Explainability for NLP

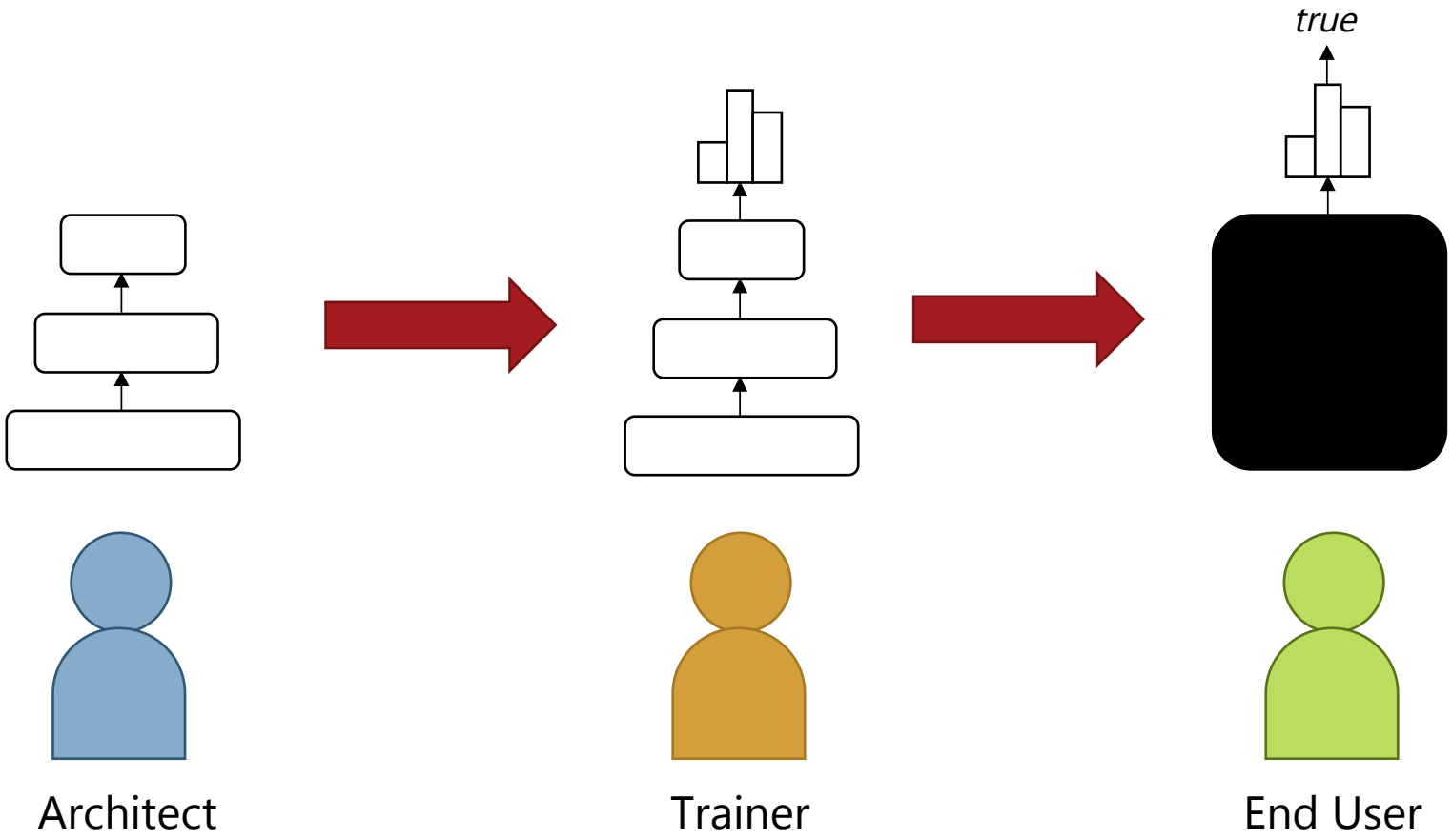
Isabelle Augenstein*
augenstein@di.ku.dk
@IAugenstein
http://isabelleaugenstein.github.io/

ALPS Winter School
22 January 2021

UNIVERSITY OF COPENHAGEN

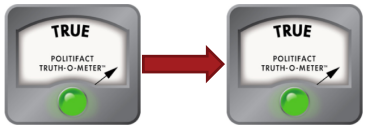
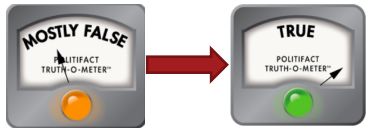
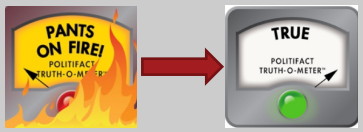
*partial slide credit: Pepa Atanasova, Nils Rethmeier

Explainability – what is it and why do we need it?

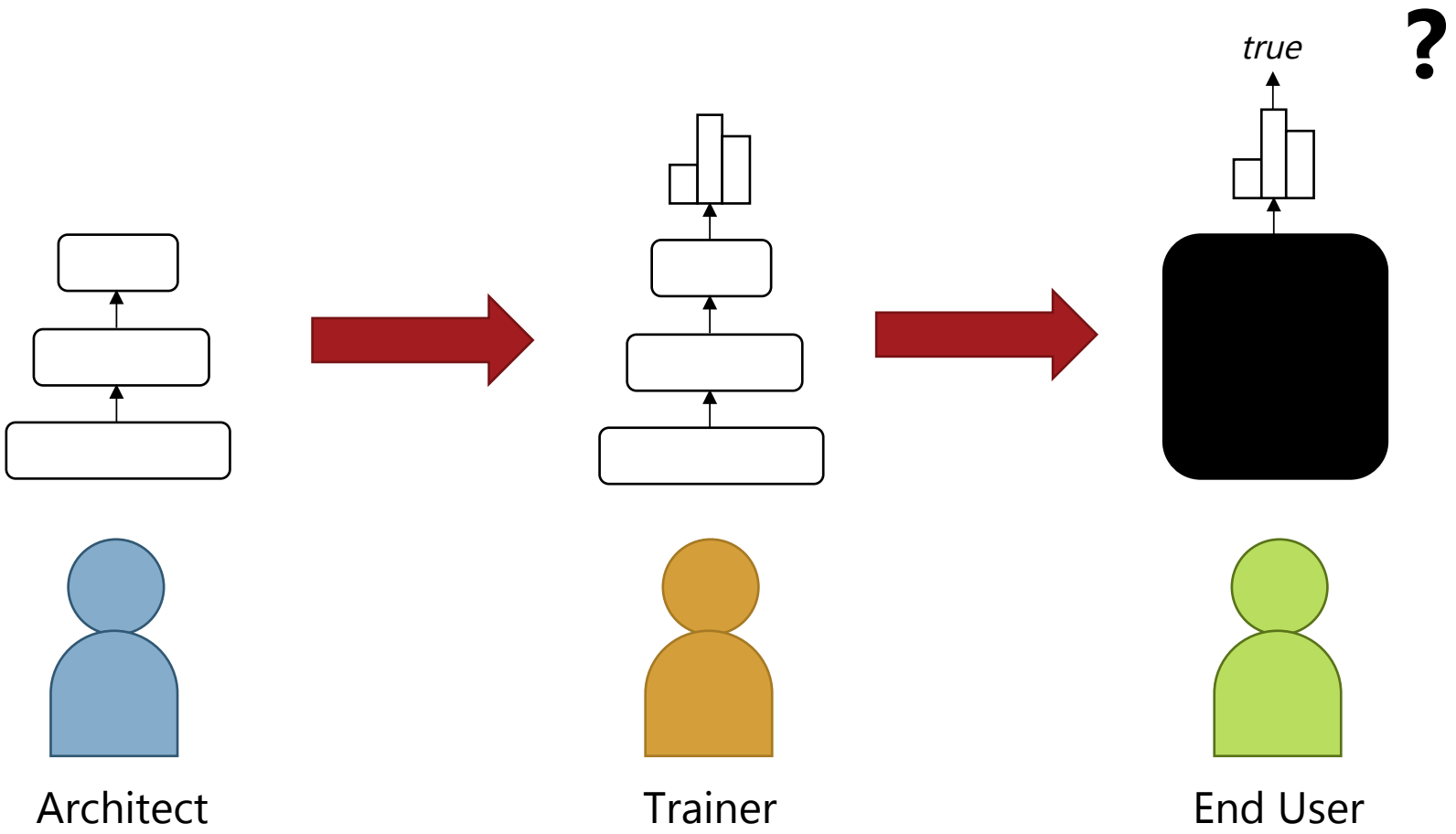


Terminology borrowed from Strobelt et al. (2018), "LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE transactions on visualization and computer graphics."

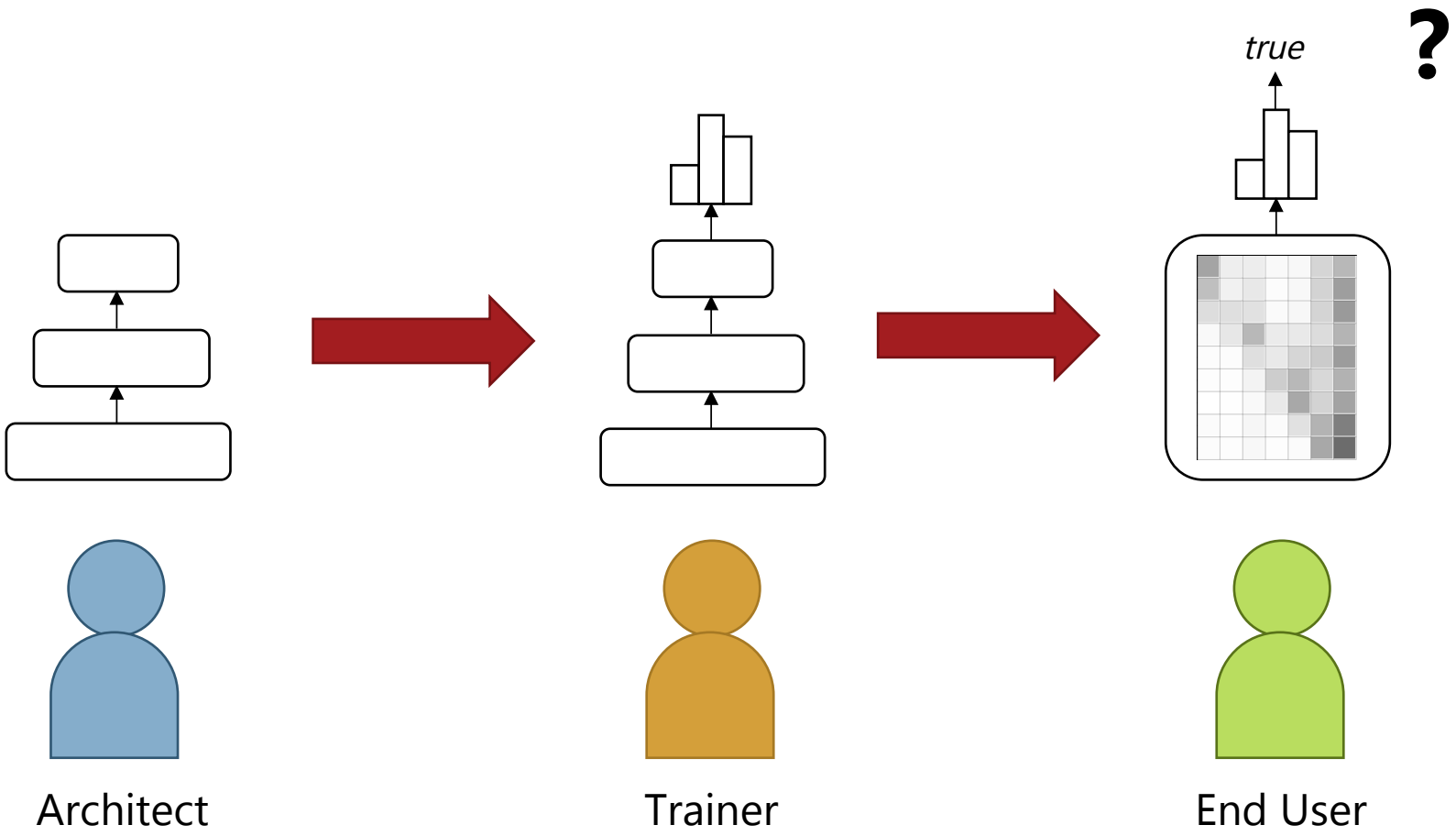
Explainability – what is it and why do we need it?

	Right prediction	Wrong prediction
Right reasons	**Claim:** "In the COVID-19 crisis, 'only 20% of African Americans had jobs where they could work from home.'" **Evidence:** "20% of black workers said they could work from home in their primary job, compared to 30% of white workers." 	**Claim:** "Children don't seem to be getting this virus." **Evidence**: "There have been no reported incidents of infection in children." 
Wrong reasons	**Claim:** "Taylor Swift had a fatal car accident." **Reason**: overfitting to spurious patterns (celebrity death hoaxes are common) 	**Claim**: "Michael Jackson is still alive, appears in daughter's selfie." **Reason**: overfitting to spurious patterns (celebrity death hoaxes are common) 

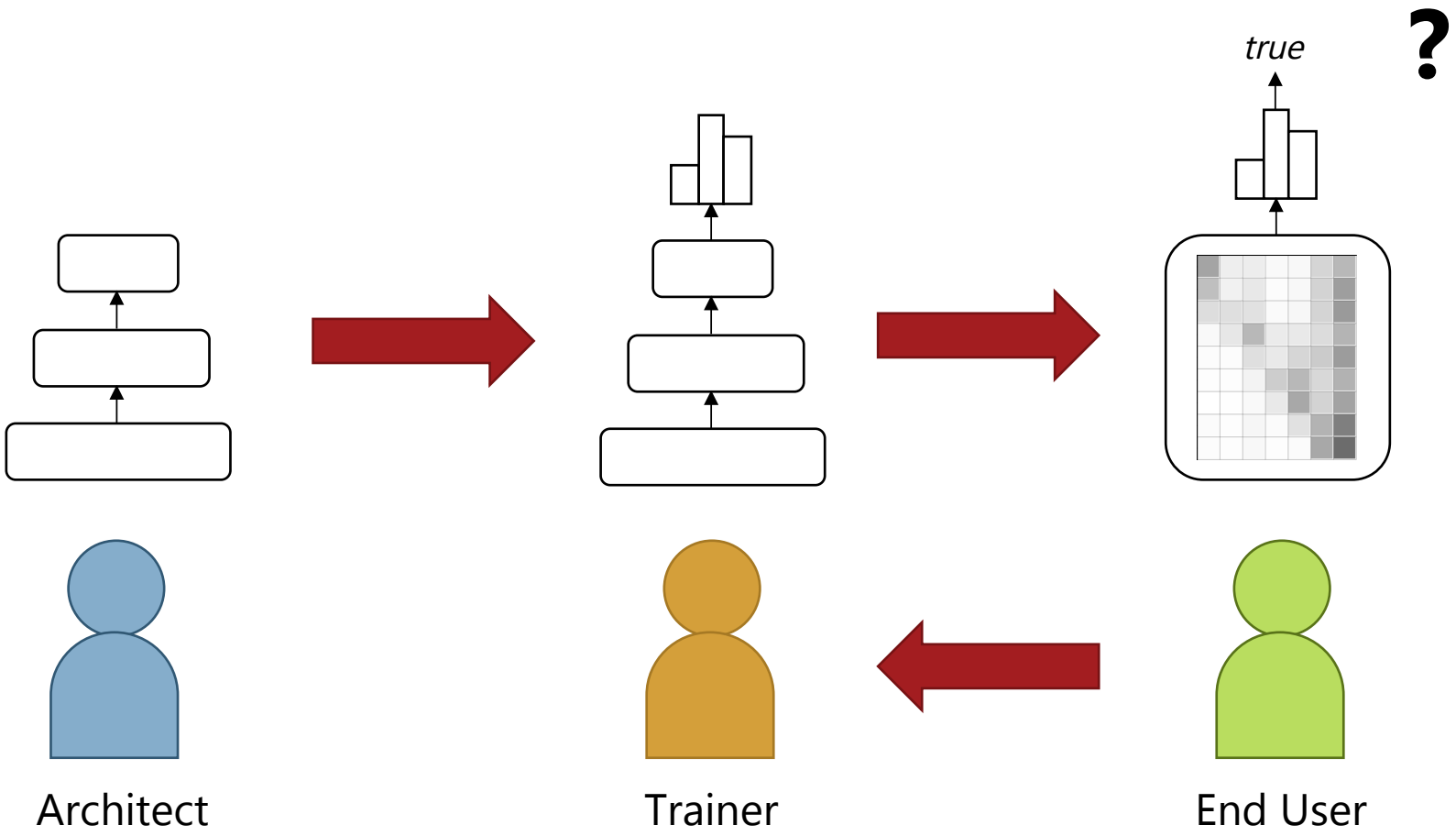
Explainability – what is it and why do we need it?



Explainability – what is it and why do we need it?



Explainability – what is it and why do we need it?



Types of Explainability

- **Model Understanding**
  - What features and parameters has a model learned?
  - How do these features and parameters relate to model outputs generally?
- **Decision Understanding**
  - How does the model arrive at predictions for specific instances?
  - Which features and parameters influence a specific prediction?

Types of Explainability

- **Black Box Explainability Methods**
  - No access to the model parameters, only predictions
  - Observing output changes via different inputs
- **White Box Explainability Methods**
  - Access to the model features and parameters
  - Correlating outputs with specific features and parameters

Types of Explainability

- **Joint Explainability Methods**
  - Explanation produced jointly with target task
- **Post-Hoc Explainability Methods**
  - Explanation produced for a trained model

Types of Explainability

- **Model Understanding**
  - What features and parameters has a model learned?
  - Methods:
    - Feature visualisation methods (white box)
    - Adversarial examples (black or white box)

- **Decision Understanding**
  - How does the model arrive at predictions for specific instances?
  - Methods
    - Probing tasks (black or white box)
    - Correlating inputs with gradients/attention weights/etc. (white box)
    - Generating text explaining predictions (white box)

Overview of Today's Talk

- **Introduction**
  - Explainability – what is it and why do we need it?

- **Part 1: Decision understanding**
  - *Instance-level* explainability for text classification and fact checking
  - Language generation based explanations
  - Evaluating instance-level explanations

- **Part 2: Model understanding**
  - *Model-wide* explainability for text classification and fact checking
  - Finding model-wide explanations
  - Visualising model-wide explanations

Part 1:

Decision Understanding

Generating Fact Checking Explanations

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein

ACL 2020

Fact Checking

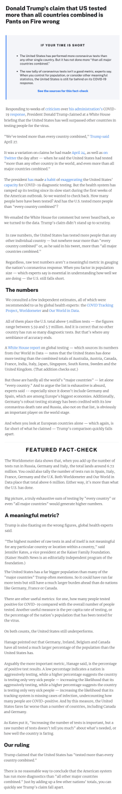
POLITIFACT
The Poynter Institute

Donald Trump

stated on April 27, 2020 in comments made during a White House briefing:

> **"We've tested more than every country combined."**

HEALTH CHECK | CORONAVIRUS | DONALD TRUMP



Donald Trump's claim that US tested more than all countries combined is Pants on Fire wrong

IF YOUR TIME IS SHORT

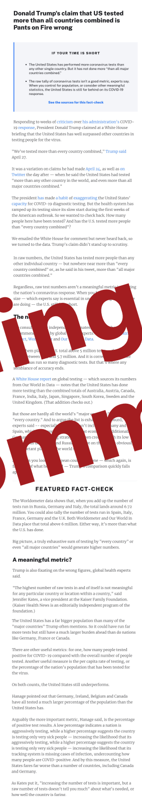
- The United States has performed more coronavirus tests than any other single country. But it has not done more "than all major countries combined."
- The raw tally of coronavirus tests isn't a good metric, experts say. When you control for population, or consider other meaningful statistics, the United States is still far behind on its COVID-19 response.

See the sources for this fact-check

https://www.politifact.com/factchecks/2020/apr/30/donald-trump/trumps-boasts-testing-spring-cherry-picking-data/

Terminology

POLITIFACT
The Poynter Institute



Donald Trump's claim that US tested more than all countries combined is Pants on Fire wrong

IF YOUR TIME IS SHORT

- The United States has performed more coronavirus tests than any other single country. But it has not done more "than all major countries combined."
- The raw tally of coronavirus tests isn't a good metric, experts say. When you control for population, or consider other meaningful statistics, the United States is still far behind on its COVID-19 response.

See the sources for this fact-check

Responding to weeks of criticism over his administration's COVID-19 response, President Donald Trump claimed at a White House briefing that the United States has well surpassed other countries in testing people for the virus.

"We've tested more than every country combined," Trump said April 27.

It was a variation on claims he had made April 24, as well as on Twitter the day after — when he said the United States had tested "more than any other country in the world, and even more than all major countries combined."

The president has made a habit of exaggerating the United States' capacity for COVID-19 diagnostic testing. But the health system has ramped up its testing since its slow start during the first weeks of the American outbreak. So we wanted to check back. How many people here have been tested? And has the U.S. tested more people than "every country combined"?

We emailed the White House for comment but never heard back, so we turned to the data. Trump's claim didn't stand up to scrutiny.

In raw numbers, the United States has tested more people than any other individual country — but nowhere near more than "every country combined" or, as he said in his tweet, more than "all major countries combined."

A White House report on global testing — which sources its numbers from Our World in Data — notes that the United States has done more testing than the combined totals of Australia, Austria, Canada, France, India, Italy, Japan, Singapore, South Korea, Sweden and the United Kingdom. (That addition checks out.)

FEATURED FACT-CHECK

The Worldometer data shows that, when you add up the number of tests run in Russia, Germany and Italy, the total lands around 6.72 million. You could also tally the number of tests run in Spain, Italy, France, Germany and the U.K. Both Worldometer and Our World in Data place that total above 6 million. Either way, it's more than what the U.S. has done.

Big picture, a truly exhaustive sum of testing by "every country" or even "all major countries" would generate higher numbers.

A meaningful metric?

Trump is also fixating on the wrong figures, global health experts said.

"The highest number of raw tests in and of itself is not meaningful for any particular country or location within a country," said Jennifer Kates, a vice president at the Kaiser Family Foundation. (Kaiser Health News is an editorially independent program of the foundation.)

The United States has a far bigger population than many of the "major countries" Trump often mentions. So it could have run far more tests but still have a much larger burden ahead than do nations like Germany, France or Canada.

There are other useful metrics: for one, how many people tested positive for COVID-19 compared with the overall number of people tested. Another useful measure is the per capita rate of testing, or the percentage of the nation's population that has been tested for the virus.

On both counts, the United States still underperforms.

Hanage pointed out that Germany, Ireland, Belgium and Canada have all tested a much larger percentage of the population than the United States has.

Arguably the more important metric, Hanage said, is the percentage of positive test results. A low percentage indicates a nation is aggressively testing, while a higher percentage suggests the country is testing only very sick people — increasing the likelihood that its aggressively testing, while a higher percentage suggests the country is testing only very sick people — increasing the likelihood that its tracking system is missing cases of infection, undercounting how many people are COVID-positive. And by this measure, the United States fares far worse than a number of countries, including Canada and Germany.

As Kates put it, "increasing the number of tests is important, but a raw number of tests doesn't tell you much" about what's needed, or how well the country is faring.

Ruling

Comments

Our ruling

Justification

Trump claimed that the United States has "tested more than every country combined."

There is no reasonable way to conclude that the American system has run more diagnostics than "all other major countries combined." Just by adding up a few other nations' totals, you can quickly see Trump's claim fall apart.

https://www.politifact.com/factchecks/2020/apr/30/donald-trump/trumps-boasts-testing-spring-cherry-picking-data/

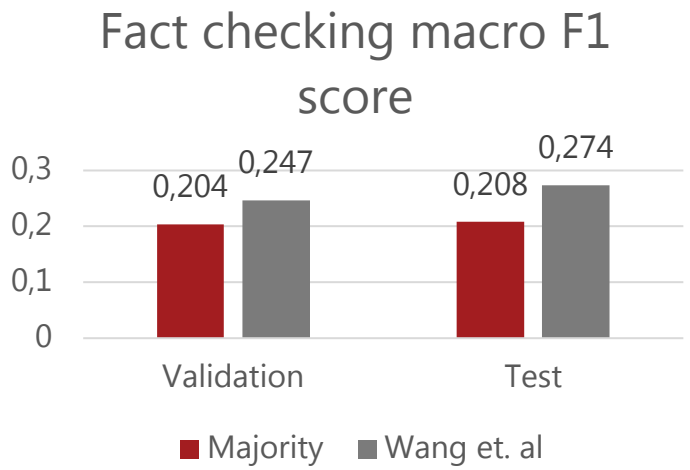
Automating Fact Checking

Statement: "The last quarter, it was just announced, our gross domestic product was below zero. Who ever heard of this? Its never below zero."
Speaker: Donald Trump
Context: presidential announcement speech
Label: Pants on Fire

Dataset Statistics	
Training set size	10,269
Validation set size	1,284
Testing set size	1,283
Avg. statement length (tokens)	17.9
Top-3 Speaker Affiliations	
Democrats	4,150
Republicans	5,687
None (e.g., FB posts)	2,185

Wang, William Yang. ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection." Proceedings ACL'2017.

Automating Fact Checking



Claim:
We've tested more than every country combined.

Metadata:
topic=healthcare, Coronavirus,
speaker=Donald Trump,
job=president,
context=White House briefing,
history={4, 10, 14, 21, 34, 14}

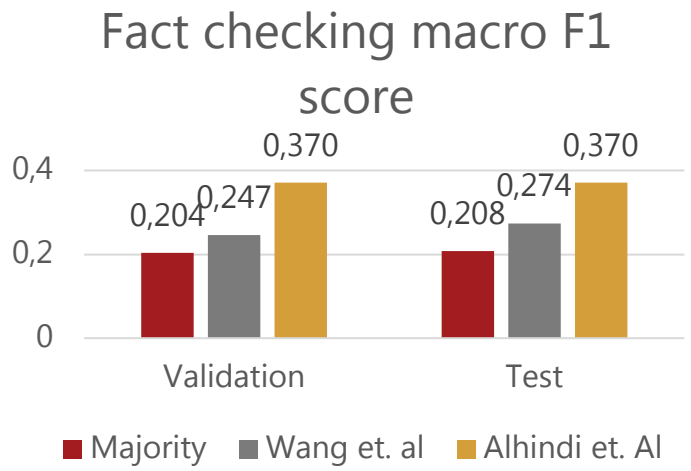
Hybrid CNN → Veracity Label

Wang, William Yang. ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection." Proceedings ACL'2017.

Automating Fact Checking

Claim:
We've tested more than every country combined.

Justification:
Trump claimed that the United States has "tested more than every country combined." There is no reasonable way to conclude that the American system has run more diagnostics than "all other major countries combined." Just by adding up a few other nations' totals, you can quickly see Trump's claim fall apart.
Plus, focusing on the 5 million figure distracts from the real issue — by any meaningful metric of diagnosing and tracking, the United States is still well behind countries like Germany and Canada.
The president's claim is not only inaccurate but also ridiculous. ~~We rate it Pants on Fire!~~



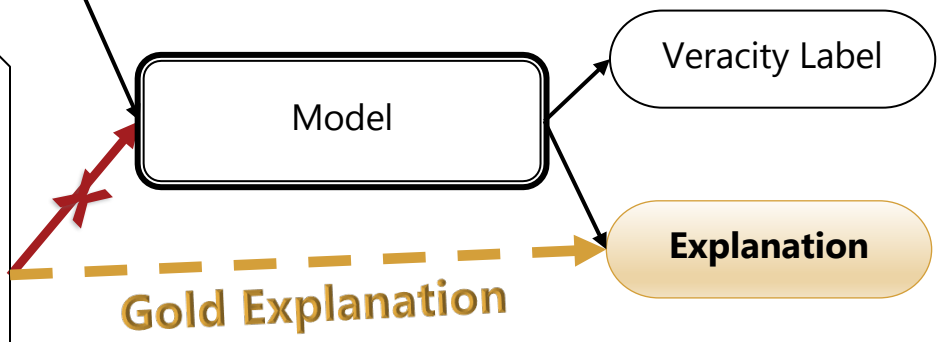
Logistic Regression → Veracity Label

Alhindi, Tariq, Savvas Petridis, and Smaranda Muresan. "Where is your Evidence: Improving Fact-checking by Justification Modeling." Proceedings of FEVER'2018.

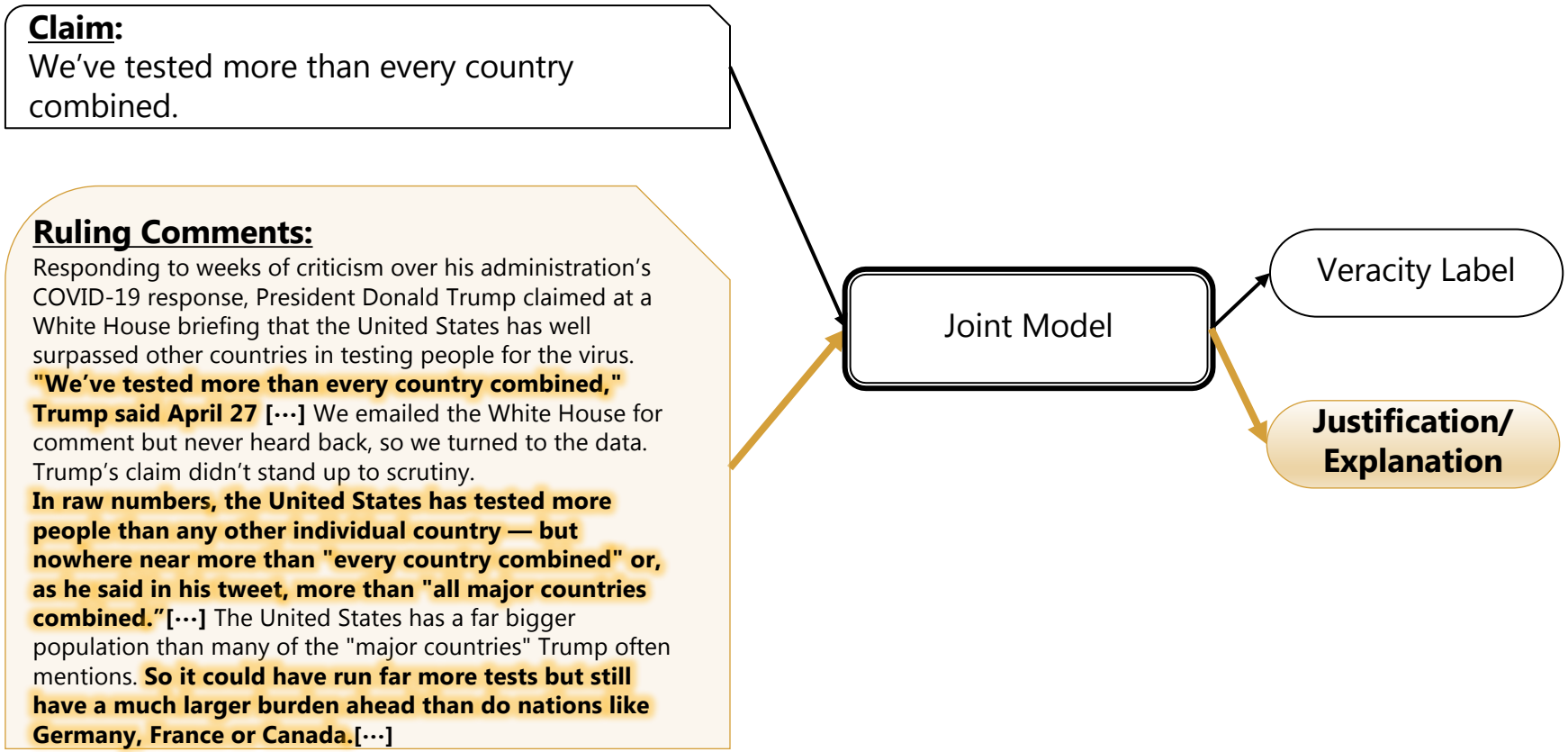
How about generating an explanation?

Claim:
We've tested more than every country combined.

Justification:
Trump claimed that the United States has "tested more than every country combined." There is no reasonable way to conclude that the American system has run more diagnostics than "all other major countries combined." Just by adding up a few other nations' totals, you can quickly see Trump's claim fall apart.
Plus, focusing on the 5 million figure distracts from the real issue — by any meaningful metric of diagnosing and tracking, the United States is still well behind countries like Germany and Canada.
The president's claim is not only inaccurate but also ridiculous. ~~We rate it Pants on Fire!~~



Generating Explanations from Ruling Comments



Related Studies on Generating Explanations

- *Camburu et. al; Rajani et. al* generate abstractive explanations
  - Short input text and explanations;
  - Large amount of annotated data.
- Real world fact checking datasets are of limited size and the input consists of long documents
- We take advantage of the LIAR-PLUS dataset:
  - Use the summary of the ruling comments as a gold explanation;
  - Formulate the problem as extractive summarization.

- *Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. "e-SNLI: Natural language inference with natural language explanations." In Advances in Neural Information Processing Systems,. 2018.*
- *Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4932-4942. 2019.*

Example of an Oracle's Gold Summary

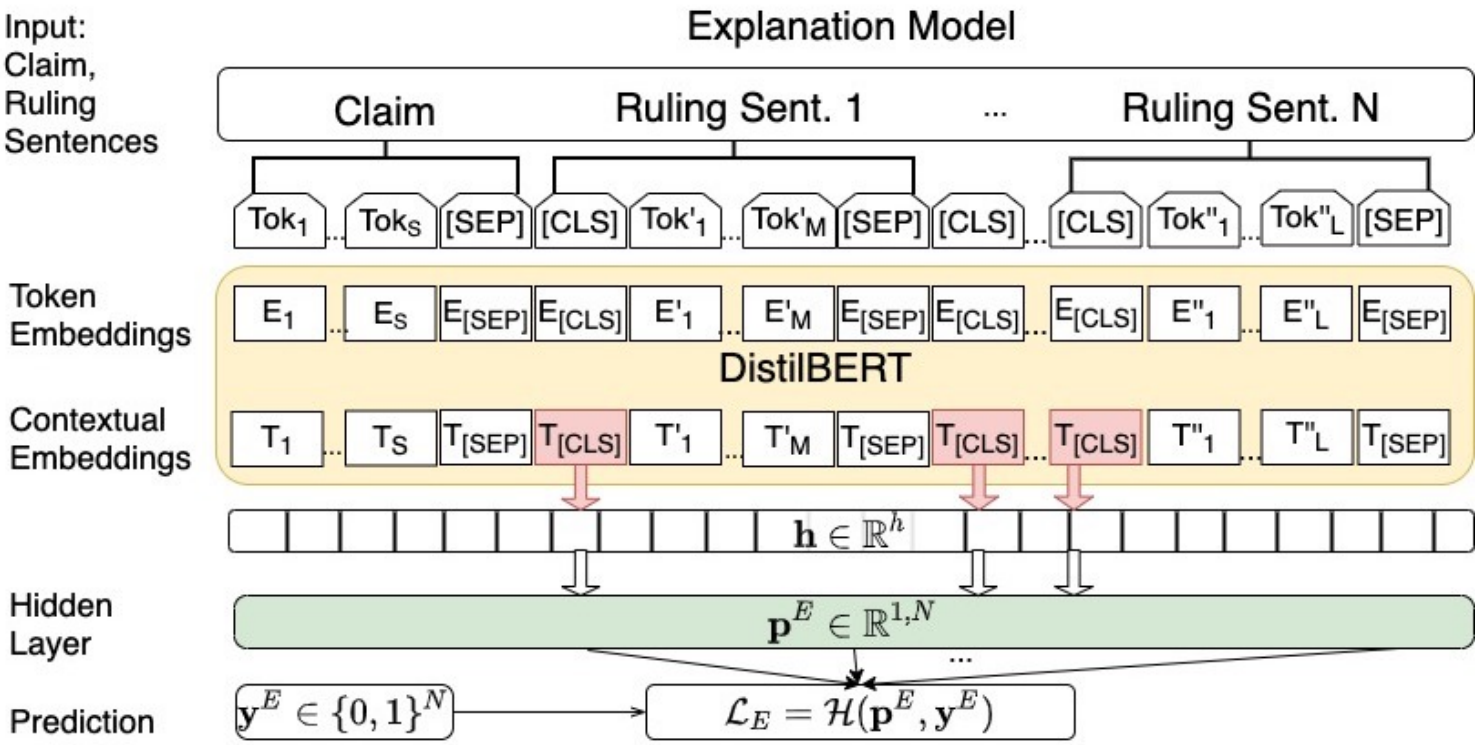
Claim: "The president promised that if he spent money on a stimulus program that unemployment would go to 5.7 percent or 6 percent. Those were his words."

Label: Mostly-False

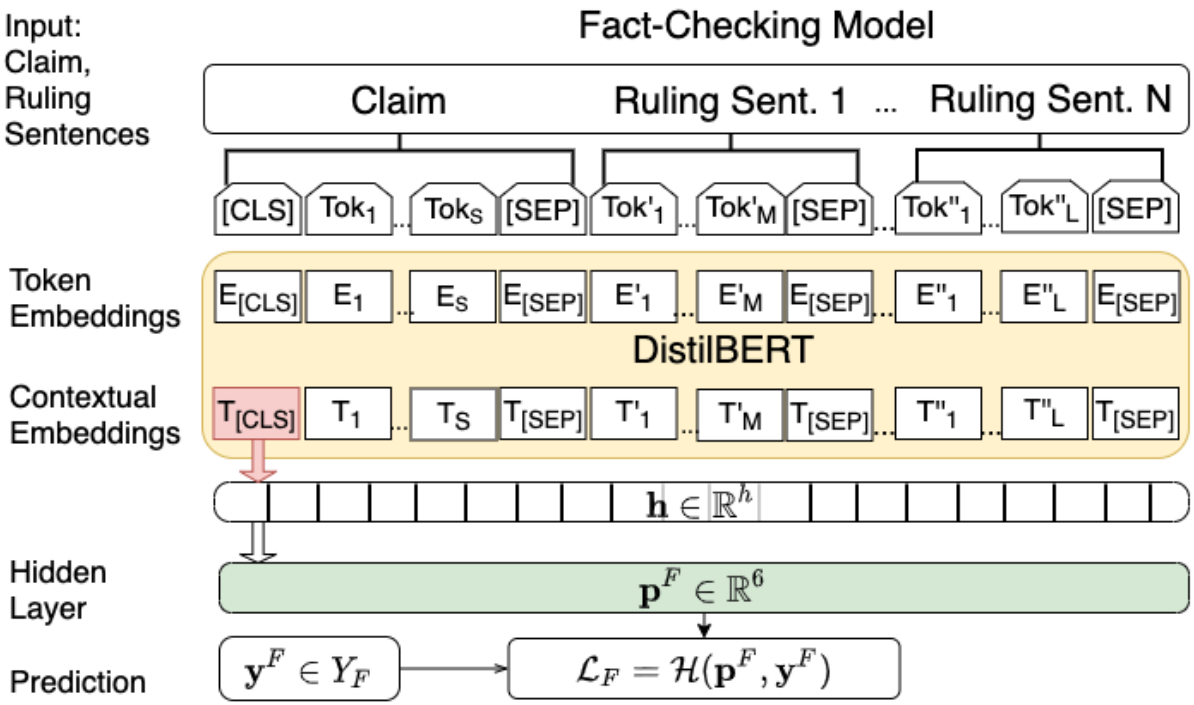
Just: **Bramnick said "the president promised that if he spent money on a stimulus program that unemployment would go to 5.7 percent or 6 percent. Those werehis words."** Two economic advisers **estimated in a 2009 report** that **with the stimulus plan**, the unemployment **rate would peak near 8 percent before dropping to** less than 6 **percent by now**. Those are critical details Bramnick's statement ignores. To comment on this ruling, go to NJ.com.

Oracle: **"The president promised that if he spent money on a stimulus program that unemployment would go to 5.7 percent or 6 percent. Those were his words,"Bramnick said** in a Sept. 7 interview on NJToday. But **with the stimulus plan, the report** projected the nation's jobless **rate would peak near 8 percent in 2009 before falling to** about 5.5 **percent by now**. So **the estimates in the report** were wrong.

Which sentences should be selected for the explanation?



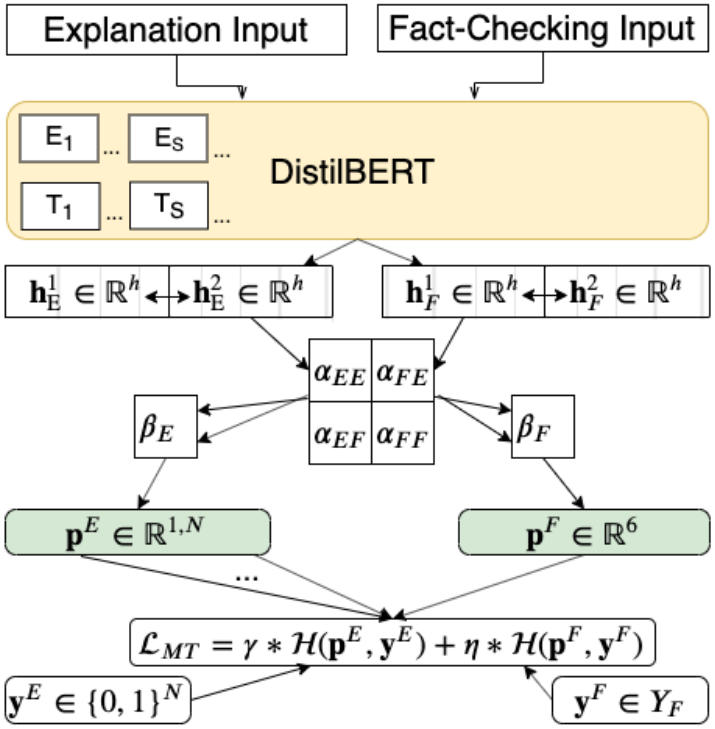
What is the veracity of the claim?



Joint Explanation and Veracity Prediction

Cross-stitch layer

Multi-task objective

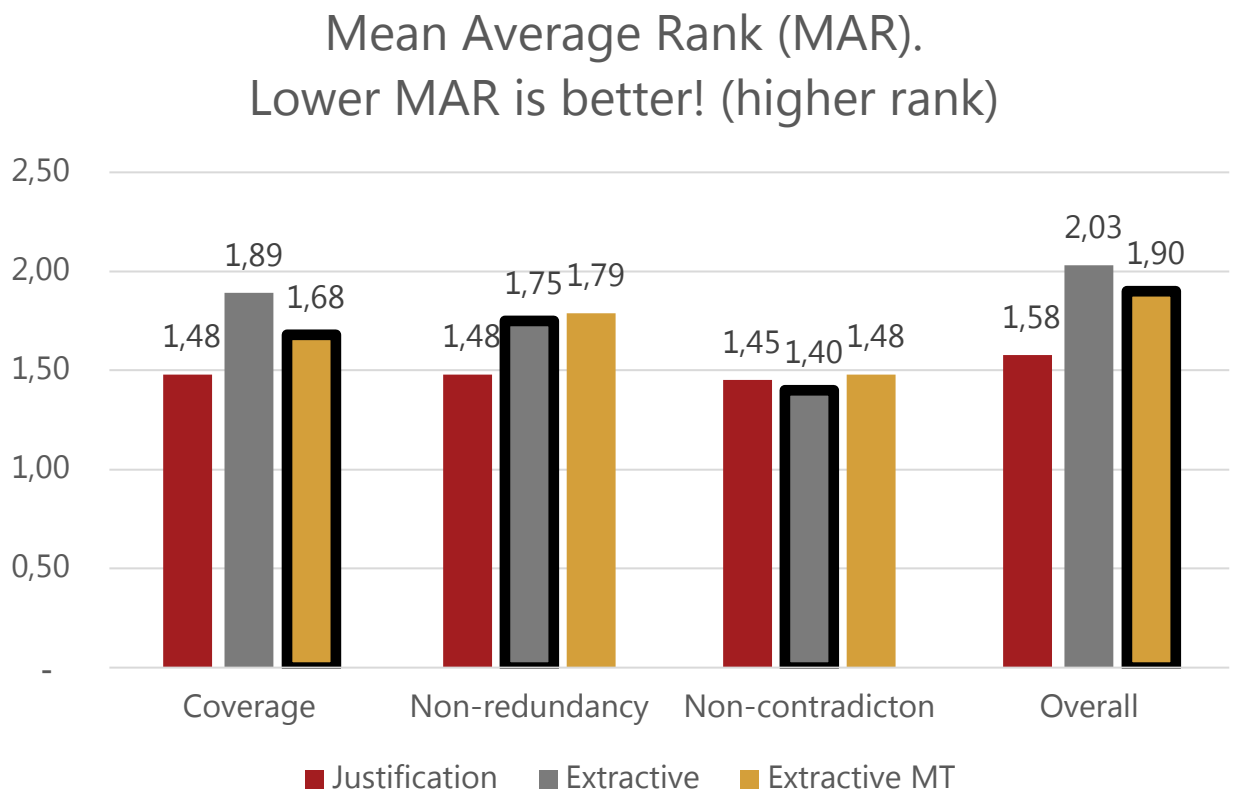


Ruder, Sebastian, et al. "Latent multi-task architecture learning." Proceedings of the AAAI'2019.

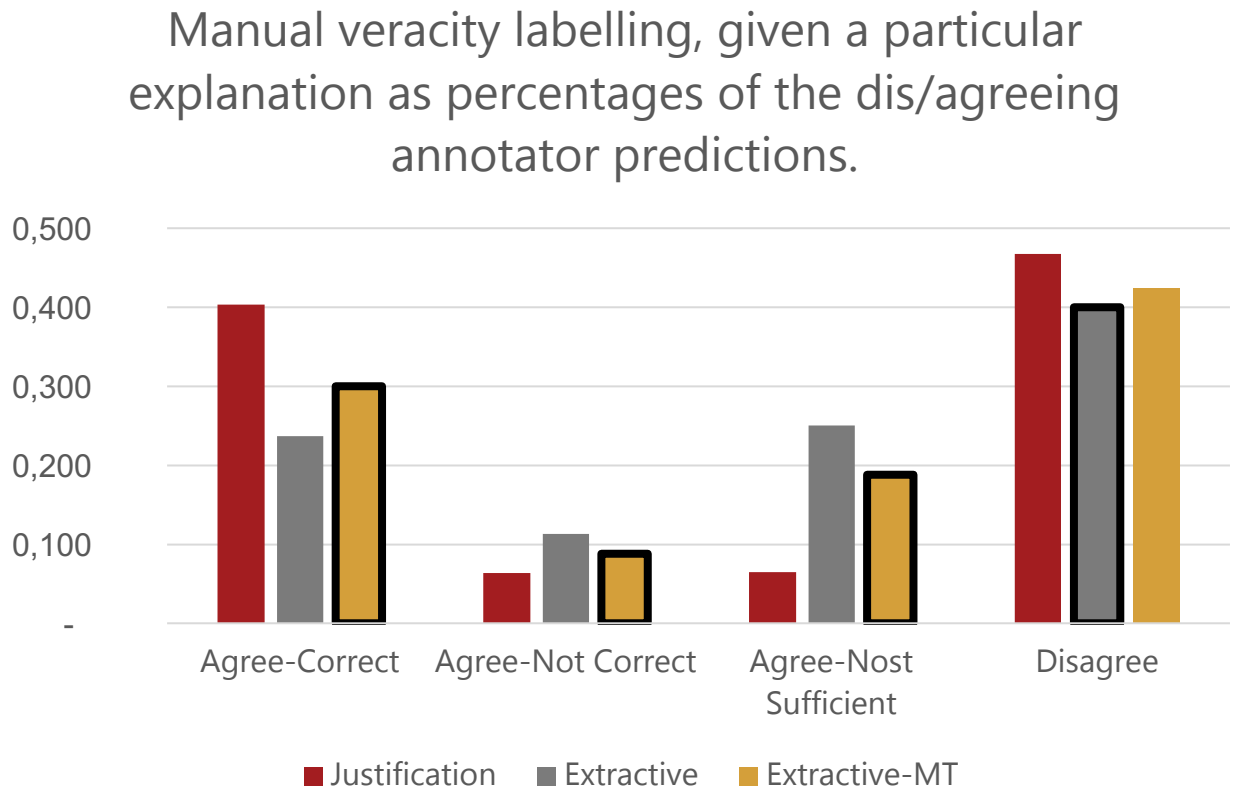
Manual Evaluation

- ***Explanation Quality***
  - ***Coverage.*** The explanation contains important, salient information and does not miss any important points that contribute to the fact check.
  - ***Non-redundancy.*** The summary does not contain any information that is redundant/repeated/not relevant to the claim and the fact check.
  - ***Non-contradiction.*** The summary does not contain any pieces of information contradictory to the claim and the fact check.
  - ***Overall***. Rank the explanations by their overall quality.
- ***Explanation Informativeness***. Provide a veracity label for a claim based on a veracity explanation coming from the justification, the Explain-MT, or the Explain-Extractive system.

Explanation Quality



Explanation Informativeness



Summary

- First study on generating veracity explanations
- Jointly training veracity prediction and explanation
  - improves the performance of the classification system
  - improves the coverage and overall performance of the generated explanations

Future Work

- Can we generate better or even abstractive explanations given limited resources?
- How to automatically evaluate the properties of the explanations?
- Can explanations be extracted from evidence pages only (lots of irrelevant and multi-modal results)?

A Diagnostic Study of Explainability Techniques for Text Classification

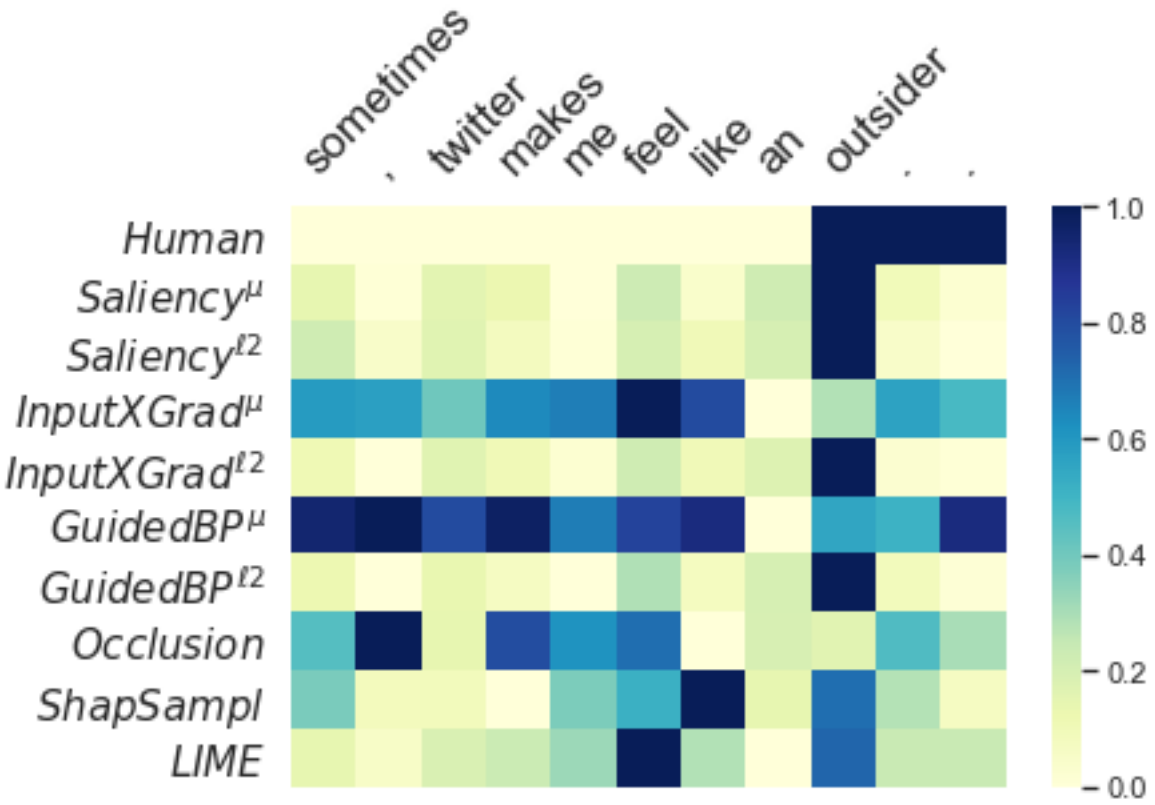
Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein

EMNLP 2020

Explainability Datasets for Decision Understanding

Dataset	Example	Size	Length
e-SNLI (Camburu et al., 2018)	*Premise:* An adult dressed in black **holds a stick.** *Hypothesis:* An adult is walking away, **empty-handed**. *Label*: contradiction	549 367 Train 9 842 Dev 9 824 Test	27.4 inst. 5.3 expl.
Movie Reviews (Zaidan et al., 2007)	*Review:* he is one of **the most exciting martial artists on the big screen**, continuing to perform his own stunts and **dazzling audiences** with his flashy kicks and punches. *Class:* Positive	1 399 Train 199 Dev 199 Test	834.9 inst. 56.18 expl.
Tweet Sentiment Extraction (TSE) [1]	*Tweet:* im soo **bored**...im deffo missing my music channels *Class:* Negative	21 983 Train 2 747 Dev 2 748 Test	20.5 inst. 9.99 expl.

Post-Hoc Explainability Methods for Decision Understanding



Example: Twitter Sentiment Extraction (TSE)

Post-Hoc Explainability for Decision Understanding: Research Questions

- How can explainability methods be evaluated?
  - Proposal: set of **diagnostic properties**

- What are characteristics of different explainability methods?
- How do explanations for models with different architectures differ?
- How do automatically and manually generated explanations differ?

Post-Hoc Explainability Methods for Decision Understanding: Gradient-Based Approaches

- Compute gradient of input w.r.t. output
- Gradient is computed for each element of vector
- Different aggregation methods used to produce one score per input token (mean average, L2 norm aggregation)

- Common approaches
  - **Saliency**
    - see above
  - **InputX-Gradient**
    - additionally multiplies gradient with input
  - **Guided Backpropagation**
    - over-writes the gradients of ReLU functions so that only non-negative gradients are backpropagated

Post-Hoc Explainability Methods for Decision Understanding: Perturbation-Based Approaches

- Replace tokens in input with other tokens to compute their relative contributions

- Common approaches
  - **Occlusion**
    - replaces each token with a baseline token and measures change in output
  - **Shapley Value Sampling**
    - computes average marginal contribution of each word across word perturbations

Post-Hoc Explainability Methods for Decision Understanding: Simplification-Based Approaches

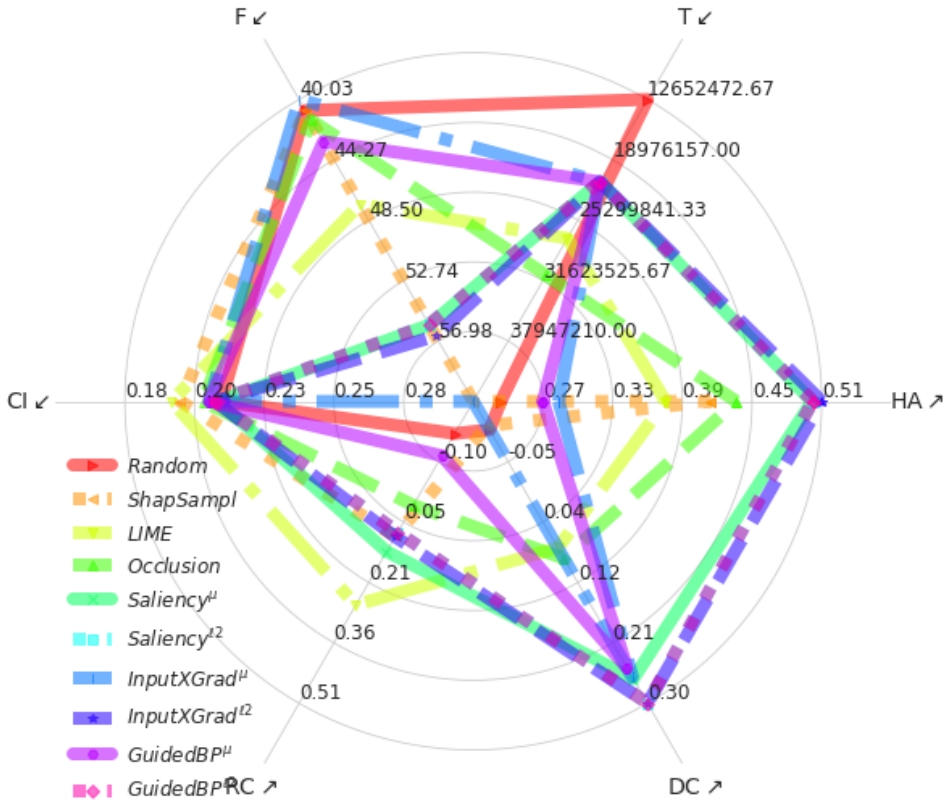
- Train local linear models to approximate local decision boundaries

- Common approaches
  - **LIME**
    - train one linear model per instance

Post-Hoc Explainability Methods for Decision Understanding: Diagnostic Properties

- **Agreement with Human Rationales (HA)**
  - Degree of overlap between human and automatic saliency scores
- **Confidence Indication (CI)**
  - Predictive power of produced explanations for model's confidence
- **Faithfulness (F)**
  - Mask most salient tokens, measure drop in performance
- **Rationale Consistency (RC)**
  - Difference between explanations for models trained with different random seeds, with model with random weights
- **Dataset Consistency (DC)**
  - Difference between explanations for similar instances

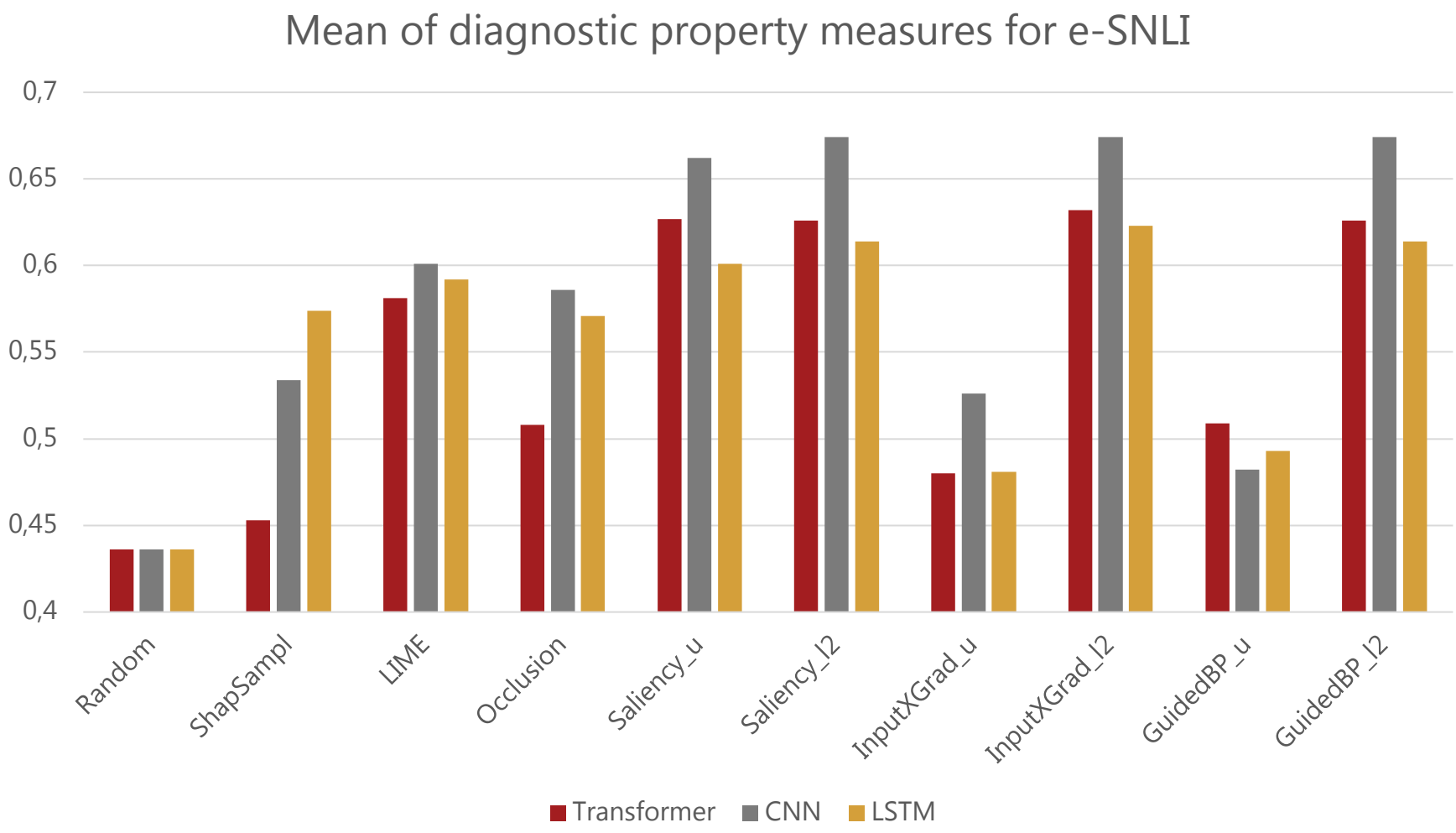
Post-Hoc Explainability Methods for Decision Understanding: Selected Results



Spider chart for Transformer model on e-SNLI

HA: Agreement with human rationales
CI: Confidence indication
F: Faithfulness
RC: Rationale Consistency
DC: Dataset Consistency

Post-Hoc Explainability Methods for Decision Understanding: Aggregated Results



Summary

- Diagnostic properties allow to assess different aspects of explainability techniques
- Gradient-based methods outperform perturbation-based and simplification-based ones for most properties across model architectures and datasets
  - Exception: Shapley Value Sampling and LIME better for Confidence Indication property
- Gradient-based methods also fastest to compute

Part 2:

Model Understanding

Overview of Today's Talk

- **Introduction**
  - Explainability – what is it and why do we need it?
- **Part 1: Decision understanding**
  - *Instance-level* explainability for text classification and fact checking
  - Language generation based explanations
  - Evaluating instance-level explanations
- **Part 2: Model understanding**
  - *Model-wide* explainability for text classification and fact checking
  - Finding model-wide explanations
  - Visualising model-wide explanations

Universal Adversarial Trigger Generation for Fact Checking

Pepa Atanasova*, Dustin Wright*,
Isabelle Augenstein

EMNLP 2020

*equal contributions

Generating Adversarial Claims

- Fact checking models can overfit to **spurious patterns**
  - Making the right predictions for the wrong reasons
  - This leads to vulnerabilities, which can be exploited by adversaries (e.g. agents spreading mis- and disinformation)

- How can one reveal such **vulnerabilities**?
  - Generating instance-level explanations for fact checking models (first part of talk)
  - Generating adversarial claims (this work)

Previous Work

- **Universal adversarial attacks** (Gao and Oates, 2019; Wallace et al, 2019)
  - Single perturbation changes that can be applied to many instances
  - Change the meaning of the input instances and thus produce label-incoherent claims
  - Are not per se semantically well-formed

- **Rule-based perturbations** (Riberio et al., 2018)
  - Semantically well-formed, but require hand-crafting patterns

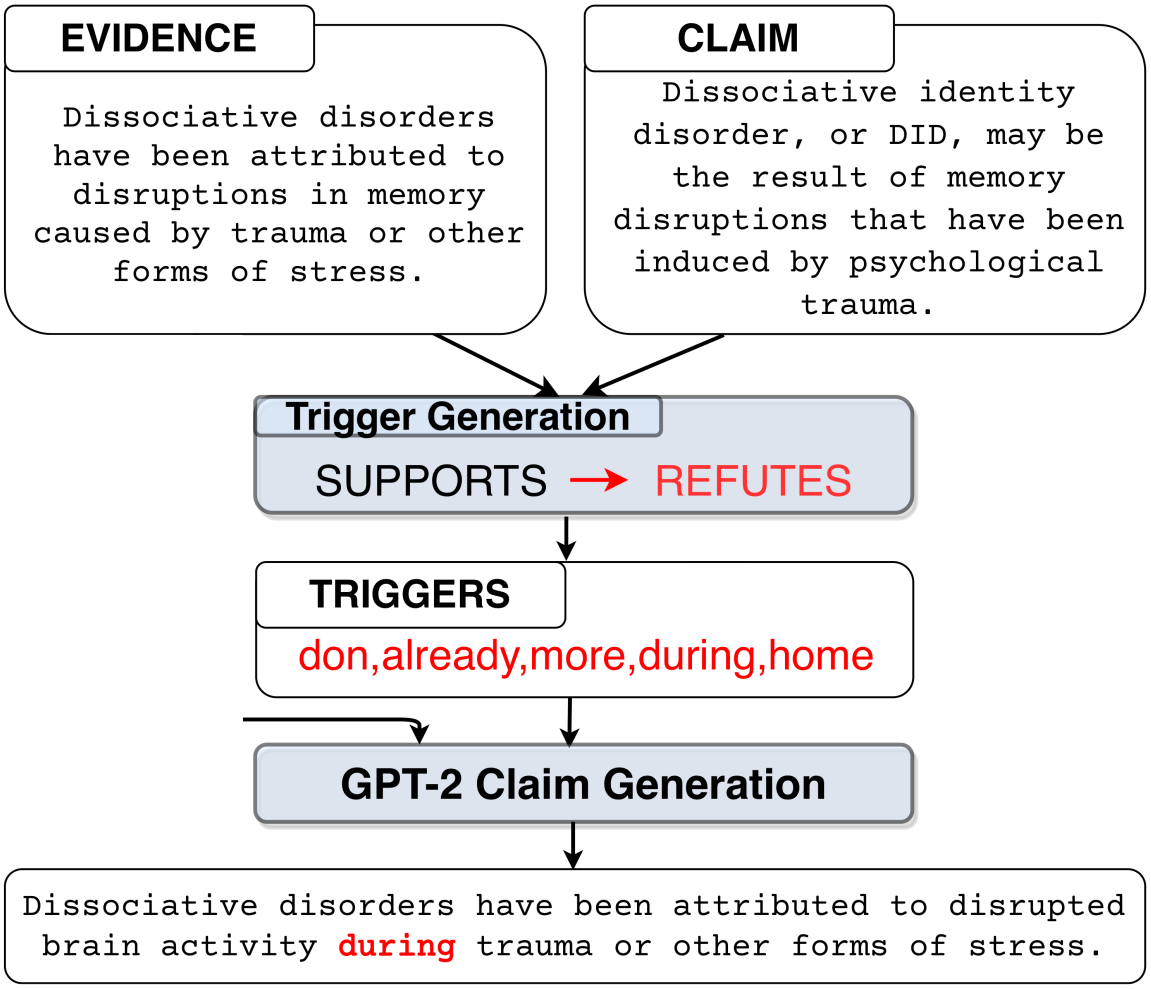
Previous Work

- **FEVER 2.0 shared task** (Thorne et al., 2019)
  - Builders / breakers setup
  - Methods of submitted systems:
    - Producing claims requiring multi-hop reasoning (Niewinski et al., 2019)
    - Generating adversarial claims manually (Kim and Allan, 2019)

Goals of this Work

- Generate claims fully automatically
- Preserve the meaning of the source text
- Produce semantically well-formed claims

Model



RoBERTa-based FEVER model to predict FC label

1) HotFlip attack model to find triggers
2) STS auxiliary model to preserve FC label

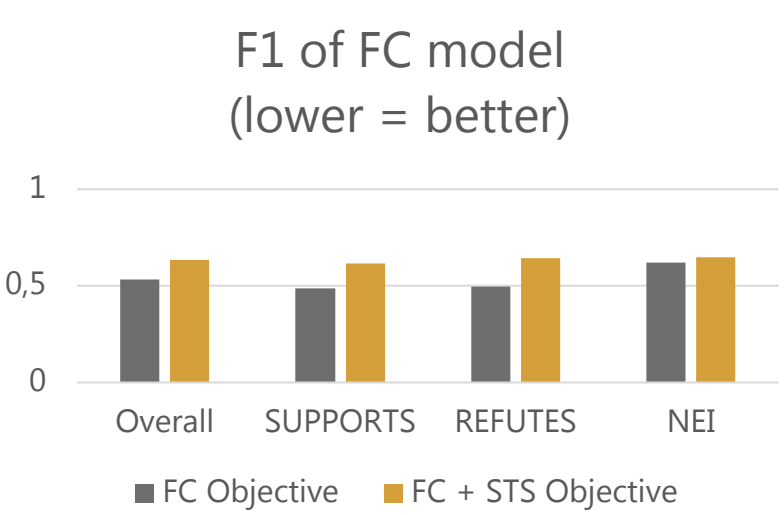
Claim generation conditioned on evidence and triggers

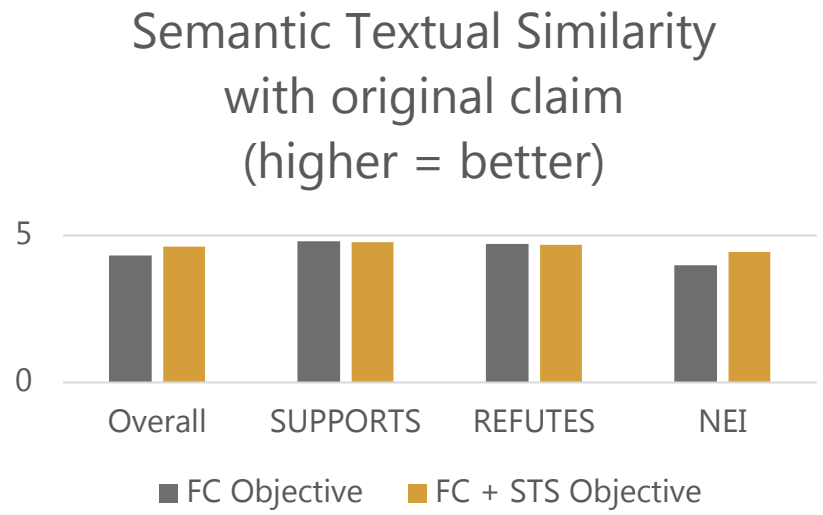
Examples of Generated Claims

Evidence	Triggers	Generated Claim
	SUPPORTS Claims	
Since the 19th century, some Romani have also migrated to the Americas.	don,already,more,during,home	Romani have moved to the Americas during the 19th century.
	REFUTES Claims	
Times Higher Education World University Rankings is an annual publication of university rankings by Times Higher Education (THE) magazine.	interested,reward,visit,consumer,conclusion	Times Higher Education World University Rankings is a consumer magazine.
	NOT ENOUGH INFO Claims	
The KGB was a military service and was governed by army laws and regulations , similar to the Soviet Army or MVD Internal Troops.	nowhere,only,none,no,nothing	The KGB was only controlled by a military service.

Results: Manual Evaluation

- How well can we generate universal adversarial triggers?





- Trade-off between how potent the attack is (reduction in F1) vs. how semantically coherent the claim is (STS)
- Reduction in F1 for both trigger generation methods
- Macro F1 of generated w.r.t. original claim: 56.6 (FC Objective); 60.7 (FC + STS Objective) -- STS Objective preserves meaning more often

Key Take-Aways

- Novel extension to the HotFlip attack for universarial adversarial trigger generation (Ebrahimi et al., 2018)
- Conditional language model, which takes trigger tokens and evidence, and generates a semantically coherent claim

- Resulting model generates semantically coherent claims containing universal triggers, which preserve the label
- Trade-off between how well-formed the claim is and how potent the attack is

TX-Ray: Quantifying and Explaining Model-Knowledge Transfer in (Un-)Supervised NLP

Nils Rethmeier, Vageesh Kumar Saxena, Isabelle Augenstein

UAI 2020

Motivation: observe knowledge acquisition of neural nets

Problems of supervised probing task evaluation setup

- only **measures expected** (probed) model knowledge and **semantics**
- is misleading when model and probe domains mismatch
- probing annotation can not scale to uncover unforeseen semantics

Goal: instead can we visualise, quantify and explore how a (language) model

- learns → RQ (1) How does self-supervision abstract knowledge?
- applies → RQ (2) How is knowledge (zero-shot) applied to new inputs X?
- adapts → RQ (3) How is knowledge adapted by supervision?

Approach: un-/ self-supervised interpretability

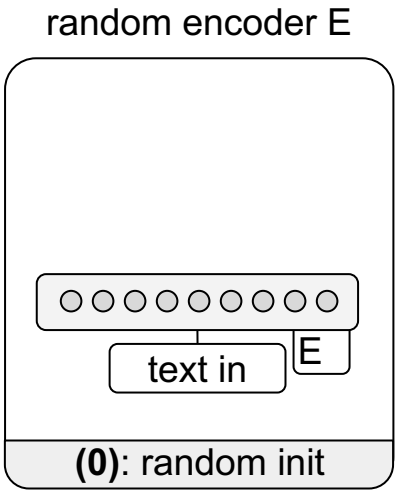
- visualise what input (features) each neuron prefers (maximally activates on) -- i.e. **activation maximisation** by Erhan, 2009 -- used on RBMs [1]

Goal: Visualise & measure transfer in neural nets

How does each neuron ◯

RQ (1) abstract/ learn (textual) **knowledge?** ◯

- during initial/ pre-training on text X*pre*

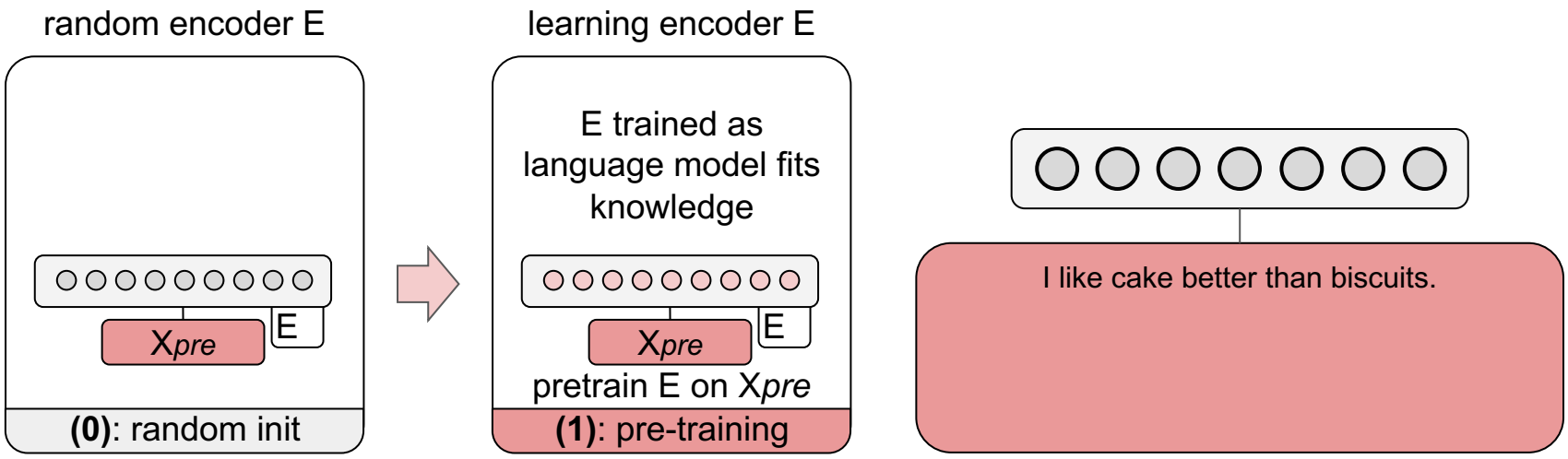


Goal: Visualise & measure transfer in neural nets

How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ○

- during initial/ pre-training on text X*pre*

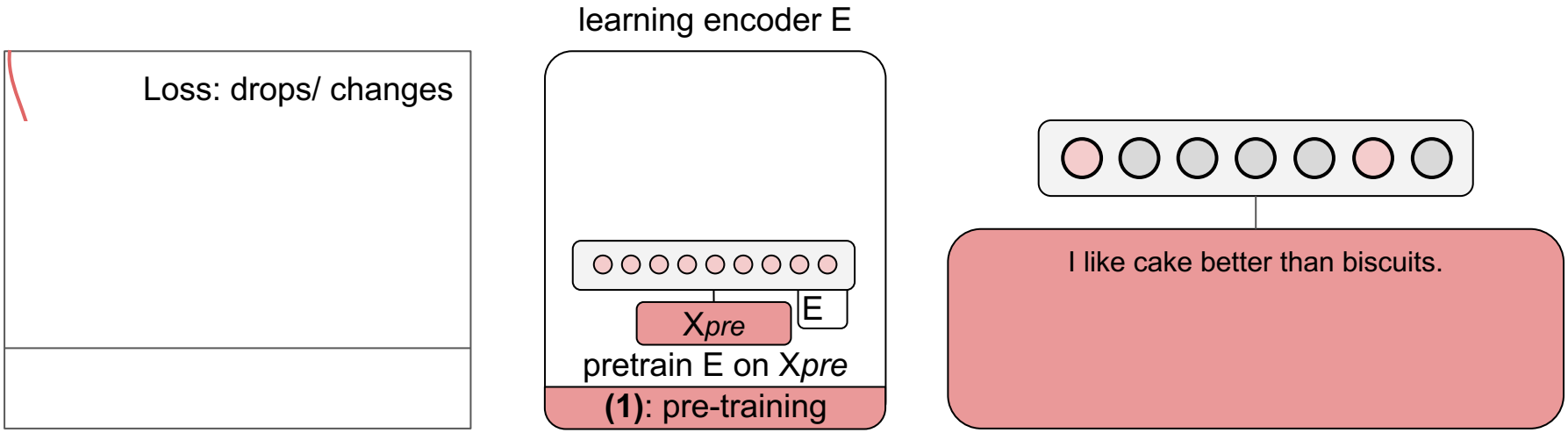


Goal: Visualise & measure transfer in neural nets

How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ○

- during initial/ pre-training on text X*pre*

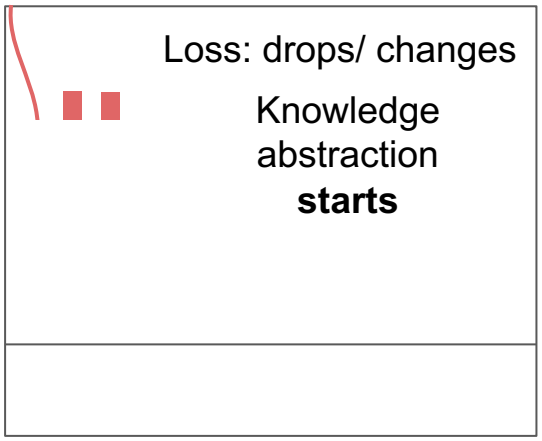


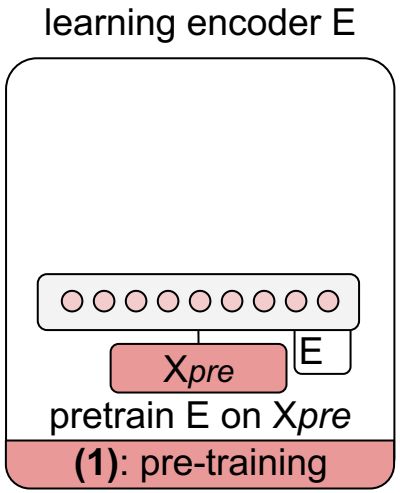
Goal: Visualise & measure transfer in neural nets

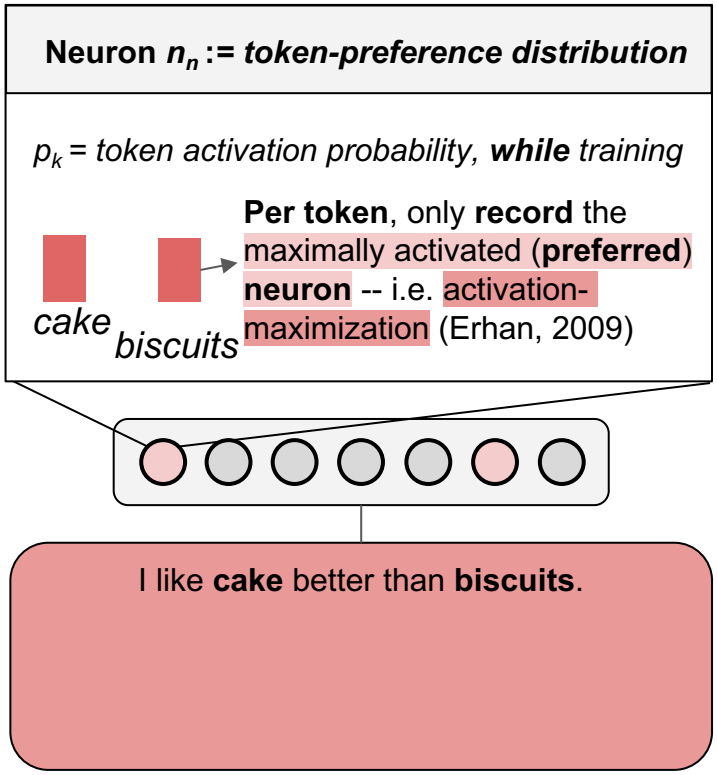
How does each neuron ◯

RQ (1) abstract/ learn (textual) **knowledge?** ◯

- during initial/ pre-training on text X*pre*





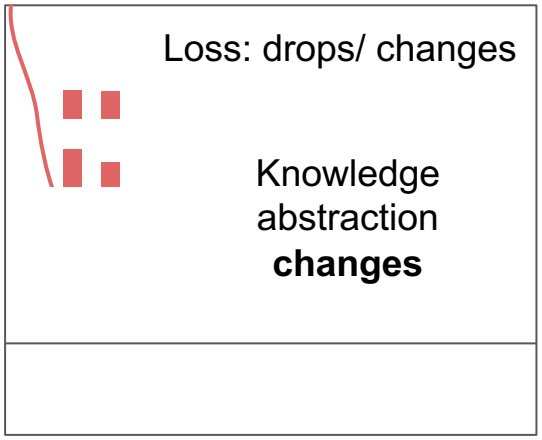


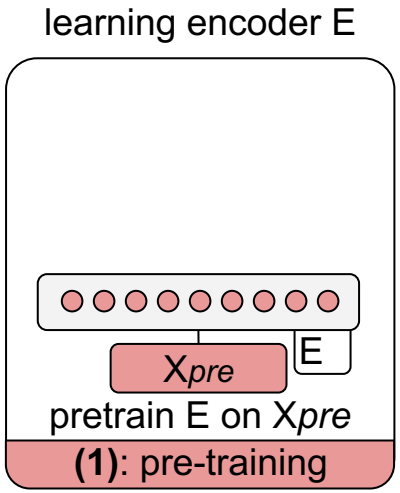
Goal: Visualise & measure transfer in neural nets

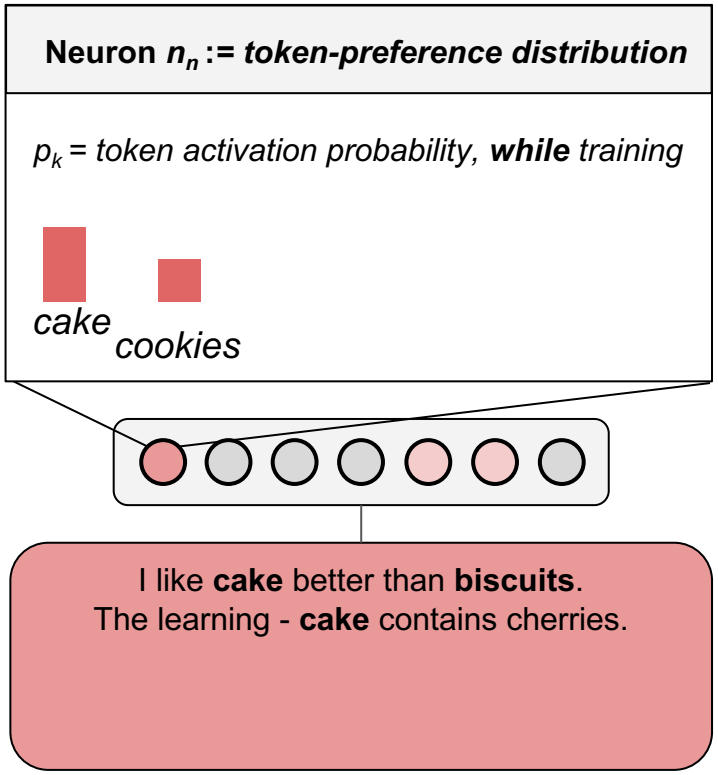
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ○

- during initial/ pre-training on text X*pre*





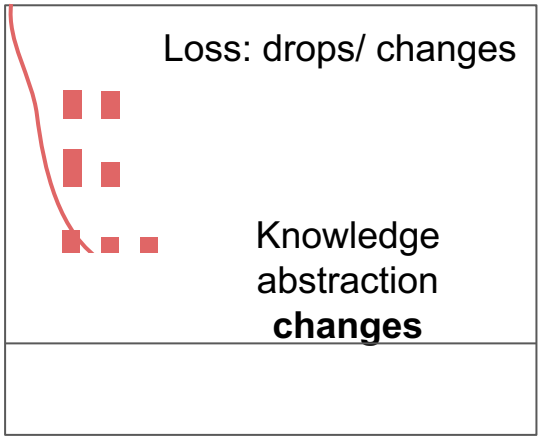


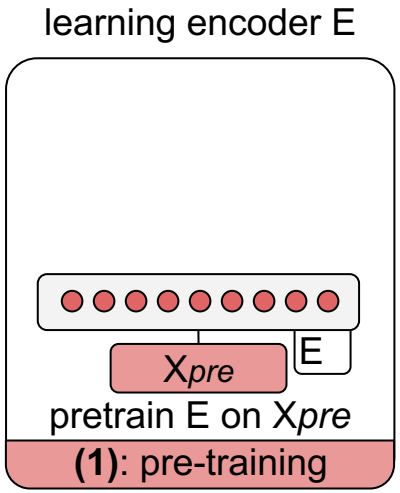
Goal: Visualise & measure transfer in neural nets

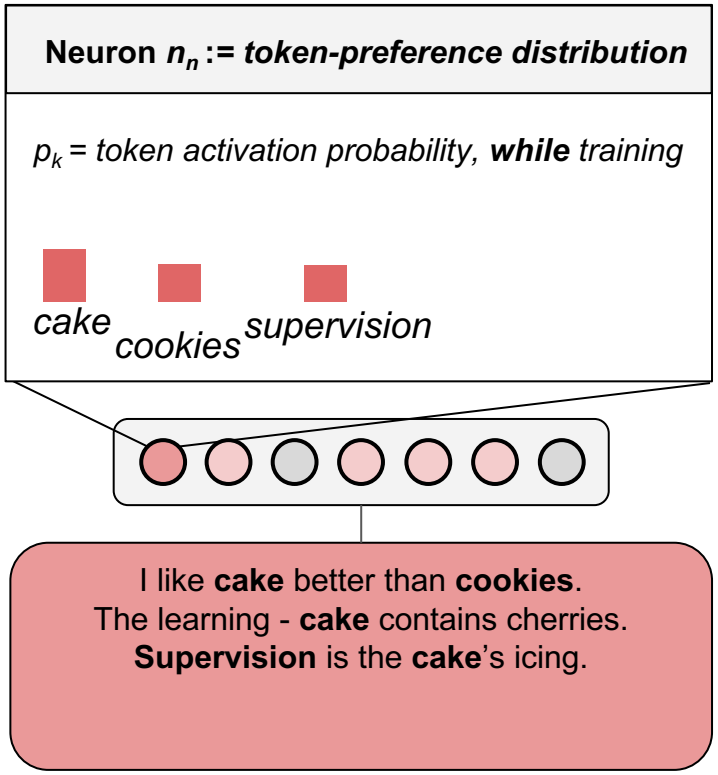
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ○

- during initial/ pre-training on text X*pre*





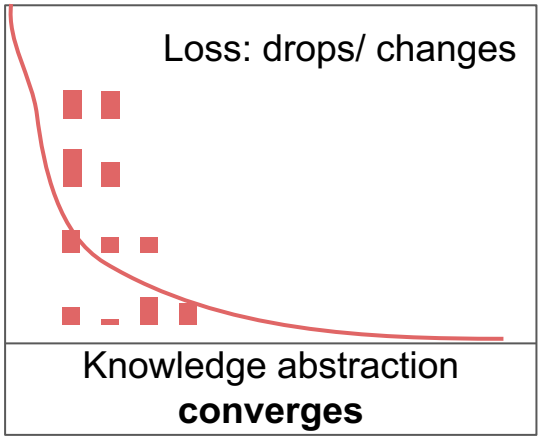


Goal: Visualise & measure transfer in neural nets

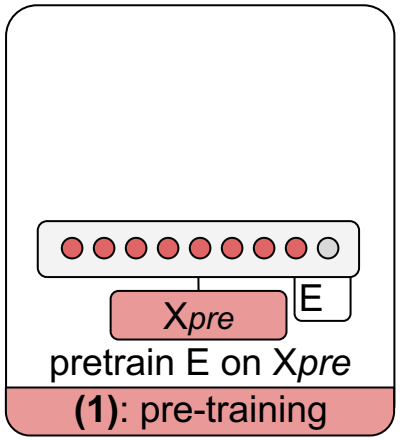
How does each neuron ○

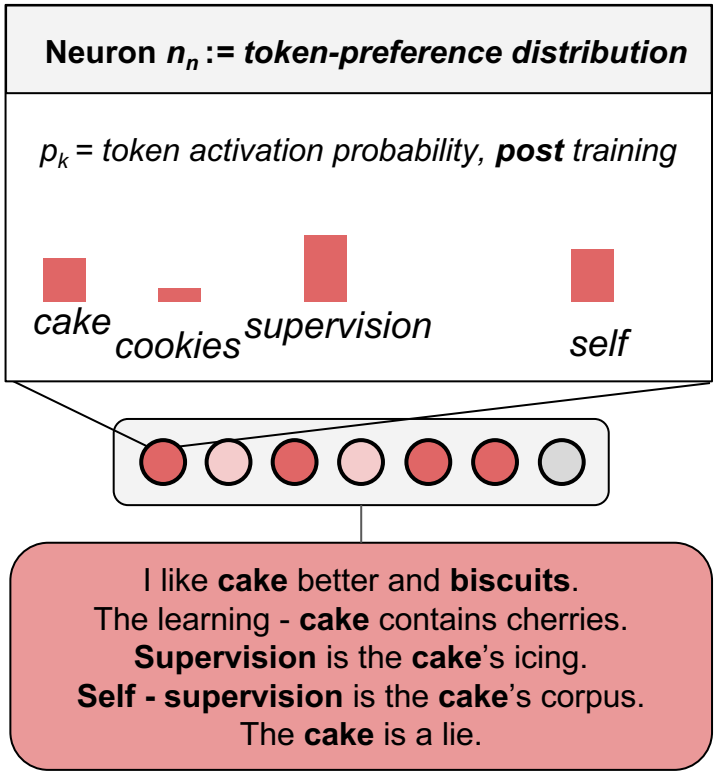
RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*



converged encoder E





Goal: Visualise & measure transfer in neural nets

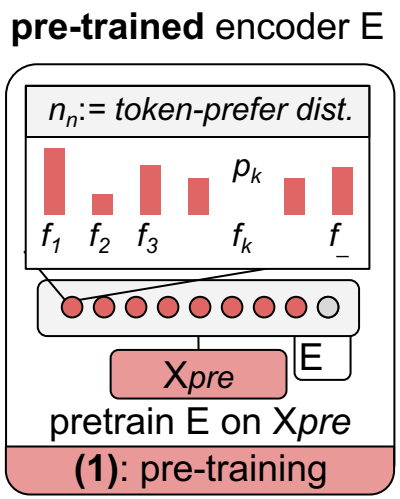
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*

Take-away (1):

(pre)-trained builds neural knowledge (feature activation distributions)



Goal: Visualise & measure transfer in neural nets

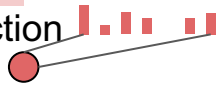
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

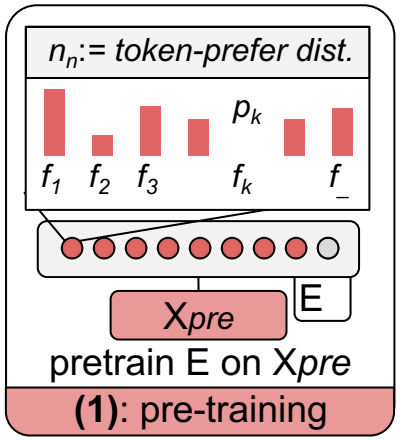
- during initial/ pre-training on text X*pre*

Take-away (1):

(pre)-trained builds neural knowledge (feature activation distributions)

Token-activation distributions visualize the knowledge abstraction  of each neuron

pre-trained encoder E



Goal: Visualise & measure transfer in neural nets

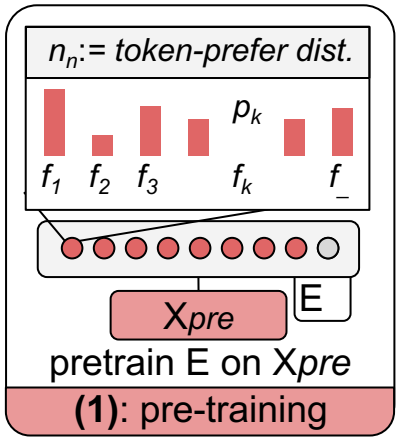
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*

RQ (2) apply learned knowledge ●

- to new domain text X*end*



Goal: Visualise & measure transfer in neural nets

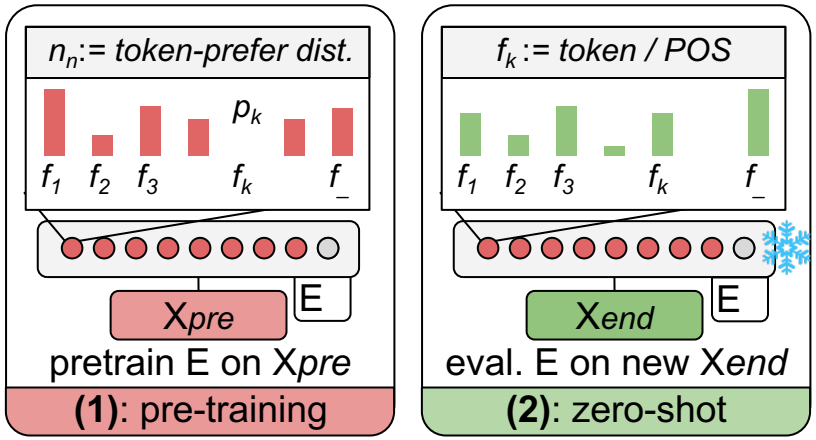
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*

RQ (2) apply learned knowledge ●

- to new domain text X*end*



- feed new text Xend to frozen pretrained encoder ●●●

Goal: Visualise & measure transfer in neural nets

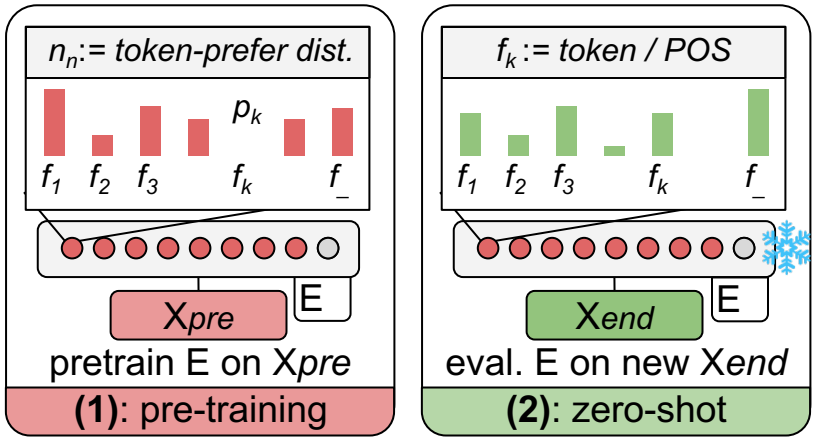
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*

RQ (2) apply learned knowledge ●

- to new domain text X*end*



- activation distribution changes
- feed new text Xend to frozen pretrained encoder

Goal: Visualize & measure transfer in neural nets

We use Hellinger H(,) distance as **change/ transfer measure** -- i.e. a symmetric KLD

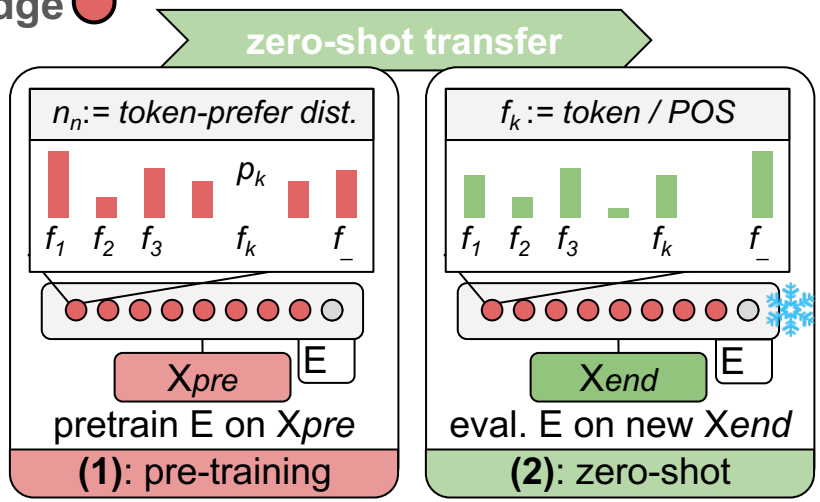
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*

RQ (2) apply learned knowledge ●

- to new domain text X*end*



- no neuron transfer if large change vs.
- neuron transfers if small change vs.
- activation distribution changes
- feed new text Xend to frozen pretrained encoder

Goal: Visualise & measure transfer in neural nets

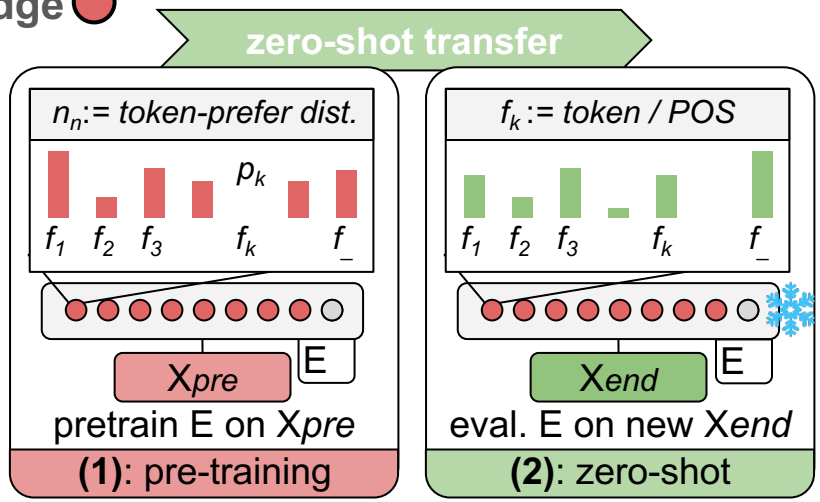
How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*

RQ (2) apply learned knowledge ●

- to new domain text X*end*



Over-specialisation:

- no neuron transfer if large change vs.

OOD-generalisation:

- neuron transfers if small change vs.

Goal: Visualise & measure transfer in neural nets

How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*

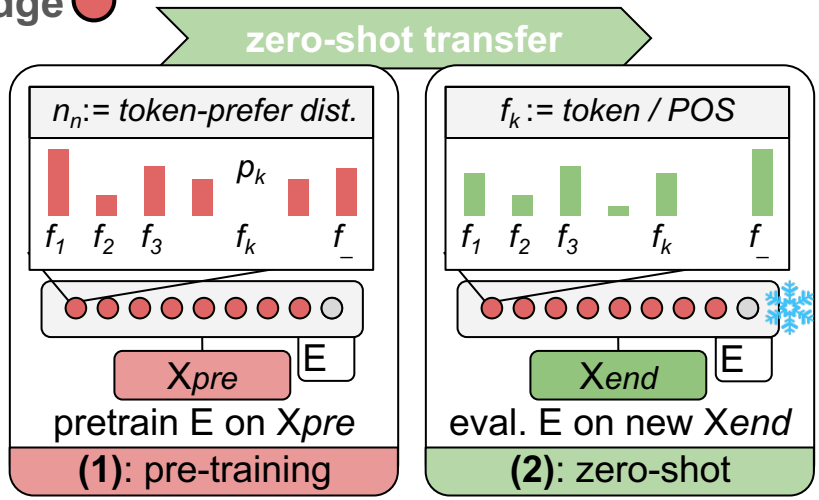
RQ (2) apply learned knowledge ●

- to new domain text X*end*

Take-away (2):

Transfer if encoder ●●● **generalises** well (activates similarly) on the new text

Shows how able each neuron is generalise to new data(-distrib.)!



Goal: Visualise & measure transfer in neural nets

How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●

- during initial/ pre-training on text X*pre*

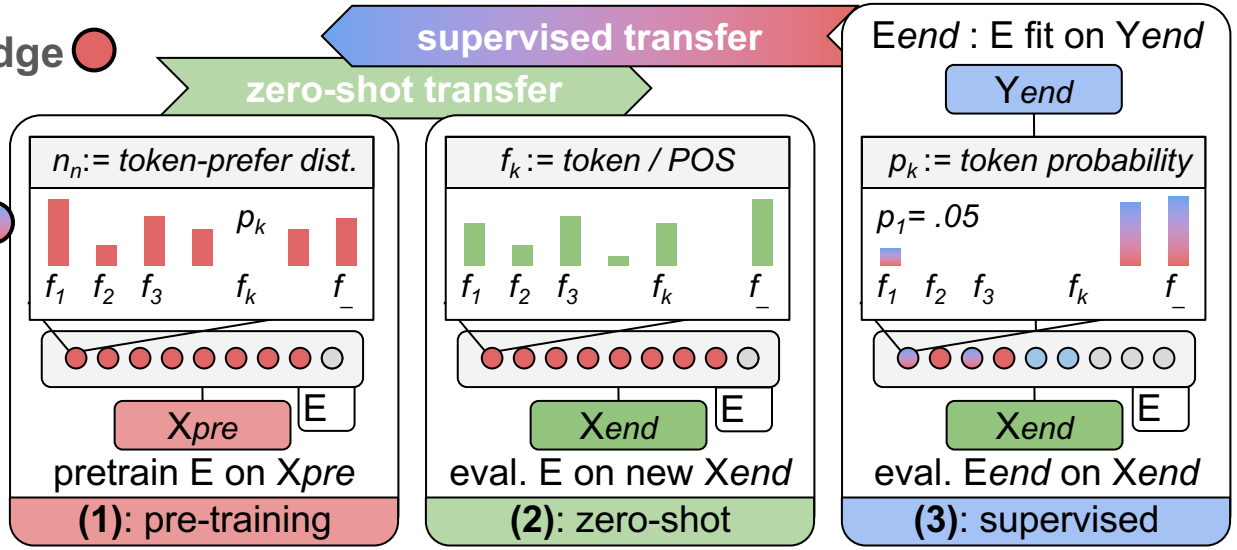
RQ (2) apply learned knowledge ●

- to new domain text X*end*

RQ (3) adapt its knowledge ●

- to suit a supervision signal Y*end* on the new domain X*end*

  (no more LM loss/ objective)



Goal: Visualise & measure transfer in neural nets

How does each neuron ○

RQ (1) abstract/ learn (textual) **knowledge?** ●
- during initial/ pre-training on text X*pre*

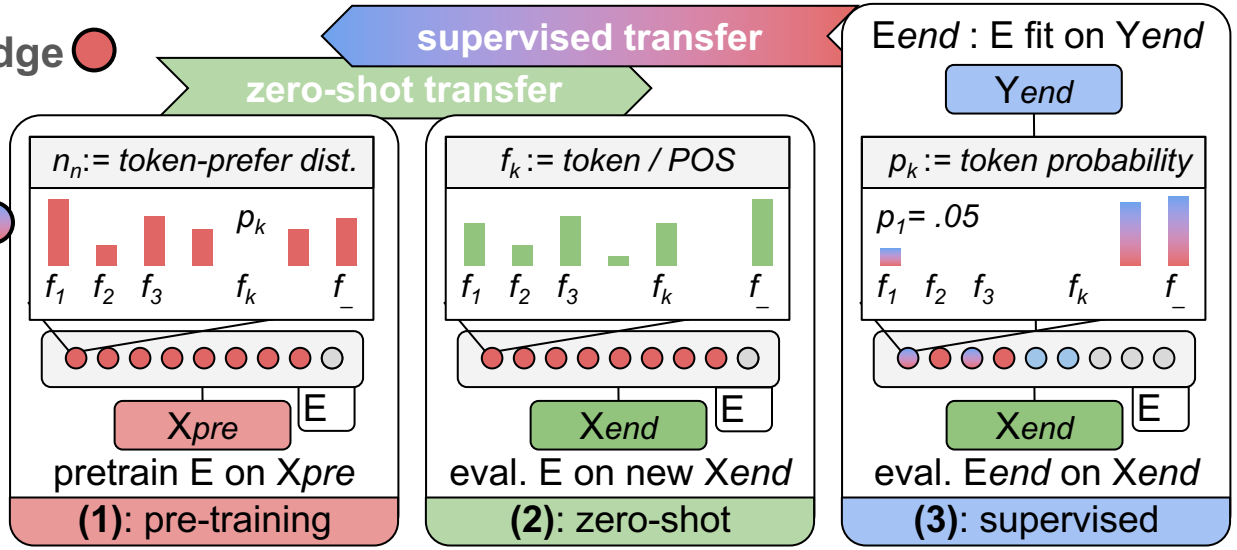
RQ (2) apply learned knowledge ●
- to new domain text X*end*

RQ (3) adapt its knowledge ●
- to suit a supervision signal Y*end* on the new domain X*end*

(no more LM loss/ objective)

Take-away (3):

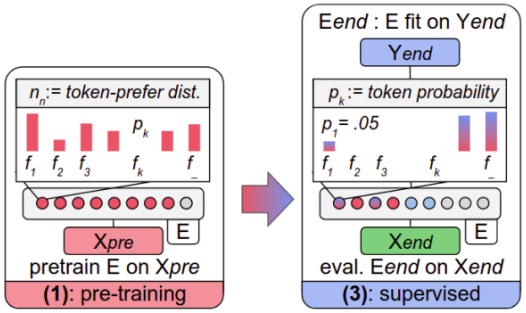
- ● supervision specialises (re-fits) knowl.
- ○ adds new knowledge
- ○ and 'avoids' (sparsifies) old knowledge



Experiment: **XAI to guide pruning**

Pretrain, then supervise

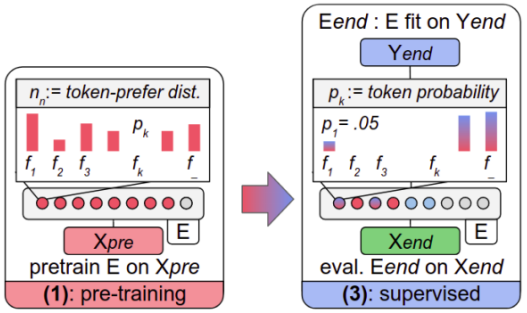
- pretrain language model on Wikitext-2 **(1)**
- fine tune to IMDB binary reviews **(3)**



Experiment: **XAI to guide pruning**

Pretrain, then supervise

- pretrain language model on Wikitext-2 **(1)**
- fine tune to IMDB binary reviews **(3)**



Prune neurons that post supervision …

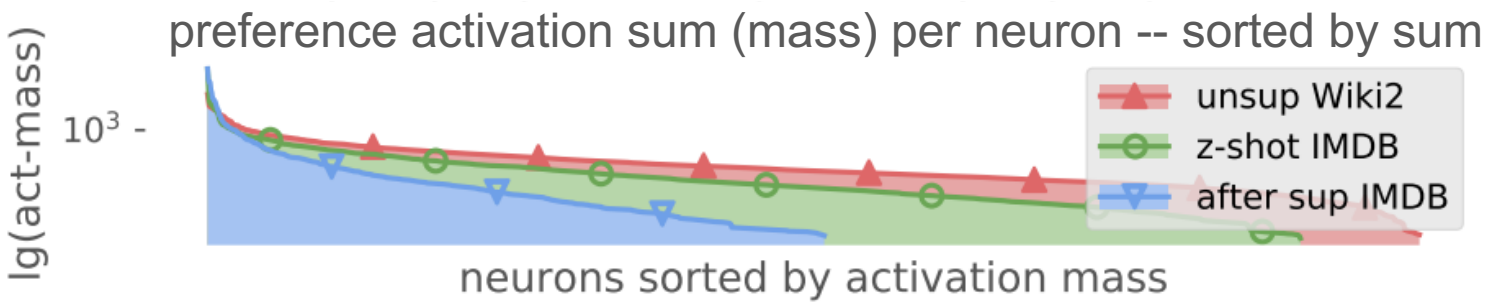
- were specialized (re-fit)
- became **preferred** (activated)
- were 'avoided' (now empty distrib.)
  → i.e, neuron has no max activations post supervision

Which neurons prunded?	% AM of 675	F1 change % train	F1 change % test	pruning effect
none = baseline	100.000	0.00	0.00	–
A: 740 **avoided**	–	3.65	2.80	↓ noise, ↑ generality
B: 20 least **prefered**	0.004	-3.79	0.00	↓ over-fitting
C: 20 top **prefered**	83.120	-4.99	-1.43	↓ generalization
D: 85 **sup added**	3.006	-3.71	-3.87	↓ sup. knowledge
activation mass (**AM**)		(+) is better		standard / TX-Ray

Take-Aways

TX-Ray can **explore** generalisation and specialisation at individual neuron-level

- self-supervised pre-training builds general knowledge
  - preference spread across many neurons -- 89% maximally active (preferred) neurons
- zero-shot application shows match of model knowledge vs new domain
  - preference less spread -- 88% of neurons preferred, partial generalisation
- supervised knowledge fine-tuning sparsifies (concentrates) activation
  - preference peaked -- only 45% of neurons preferred, many become domain over-specialised



Wrap-Up

Overall Take-Aways

- **Why** explainability?
  - understanding if a model is right for the right reasons
- Generated explanations can **help users understand**:
  - inner workings of a model (model understanding)
  - how a model arrived at a prediction (decision understanding)
- Explainability can **enable**:
  - human-in-the-loop model development
  - human-in-the-loop data selection

Overall Take-Aways

- **Caveats**:
  - There can be more than one correct explanation
  - Different explainability methods provide different explanations

- Different streams of explainability methods have different **benefits and downsides**
  - Black box vs. white box
  - Hypothesis testing vs. bottom-up understanding
  - Requirements for annotated training data
  - Joint vs. post-hoc explanation generation
  - One-time analysis vs. continuous monitoring
  - Perform well w.r.t. different properties

Where to from here?

- Making explanations useful to **model trainers** and **end users**
  - How to interpret different explanations?
  - Explanations for models with large number of parameters
  - What-if analyses

- Some research on generating explanations, relatively little work on understanding **in what context they are useful**
  - (Automatically) evaluating explanations
  - Human-in-the-loop development

Thank you!

isabelleaugenstein.github.io
augenstein@di.ku.dk
@IAugenstein
github.com/isabelleaugenstein

Thanks to my PhD students and collaborators!

Pepa Atanasova

Dustin Wright

Jakob Grue Simonsen

Christina Lioma

Nils Rethmeier

Vageesh Kumar Saxena

CopeNLU

https://copenlu.github.io/

Presented Papers

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein. *Generating Fact Checking Explanations*. In Proceedings of ACL 2020.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, Isabelle Augenstein. *A Diagnostic Study of Explainability Techniques for Text Classification*. EMNLP 2020.

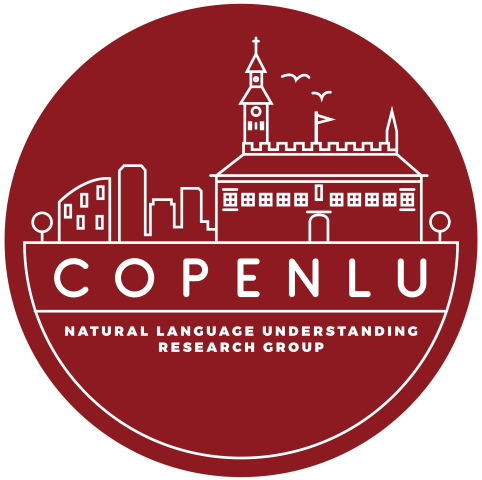
Pepa Atanasova*, Dustin Wright*, Isabelle Augenstein. *Generating Label Cohesive and Well-Formed Adversarial Claims*. EMNLP 2020.

Nils Rethmeier, Vageesh Kumar Saxena, Isabelle Augenstein. *TX-Ray: Quantifying and Explaining Model-Knowledge Transfer in (Un)Supervised NLP*. In Proceedings of the Conference on UAI 2020.

*equal contributions

Hiring

2 PhD students, 1 postdoc – explainable stance detection
- funded by Danish Research Council (DFF) Sapere Aude grant
- application deadline PhD: 31 January 2021
- start date: Summer/Autumn 2021
- PhD: https://employment.ku.dk/faculty/?show=153150
- Postdoc: https://tinyurl.com/yd6jw5nm



1 postdoc – explainable IE & QA
- funded by Innovation Foundation Denmark
- application deadline: 28 February 2021
- start date: Autumn 2020
- Postdoc: https://employment.ku.dk/faculty/?show=153308

Reading Materials on Explainability

Decision Understanding

- Camburu, O. M., Shillingford, B., Minervini, P., Lukasiewicz, T., & Blunsom, P. (2020, July). Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4157-4165).
  - Reverse explanations to help prevent inconsistencies

- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020, February). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).
  - Perturbation-based explanation methods be fooled easily, are potentially unreliable

Reading Materials on Explainability

Model Understanding

- Geva, M., Schuster, R., Berant, J., & Levy, O. (2020). Transformer Feed-Forward Layers Are Key-Value Memories. *arXiv preprint arXiv:2012.14913*.
  - how to use activation maximisation to analyse memorisation in Transformers

- Durrani, N., Sajjad, H., Dalvi, F., & Belinkov, Y. (2020). Analyzing Individual Neurons in Pre-trained Language Models. *arXiv preprint arXiv:2010.02695*.
  - pruning neurons and how many neurons are used per task

Reading Materials on Explainability

Surveys, Opinion Papers & Benchmarks

- Eraser benchmark with explainability datasets: https://www.eraserbenchmark.com/
- Jacovi, A., & Goldberg, Y. (2020). Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? In *ACL 2020*.
- Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2020). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *arXiv preprint arXiv:2010.07487*.
- Kotonya, N., & Toni, F. (2020). Explainable Automated Fact-Checking: A Survey. *In Coling 2020*.

Reading Materials on Explainability

Tutorials

- Yonatan Belinkov, Sebastian Gehrmann, Ellie Pavlick. Interpretability and Analysis in Neural NLP. Tutorial at *ACL 2020*. https://www.aclweb.org/anthology/2020.acl-tutorials.1/
- Eric Wallace, Matt Gardner, Sameer Singh. Interpreting Predictions of NLP Models. Tutorial at *EMNLP 2020*. https://www.aclweb.org/anthology/2020.emnlp-tutorials.3/
- Hima Lakkaraju, Julius Adebayo, Sameer. Explaining Machine Learning Predictions -- State-of-the-art, Challenges, and Opportunities. Tutorials at *NeurIPS 2020, AAAI 2021*. https://explainml-tutorial.github.io/
- Jay Alammar. Interfaces for Explaining Transformer Language Models. *Blog, 2020*. https://jalammar.github.io/explaining-transformers/