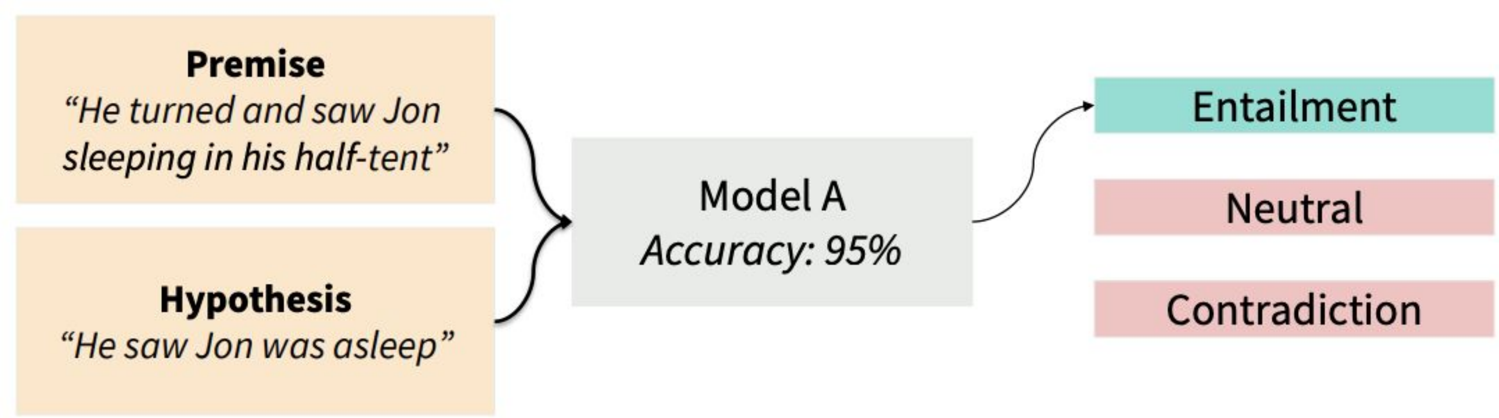
output prediction

# Analysis studies

Today we will take a look at some analysis work, and hopefully, we can develop an intuition why analysis is important and how model are being inspected at

# Diagnostic Test Set [McCoy et al., 2019]

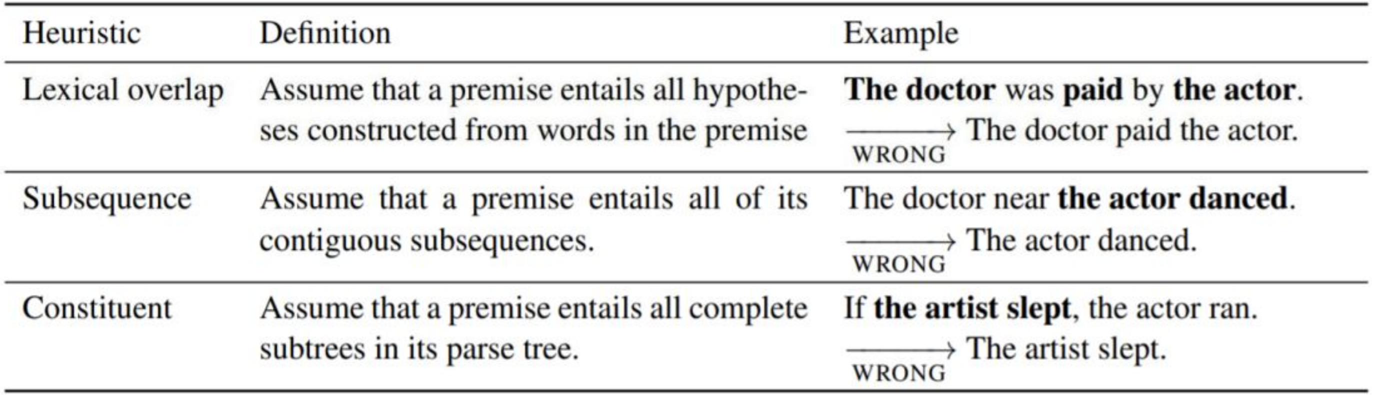
Recall the **natural language inference task**, as encoded in the **Multi-NLI dataset**.



A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, Williams et al., 2018, https://cims.nyu.edu/~sbowman/multinli/paper.pdf

# Diagnostic Test Set

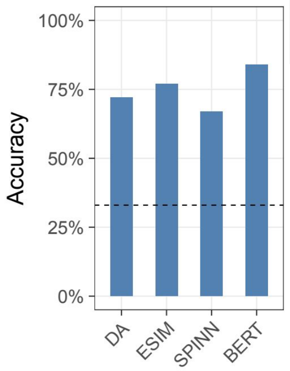
McCoy et al., 2019 developed a diagnostic test set called **HANs** (Heuristic Analysis for NLI Systems) to analyze specific skill or capacity of the model



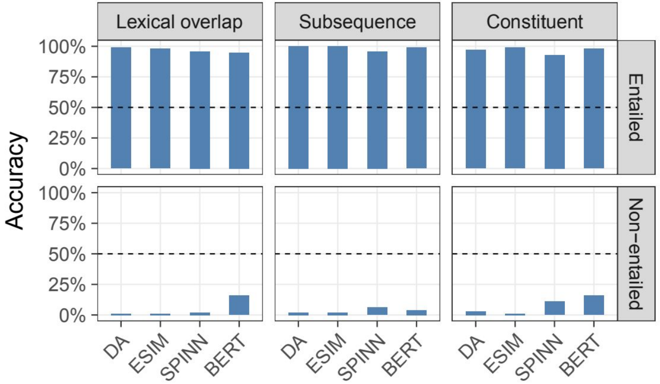
| Heuristic | Definition | Example |
|---|---|---|
| Lexical overlap | Assume that a premise entails all hypotheses constructed from words in the premise | **The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor. |
| Subsequence | Assume that a premise entails all of its contiguous subsequences. | The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced. |
| Constituent | Assume that a premise entails all complete subtrees in its parse tree. | If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept. |

# Diagnostic Test Set

McCoy et al., 2019 took 4 strong MNL models, with the following accuracies on the original test set (in-domain)



Evaluating on HANS, for the entailment conditions, accuracy is very high. But on the non-entailment conditions, accuracy is very very low.



# Syntax Sensitive Dependencies [Kuncoro et al., 2018]

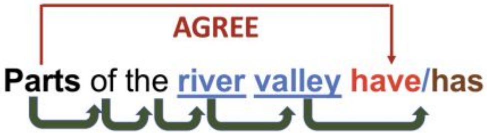
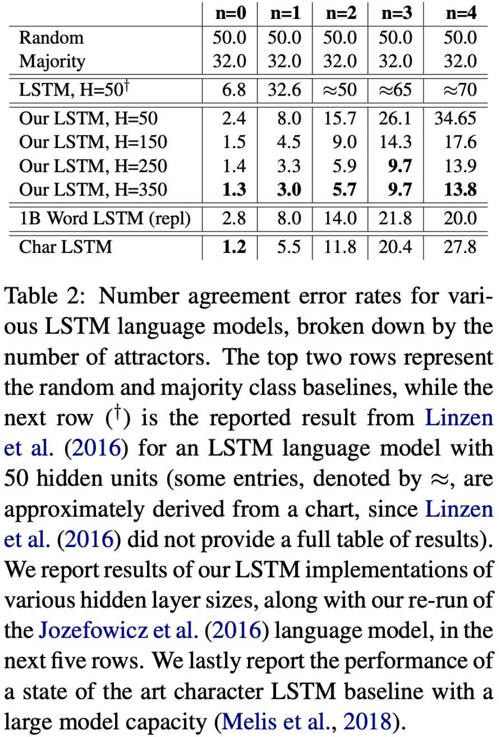


Figure 1: An example of the number agreement task with two attractors and a subject-verb distance of five.

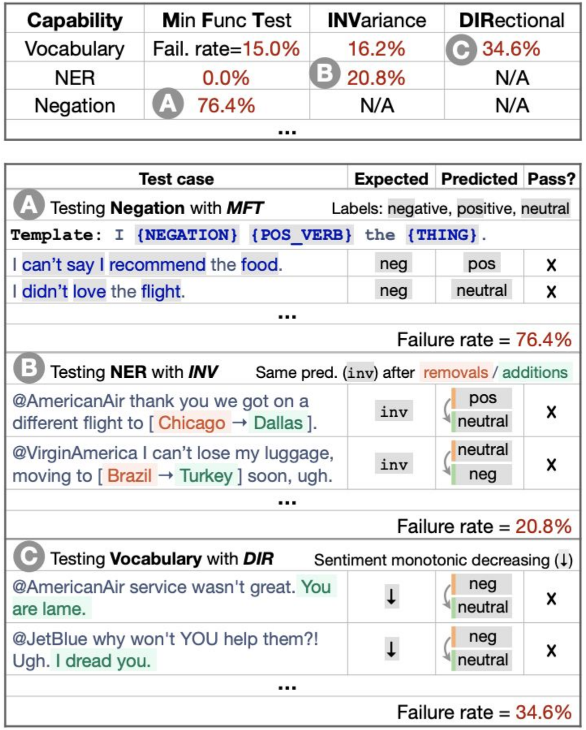


| | n=0 | n=1 | n=2 | n=3 | n=4 |
|---|---|---|---|---|---|
| Random | 50.0 | 50.0 | 50.0 | 50.0 | 50.0 |
| Majority | 32.0 | 32.0 | 32.0 | 32.0 | 32.0 |
| LSTM, H=50† | 6.8 | 32.6 | ≈50 | ≈65 | ≈70 |
| Our LSTM, H=50 | 2.4 | 8.0 | 15.7 | 26.1 | 34.65 |
| Our LSTM, H=150 | 1.5 | 4.5 | 9.0 | 14.3 | 17.6 |
| Our LSTM, H=250 | 1.4 | 3.3 | 5.9 | **9.7** | 13.9 |
| Our LSTM, H=350 | **1.3** | **3.0** | **5.7** | **9.7** | **13.8** |
| 1B Word LSTM (repl) | 2.8 | 8.0 | 14.0 | 21.8 | 20.0 |
| Char LSTM | **1.2** | 5.5 | 11.8 | 20.4 | 27.8 |

Table 2: Number agreement error rates for various LSTM language models, broken down by the number of attractors. The top two rows represent the random and majority class baselines, while the next row (†) is the reported result from Linzen et al. (2016) for an LSTM language model with 50 hidden units (some entries, denoted by ≈, are approximately derived from a chart, since Linzen et al. (2016) did not provide a full table of results). We report results of our LSTM implementations of various hidden layer sizes, along with our re-run of the Jozefowicz et al. (2016) language model, in the next five rows. We lastly report the performance of a state of the art character LSTM baseline with a large model capacity (Melis et al., 2018).

LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better, Kuncoro et al., 2018, https://aclanthology.org/P18-1132.pdf

# Unit Test Suites  [Ribeiro et al., 2020]



| Capability | Min Func Test | INVariance | DIRectional |
|---|---|---|---|
| Vocabulary | Fail. rate=15.0% | 16.2% | C 34.6% |
| NER | 0.0% | B 20.8% | N/A |
| Negation | A 76.4% | N/A | N/A |
| ... | | | |

| Test case | Expected | Predicted | Pass? |
|---|---|---|---|
| A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral | | | |
| `Template: I {NEGATION} {POS_VERB} the {THING}.` | | | |
| I can’t say I recommend the food. | neg | pos | X |
| I didn’t love the flight. | neg | neutral | X |
| ... | | | |
| Failure rate = 76.4% | | | |
| B Testing **NER** with ***INV*** Same pred. (`inv`) after removals / additions | | | |
| @AmericanAir thank you we got on a different flight to [ Chicago → Dallas ]. | `inv` | pos<br>neutral | X |
| @VirginAmerica I can’t lose my luggage, moving to [ Brazil → Turkey ] soon, ugh. | `inv` | neutral<br>neg | X |
| ... | | | |
| Failure rate = 20.8% | | | |
| C Testing **Vocabulary** with ***DIR*** Sentiment monotonic decreasing (↓) | | | |
| @AmericanAir service wasn't great. You are lame. | ↓ | neg<br>neutral | X |
| @JetBlue why won't YOU help them?! Ugh. I dread you. | ↓ | neg<br>neutral | X |
| ... | | | |
| Failure rate = 34.6% | | | |

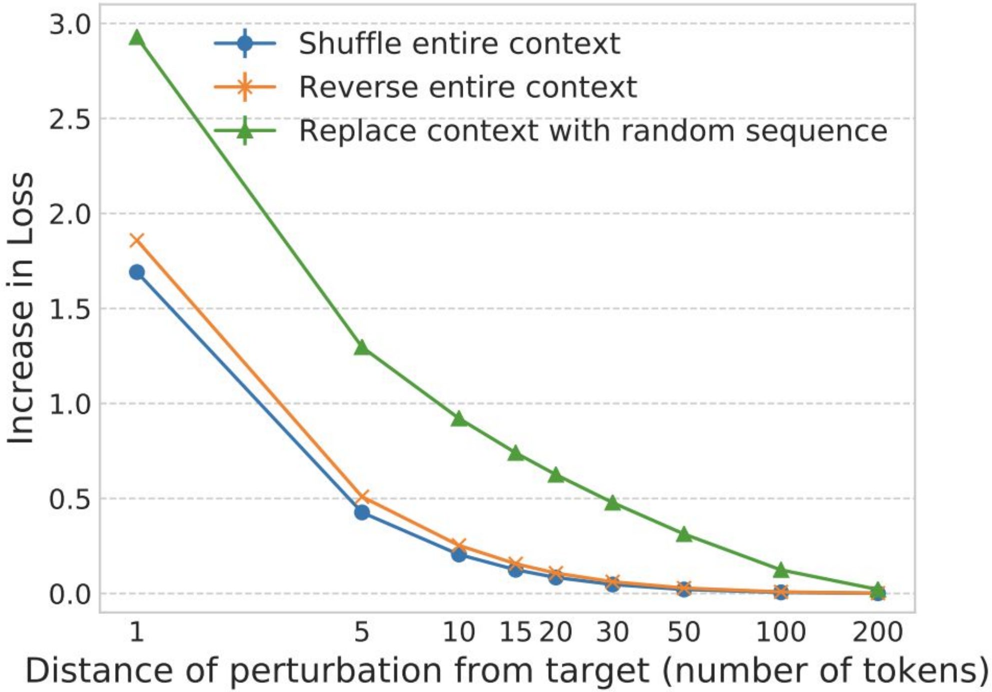
Ribeiro et al., 2020 showed **ML engineers working on a sentiment analysis product** an interface with categories of linguistic capabilities and types of test.

They found a **bunch of bugs** using this method!

- MFT - minimum functionality test
- INV - invariance test
- DIR - directional expectation test.

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, Ribeiro et al., 2020,https://arxiv.org/pdf/2005.04118.pdf

# Does LSTM really use long-distance context? [Khandelwal et al., 2018]



- shuffle or remove all contexts farther than $k$ words away for multiple values of $k$ and see at which $k$ the model's predictions start to get worse!

- Seems like history around 50 words apart is not as impactful

Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context, Khandelwal et al., 2018,https://arxiv.org/pdf/1805.04623.pdf

# Knowledge Evaluation  [Petroni et al., 2020]

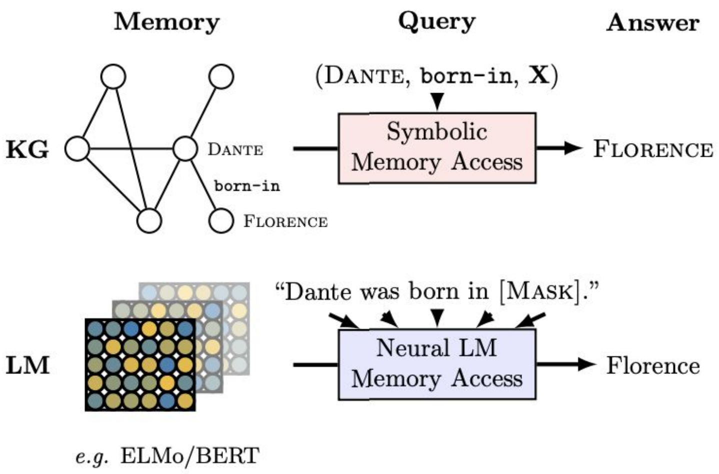


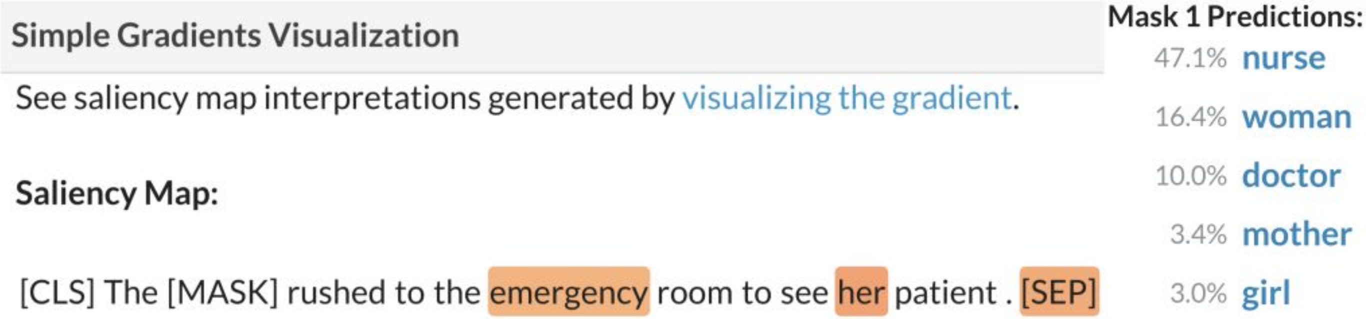
Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

- **How much does pretrained model retain knowledge**?

- Petroni et al., 2020 studies how well pretrained model answers some of NLP tasks questions
  - Amazed that **without finetuning**, BERT contains relational knowledge competitive with traditional NLP methods (oracle-based knowledge)
  - BERT also does remarkably well on open-domain question answering

Language Models as Knowledge Bases?, Petroni  et al., 2020, https://arxiv.org/pdf/1909.01066.pdf

# Saliency maps [Wallace et al., 2019]



**Simple Gradients Visualization**

See saliency map interpretations generated by visualizing the gradient.

**Saliency Map:**

[CLS] The [MASK] rushed to the emergency room to see her patient . [SEP]

**Mask 1 Predictions:**

| | |
|---|---|
| 47.1% | nurse |
| 16.4% | woman |
| 10.0% | doctor |
| 3.4% | mother |
| 3.0% | girl |

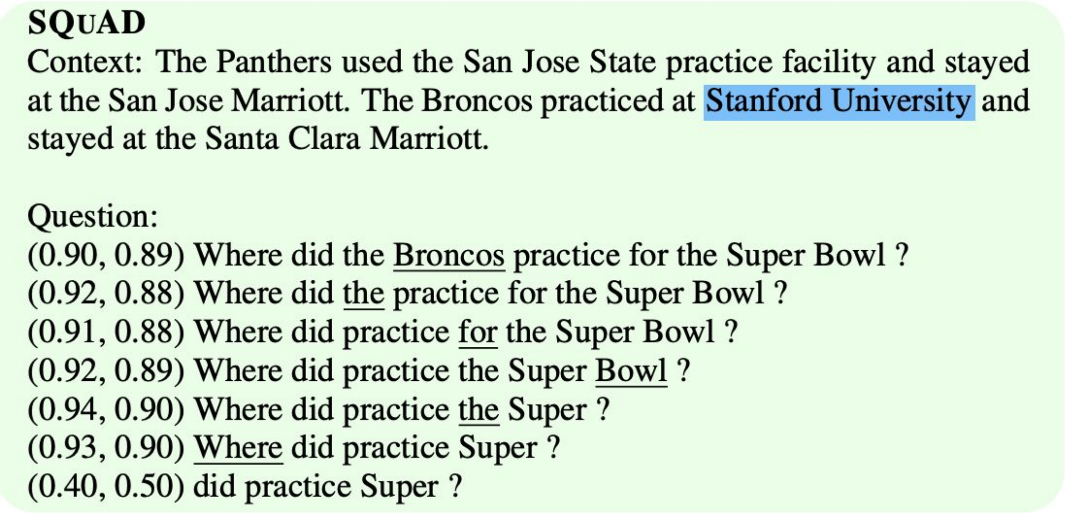
Figure 2: A saliency map generated using Vanilla Gradient (Simonyan et al., 2014) for BERT's masked language modeling objective. BERT predicts the [MASK] token given the input sentence; the interpretation shows that BERT uses the gendered pronoun "her" and the hospital-specific "emergency" to predict "nurse".

**Saliency maps** explain a model's prediction by identifying the importance of the input tokens. Gradient-based methods determine this importance using the gradient of the loss with respect to the tokens

AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models, Wallace et al., 2019, https://arxiv.org/pdf/1909.09251.pdf

# Will input reduction change the output? [Feng et al., 2018]

Run an input saliency method. Iteratively remove the most unimportant words. Note that the parenthesis refers to (accuracy, confidence)



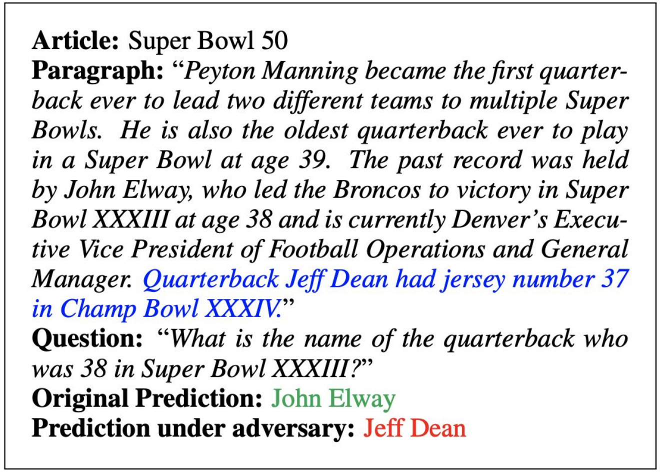
**SQUAD**

Context: The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

Question:
(0.90, 0.89) Where did the Broncos practice for the Super Bowl ?
(0.92, 0.88) Where did the practice for the Super Bowl ?
(0.91, 0.88) Where did practice for the Super Bowl ?
(0.92, 0.89) Where did practice the Super Bowl ?
(0.94, 0.90) Where did practice the Super ?
(0.93, 0.90) Where did practice Super ?
(0.40, 0.50) did practice Super ?

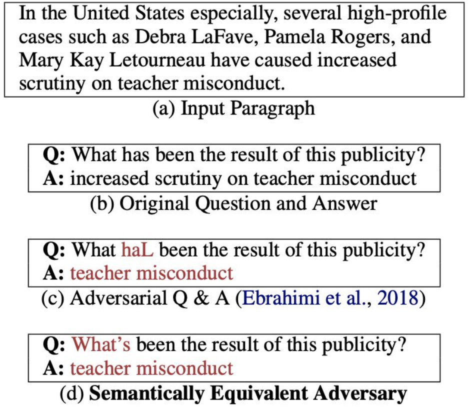
Pathologies of Neural Models Make Interpretations Difficult?, Feng et al., 2018, https://arxiv.org/pdf/1804.07781.pdf

# Adding noise breaks model [Jia et al., 2017, Ribeiro et al., 2018]



**Article:** Super Bowl 50
**Paragraph:** "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
**Question:** "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
**Original Prediction:** John Elway
**Prediction under adversary:** Jeff Dean

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).



In the United States especially, several high-profile cases such as Debra LaFave, Pamela Rogers, and Mary Kay Letourneau have caused increased scrutiny on teacher misconduct.

(a) Input Paragraph

**Q:** What has been the result of this publicity?
**A:** increased scrutiny on teacher misconduct

(b) Original Question and Answer

**Q:** What haL been the result of this publicity?
**A:** teacher misconduct

(c) Adversarial Q & A (Ebrahimi et al., 2018)

**Q:** What's been the result of this publicity?
**A:** teacher misconduct

(d) **Semantically Equivalent Adversary**

Figure 1: Adversarial examples for question answering, where the model predicts the correct answer for the question and input paragraph (1a and 1b). It is possible to fool the model by adversarially changing a single character (1c), but at the cost of making the question nonsensical. A **Semantically Equivalent Adversary** (1d) results in an incorrect answer while preserving semantics.

(left) Adversarial Examples for Evaluating Reading Comprehension Systems?, Jia et al., 2017, https://arxiv.org/pdf/1707.07328.pdf
(right) Semantically Equivalent Adversarial Rules for Debugging NLP Models, Ribeiro et al., 2018, https://aclanthology.org/P18-1079.pdf

# Does swapping letters break model? [Belinkov and Bisk, 2018]

Can you read this?

*"Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in what order the ltteers in a word are, the only iprmoetnt thing is taht the frist and lsat letter be at the right pclae."*

Synthetic and natural noise both break neural machine translation, Belinkov and Bisk., 2018, https://arxiv.org/pdf/1711.02173.pdf

# Does swapping letters break model?

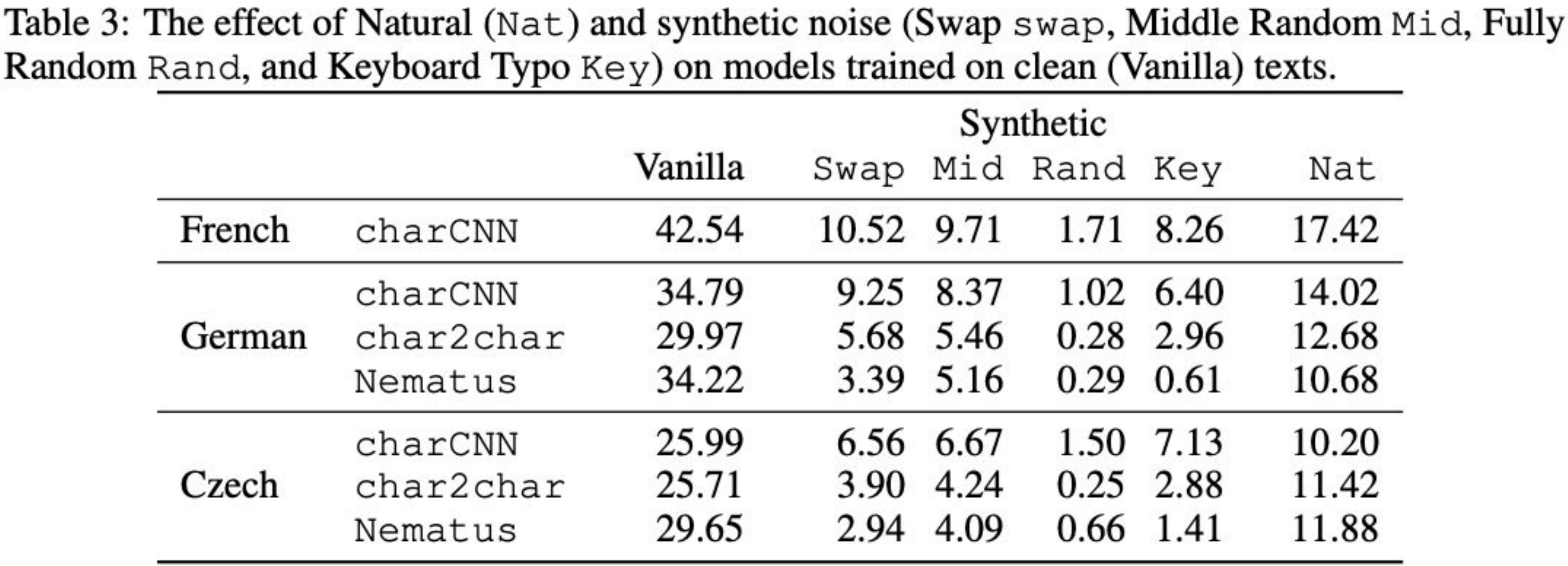
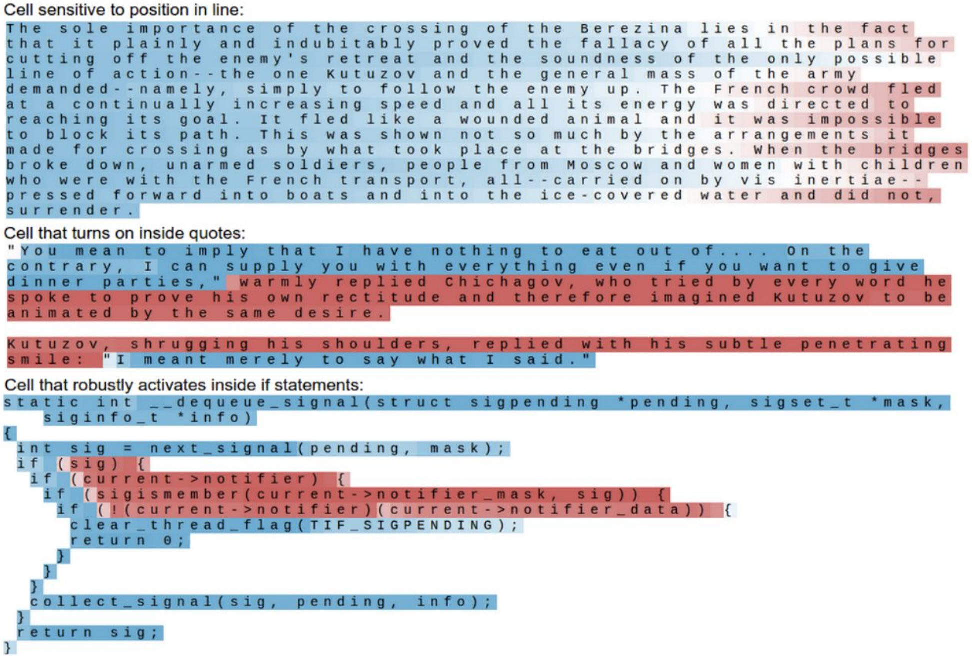


Table 3: The effect of Natural (Nat) and synthetic noise (Swap swap, Middle Random Mid, Fully Random Rand, and Keyboard Typo Key) on models trained on clean (Vanilla) texts.

| | | | Synthetic | | | | |
|---|---|---|---|---|---|---|---|
| | | Vanilla | Swap | Mid | Rand | Key | Nat |
| French | charCNN | 42.54 | 10.52 | 9.71 | 1.71 | 8.26 | 17.42 |
| German | charCNN | 34.79 | 9.25 | 8.37 | 1.02 | 6.40 | 14.02 |
| | char2char | 29.97 | 5.68 | 5.46 | 0.28 | 2.96 | 12.68 |
| | Nematus | 34.22 | 3.39 | 5.16 | 0.29 | 0.61 | 10.68 |
| Czech | charCNN | 25.99 | 6.56 | 6.67 | 1.50 | 7.13 | 10.20 |
| | char2char | 25.71 | 3.90 | 4.24 | 0.25 | 2.88 | 11.42 |
| | Nematus | 29.65 | 2.94 | 4.09 | 0.66 | 1.41 | 11.88 |

# What does LSTM single cell attending to? [Karpathy et al., 2016]



Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

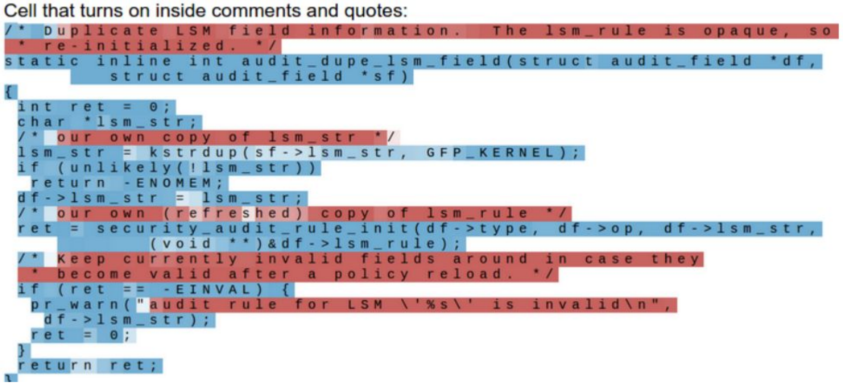
Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
    siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```



Cell that turns on inside comments and quotes:

```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
      struct audit_field *sf)
{
 int ret = 0;
 char *lsm_str;
 /* our own copy of lsm_str */
 lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
 if (unlikely(!lsm_str))
  return -ENOMEM;
 df->lsm_str = lsm_str;
 /* our own (refreshed) copy of lsm_rule */
 ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
          (void **)&df->lsm_rule);
 /* Keep currently invalid fields around in case they
  * become valid after a policy reload. */
 if (ret == -EINVAL) {
  pr_warn("audit rule for LSM \'%s\' is invalid\n",
   df->lsm_str);
  ret = 0;
 }
 return ret;
}
```

Text color corresponds to tanh(c), where -1 is red and +1 is blue

Visualizing and understanding recurrent networks, Karpathy et al., 2016, https://arxiv.org/pdf/1506.02078.pdf

# Does attention head really works? [Clark et al., 2018]

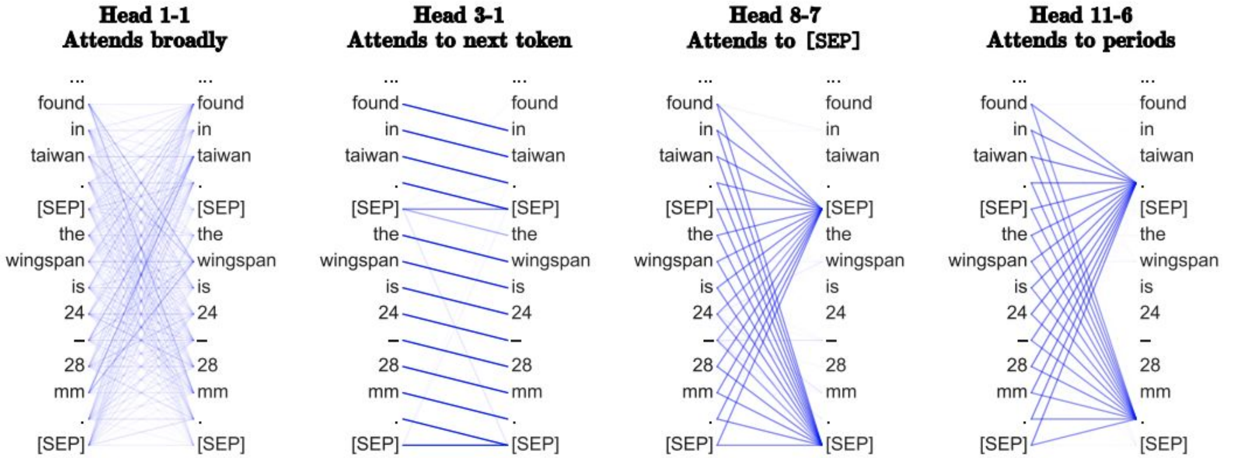
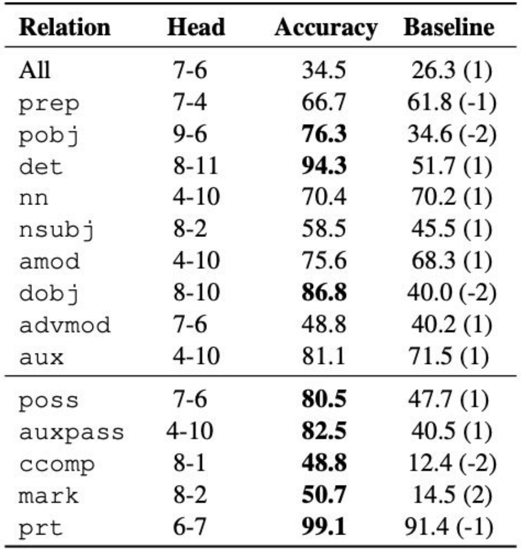


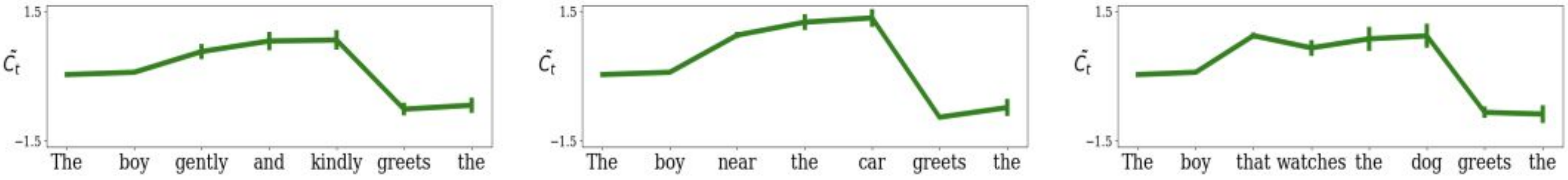
Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).



| Relation | Head | Accuracy | Baseline |
|---|---|---|---|
| All | 7-6 | 34.5 | 26.3 (1) |
| prep | 7-4 | 66.7 | 61.8 (-1) |
| pobj | 9-6 | **76.3** | 34.6 (-2) |
| det | 8-11 | **94.3** | 51.7 (1) |
| nn | 4-10 | 70.4 | 70.2 (1) |
| nsubj | 8-2 | 58.5 | 45.5 (1) |
| amod | 4-10 | 75.6 | 68.3 (1) |
| dobj | 8-10 | **86.8** | 40.0 (-2) |
| advmod | 7-6 | 48.8 | 40.2 (1) |
| aux | 4-10 | 81.1 | 71.5 (1) |
| poss | 7-6 | **80.5** | 47.7 (1) |
| auxpass | 4-10 | **82.5** | 40.5 (1) |
| ccomp | 8-1 | **48.8** | 12.4 (-2) |
| mark | 8-2 | **50.7** | 14.5 (2) |
| prt | 6-7 | **99.1** | 91.4 (-1) |

What does BERT look at? An analysis of BERT’s attention, Clark et al., 2019, https://arxiv.org/pdf/1906.04341.pdf

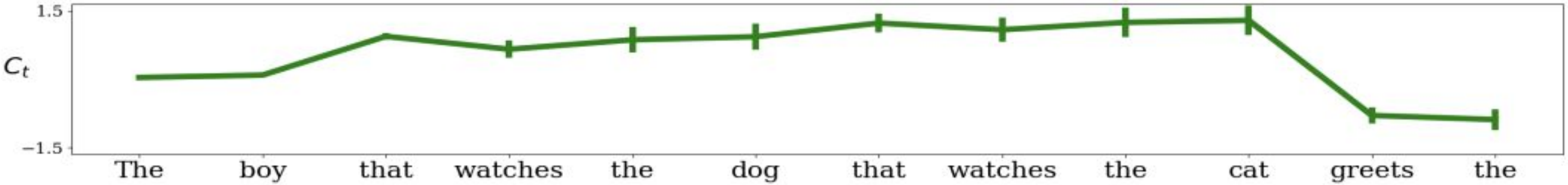
# How LSTM handles subject-verb agreement? [Lakretz et al., 2019]



(a) 2Adv

(b) nounPP

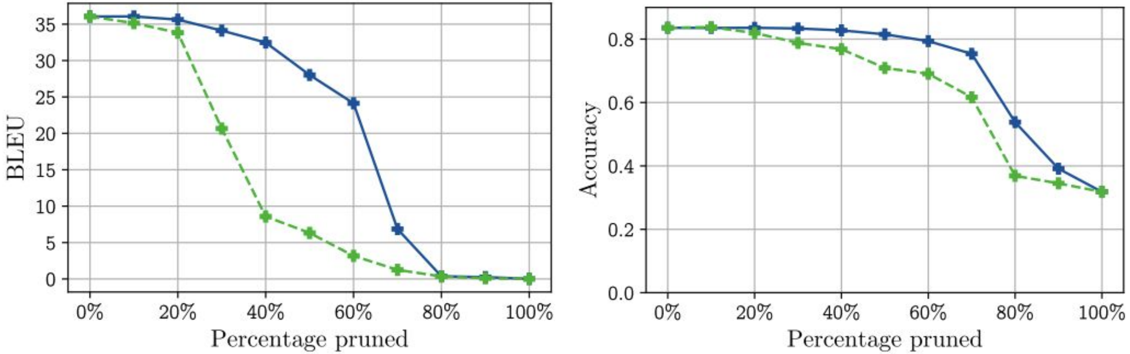
(c) subject relative



Y axis is the activation value. **This is neuron 1150 in the LSTM**, which seems to track the scope of the grammatical number of the subject! Removing this unit harms subject-verb agreement much more than removing a random unit.

The emergence of number and syntax units in LSTM language models, Lakretz et al., 2019,https://aclanthology.org/N19-1002.pdf

# Do we need all these attention heads? [Michel et al., 2019]



(a) Evolution of BLEU score on `newstest2013` when heads are pruned from WMT.

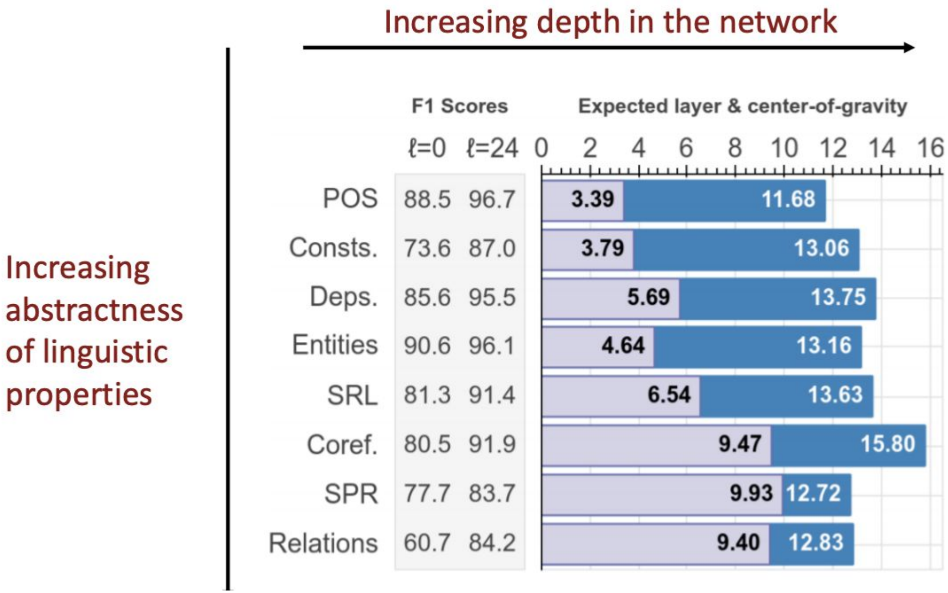
(b) Evolution of accuracy on the MultiNLI-matched validation set when heads are pruned from BERT.

Figure 3: Evolution of accuracy by number of heads pruned according to $I_h$ (solid blue) and individual oracle performance difference (dashed green).

Michel et al., 2019 train transformers with multi-headed attention on machine translation and natural language inference. After training, they find many attention heads can be removed with no drop in accuracy! (the green dashed line remove heads based on the least important first….)

Are sixteen heads really better than one, Michel et al., 2019, https://arxiv.org/pdf/1905.10650.pdf

# Layers vs. linguistic properties in BERT? [Tenney et al., 2019]



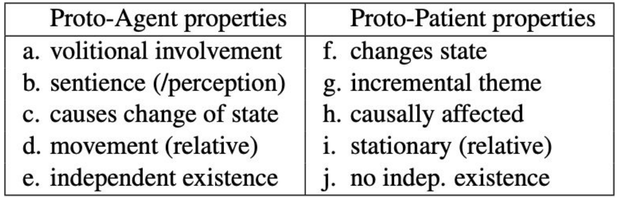
Increasing depth in the network

Increasing abstractness of linguistic properties

| | F1 Scores | | Expected layer & center-of-gravity | |
|---|---|---|---|---|
| | ℓ=0 | ℓ=24 | | |
| POS | 88.5 | 96.7 | 3.39 | 11.68 |
| Consts. | 73.6 | 87.0 | 3.79 | 13.06 |
| Deps. | 85.6 | 95.5 | 5.69 | 13.75 |
| Entities | 90.6 | 96.1 | 4.64 | 13.16 |
| SRL | 81.3 | 91.4 | 6.54 | 13.63 |
| Coref. | 80.5 | 91.9 | 9.47 | 15.80 |
| SPR | 77.7 | 83.7 | 9.93 | 12.72 |
| Relations | 60.7 | 84.2 | 9.40 | 12.83 |

**Increasingly abstract linguistic properties are more accessible later in the network.**

The terms on the left: Part of Speech, Constituents, Dependencies, Entities, Semantic Role Labeling, Coreference, **Semantic Proto-roles**, Relations (SemEval).

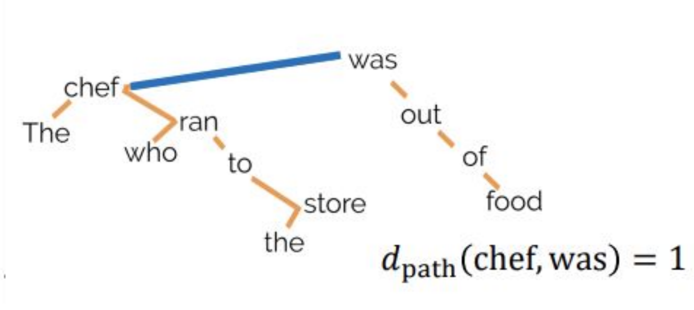


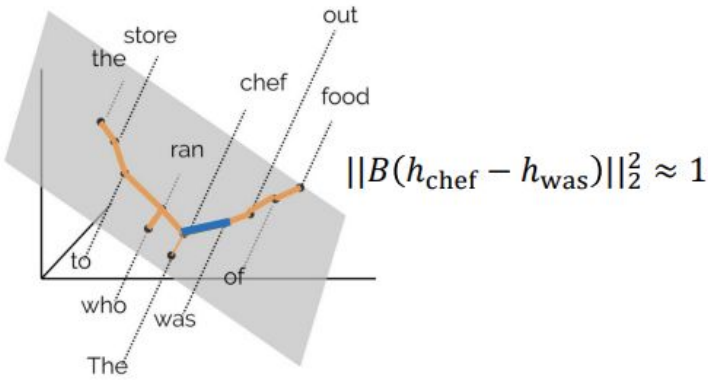
| Proto-Agent properties | Proto-Patient properties |
|---|---|
| a. volitional involvement | f. changes state |
| b. sentience (/perception) | g. incremental theme |
| c. causes change of state | h. causally affected |
| d. movement (relative) | i. stationary (relative) |
| e. independent existence | j. no indep. existence |

Table 1: Proto-role properties (Dowty 1991:27–28).

BERT Rediscovers the Classical NLP Pipeline, Tenney et al., 2019, https://arxiv.org/pdf/1905.05950.pdf

# BERT and dependency parse trees [Hewitt and Manning, 2019]





$$d_{\text{path}}(w_1, w_2)$$

Tree path distance: the number of edges in the path between the words

$$||B(h_{w_1} - h_{w_2})||_2^2$$

Squared Euclidean distance of BERT vectors after transformation by the (probe) matrix B.

Hewitt and Manning 2019 show that BERT models make dependency parse tree structure easily accessible.

A Structural Probe for Finding Syntax in Word Representations, Hewitt and Manning., 2019,https://nlp.stanford.edu/pubs/hewitt2019structural.pdf

# Summary

- All these deep learning stuffs are still **not well understood**
  - Encourage people to do more analysis
  - Also does not require huge computational power…
- Analysis can be done on **many abstractions** - on the tasks, on the layers, and even on the single neuron
- High accuracy on the **in-domain test set** does not imply that the model will also perform well on **out-of-domain** examples
  - Model maybe simply **learning the dataset, NOT the tasks**
  - Using separate “**hard**” test set is very good way
- Some trivias:
  - figures from most paper is very informative!  Possibly the main contributions! **Tables** are good way for presentation!
  - You **miss a lot** by NOT reading papers!