Special Topic Presentation

Probing Pre-trained Language Models: A Case Study of Coordination Using Causal Mediation Analysis

Presenters: Pangbo Ban, Yifan Jiang, Tianran Liu (group 6)
May 25, 2022

Outline

1. [paper] ConjNLI: Natural Language Inference Over Conjunctive Sentences (Saha et al., EMNLP 2020)
2. [paper] Investigating gender bias in language models using causal mediation analysis (Vig et al., NeurIPS 2020)
3. An overview of our project: Probing Pre-trained Language Models: A Case Study of Coordination Using Causal Mediation Analysis

Paper #1

ConjNLI: Natural Language Inference Over Conjunctive Sentences (Saha et al., 2020)

Natural Language Inference

- A task to decide given a premise, whether a hypothesis is true (entailment), false (contradiction), or unknown (neutral).

A man inspects the uniform of a figure in some East Asian country.	**contradiction** C C C C C	The man is sleeping
An older and younger man smiling.	**neutral** N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	**contradiction** C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	**entailment** E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** N N E C N	A happy woman in a fairy costume holds an umbrella.

Examples from Bowman et al.(2015)

Motivations

- Most examples in existing NLI datasets do not require inferences over the conjuncts that are connected by the coordinating word.
    - e.g. *Man and woman are sitting <u>on the sidewalk</u>.* & *Man and woman are sitting.*
- There are almost no examples with non-boolean conjunctions.
    - Boolean coordination: “A and B” is true -> A is true and B is true
    - e.g. *A total of five men <u>and women</u> are sitting.* & *A total of five men are sitting.*
- State-of-the-art models such as BERT and RoBERTa often fail to make inferences for samples with non-boolean conjuncts.

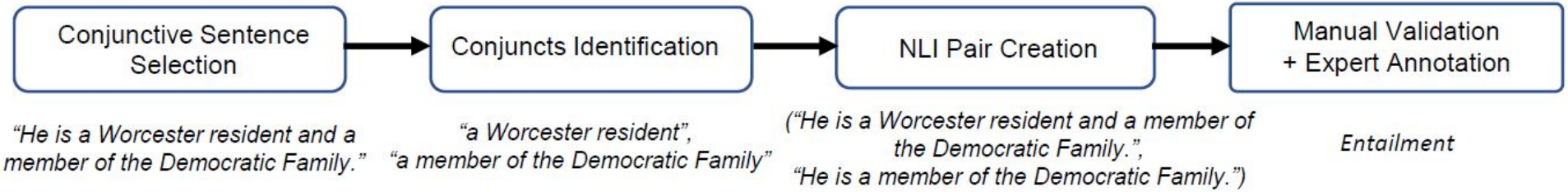
Examples from Bowman et al.(2015)

Contributions

- This paper introduces ConjNLI, a new adversarial dataset for NLI over diverse and challenging conjunctive sentences.
- They also present some initial model advancements that attempt to alleviate some of these challenges in ConjNLI.
    - Iterative adversarial fine-tuning RoBERTa
    - Initial predicate-aware (SRL) RoBERTa
    - Predicate-aware RoBERTa with adversarial fine-tuning

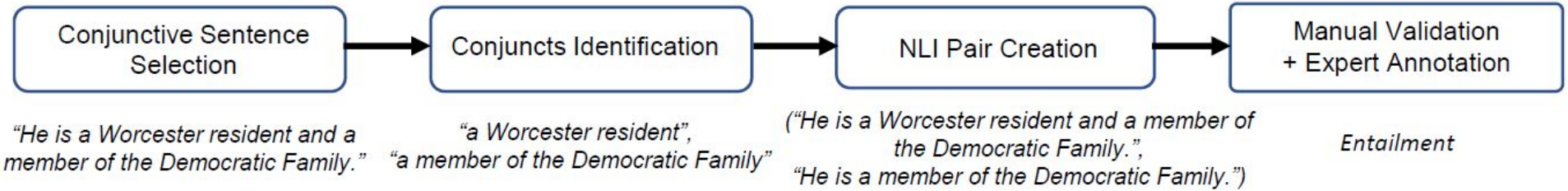
Data Creation

- **Conjunctive Sentence Selection**
  - Wikipedia
- **Conjuncts Identification**
  - Constituency parser (AllenNLP)
- **NLI Pair Creation**
  - Remove, add or replace one of the two conjuncts to obtain another sentence.



Data Creation (continued)

- **Manual Validation & Expert Annotation**
  - Pairs are first manually verified for grammaticality, then annotated by two English-speaking experts (with prior experience in NLI and NLP).
  - Round #1: each annotator annotated the examples independently.
  - Round #2: disagreements are discussed to resolve final labels.
  - Inter-annotator agreement: 0.83 (Cohen's Kappa)



Data Analysis

	Ent	Neu	Contra	Total
Conj Dev	204	281	138	623
Conj Test	332	467	201	1000
Conj All	536	748	339	1623

Table 2: Dataset splits of CONJNLI.

	and	or	but	multiple	quant	neg
Conj Dev	320	293	99	152	131	70
Conj Test	537	471	135	229	175	101
Conj All	857	764	234	381	306	171

Table 3: Data analysis by conjunction types, presence of quantifiers and negations.

Sentence	CT
Historically, the Commission was run by **three commissioners** or **fewer**.	NP + Adj
Terry Phelps and **Raffaella Reggi** were the defending champions but did not compete that year.	NP + NP
Terry Phelps and Raffaella Reggi **were the defending champions** but **did not compete that year**.	VP + VP
It is for Orienteers **in** or **around** North Staffordshire and South Cheshire.	Prep + Prep
It is a white solid, but **impure samples can appear yellowish**.	Clause + Clause
Pantun were originally not written down, the bards often being **illiterate** and **in many cases blind**.	Adj + PP
A queue is **an example of a linear data structure**, or **more abstractly a sequential collection**.	NP + AdvP

Table 4: CONJNLI sentences consist of varied syntactic conjunct categories (bolded). CT = Conjunct Types, NP = Noun Phrase, VP = Verb Phrase, AdvP = Adverbial Phrase.

Recap: Contributions

- This paper introduces ConjNLI, a new adversarial dataset for NLI over diverse and challenging conjunctive sentences.
- They also present some initial model advancements that attempt to alleviate some of these challenges in ConjNLI.
    - Iterative adversarial fine-tuning RoBERTa
    - Initial predicate-aware (SRL) RoBERTa
    - Predicate-aware RoBERTa with adversarial fine-tuning

Methods

- **Iterative Adversarial Fine-Tuning**
  - Automated Adversarial Training Data Creation (15k)
    - expert human annotation phase -> automated boolean rules + some heuristics for non-boolean semantics
    - For “boolean and”, “A and B” is true -> A and B are individually true.
      - a conjunct is removed -> entailment
        - a conjunct is removed from a named entity -> neutral (Hoeksema, 1988; Krifka, 1990).
        - e.g. *Franklin and Marshall College* & *Franklin College* -> neutral

Methods (continued)

- **Iterative Adversarial Fine-Tuning**
  - Automated Adversarial Training Data Creation
    - expert human annotation phase -> automated boolean rules + some heuristics for non-boolean semantics
    - For “boolean and”, “A and B” is true -> A and B are individually true.
      - a conjunct is removed -> entailment
      - a conjunct is added -> neutral
      - a conjunct is replaced -> contradiction

Methods (continued)

- **Iterative Adversarial Fine-Tuning**
  - Automated Adversarial Training Data Creation
    - expert human annotation phase -> automated boolean rules + some heuristics for non-boolean semantics
    - For “non-boolean and”, look for trigger words like “total”, “group”, “combined”, etc.
      - e.g. *In total, the flooding and landslides killed 3,185 people in China.* & *In total, the landslides killed 3,185 people in China.* -> contradiction

Methods (continued)

- **Iterative Adversarial Fine-Tuning**
  - Algorithm for Iterative Adversarial Fine-Tuning

Algorithm 1 Iterative Adversarial Fine-Tuning

1: $model$ = finetune($RoBERTa$, $MNLI_{train}$)
2: adv_train = get_adv_data()
3: k = len(adv_{train})
4: **for** $e = 1$ to num_epochs **do**
5: $MNLI\_small$ = sample_data($MNLI_{train}$, k)
6: all_data = $MNLI\_small \bigcup adv\_train$
7: Shuffle all_data
8: $model$ = finetune($model$, all_data)
9: **end for**

Methods (continued)

- **Initial Predicate-Aware (SRL) RoBERTa**
  - Semantic Role Labeling

Premise	Hypothesis	Label	SRL Tags
It premiered on 27 June 2016 and airs Mon-Fri 10-11pm IST.	It premiered on 28 June 2016 and airs Mon-Fri 10-11pm IST.	contra	ARG1:“It”, Verb:“premiered”, Temporal:“on 27 June 2016”
He also played in the East-West Shrine Game and was named MVP of the Senior Bowl.	He also played in the North-South Shrine Game and was named MVP of the Senior Bowl.	neutral	ARG1: “He”, Discource:“also”, Verb:“played”, Location:“in the East-West Shrine Game”.

Table 5: Two examples from CONJNLI where SRL tags can help the model predict the correct label.

Methods (continued)

- **Initial Predicate-Aware (SRL) RoBERTa**
  - Motivation: late fusion of syntactic information for NLI (Pang et al., 2019)
- **Predicate-aware RoBERTa with adversarial fine-tuning**
  - Combining them together!

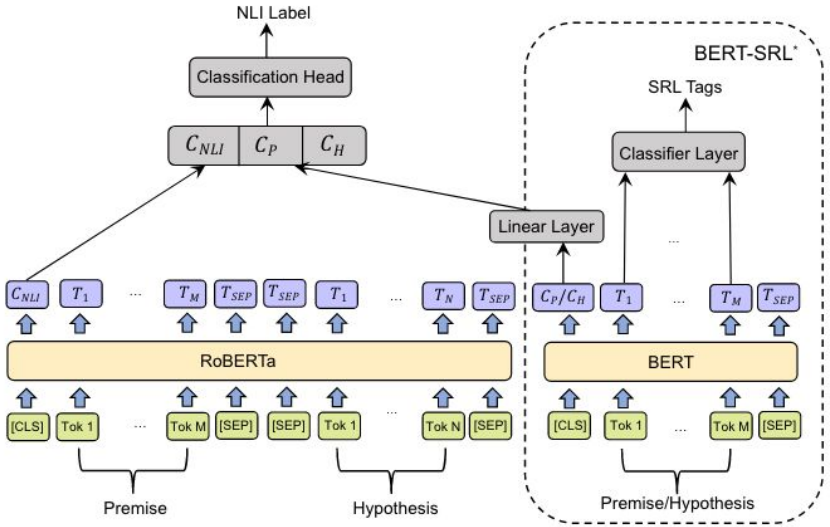


Figure 2: Architecture diagram of predicate-aware RoBERTa model for CONJNLI. * = BERT-SRL weights are frozen while fine-tuning on the NLI task.

Results

	Conj Dev	MNLI Dev	Conj Test
BERT	58.10	84.10/83.90	61.40
RoBERTa	64.68	87.56/87.51	65.50
PA	64.88	**87.75/87.63**	66.30
IAFT	**69.18**	86.93/86.81	**67.90**
PA-IAFT	68.89	87.07/86.93	67.10

Table 8: Comparison of all our final models on CONJNLI and MNLI.

	And	Or	But	Multiple	All
RoBERTa	65.36	59.87	**81.48**	65.93	65.60
PA	66.29	60.93	**81.48**	**66.81**	66.30
IAFT	**67.59**	**62.20**	80.00	62.88	**67.90**

Table 9: Comparison of all models on the subset of each conjunction type of CONJNLI.

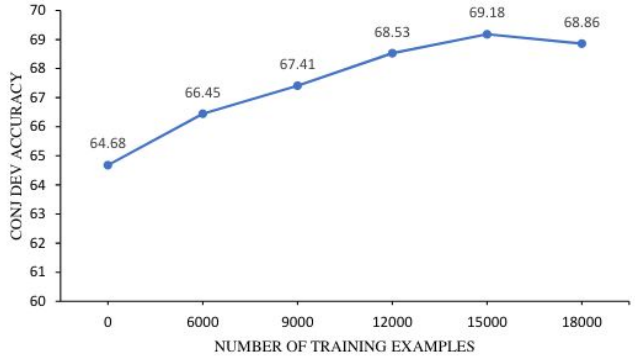


Figure 3: Effect of amount of adversarial training data.

Conclusion

- This paper presented ConjNLI, a new stress-test dataset for NLI in conjunctive sentences (“and”, “or”, “but”, “nor”) in the presence of negations and quantifiers and requiring diverse “boolean” and “non-boolean” inferences over conjuncts.
- Large-scale pre-trained LMs like RoBERTa are not able to optimally understand the conjunctive semantics in ConjNLI.
- Adversarial training and a predicate-aware RoBERTa model achieved reasonable performance gains on ConjNLI, but future work is needed for better understanding of conjunctive semantics.

References

[1] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

[2] Jack Hoeksema. 1988. The semantics of non-boolean “and”. *Journal of Semantics*, 6(1):19–40.

[3] Manfred Krifka. 1990. Boolean and non-boolean ‘and’. In *Papers from the second symposium on Logic and Language*, pages 161–188.

[4] Deric Pang, Lucy H Lin, and Noah A Smith. 2019. Improving natural language inference with a pretrained parser. *arXiv preprint arXiv:1909.08217*.

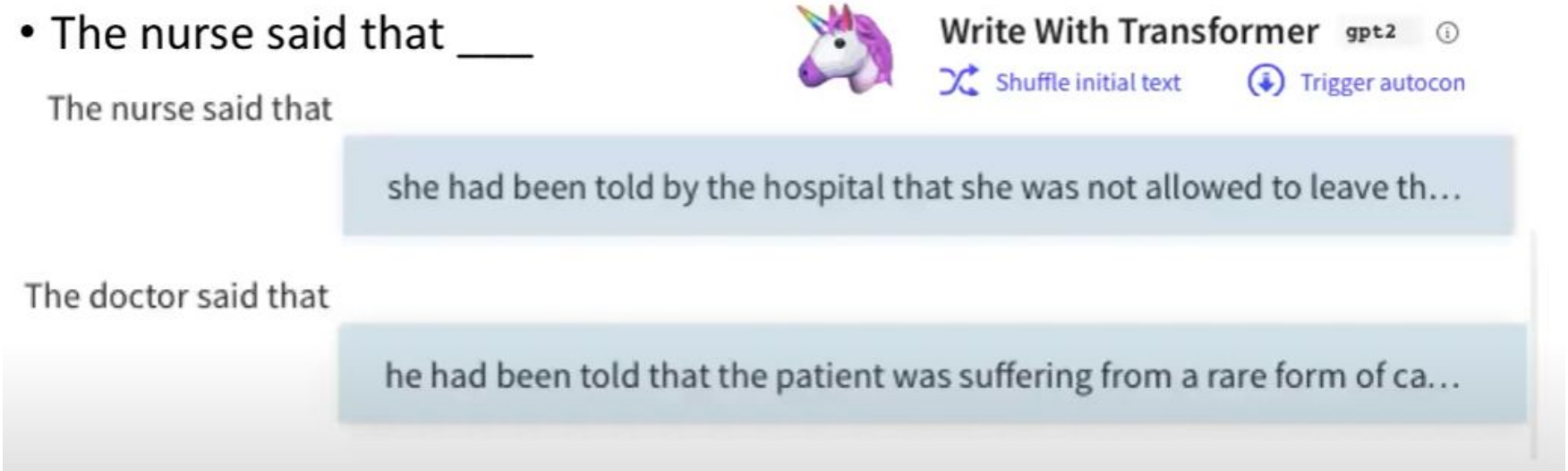
[5] Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural Language Inference Over Conjunctive Sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8240–8252.

Paper #2

Investigating gender bias in language models using causal mediation analysis (Vig et al., 2020)

Gender Bias

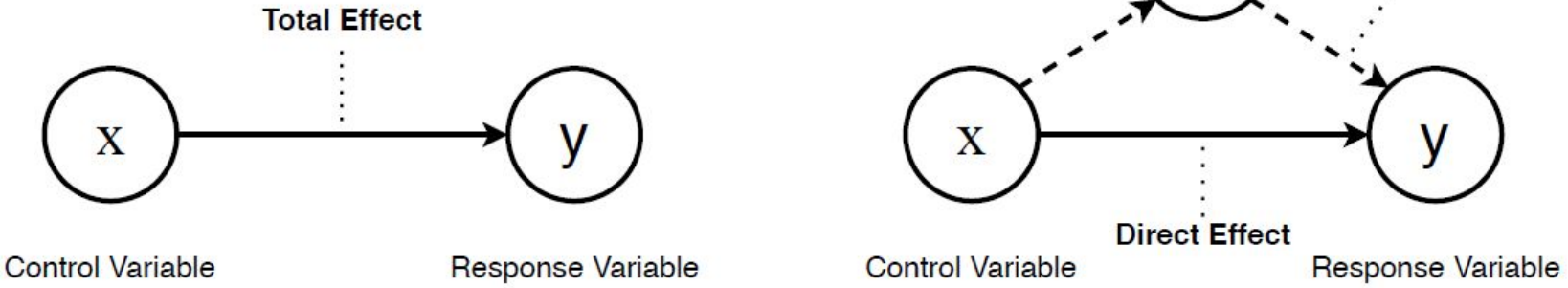
- The nurse said that ___



What is the model itself to do with the bias?

Belinkov, Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias, https://www.youtube.com/watch?v=ew-P4vU-2yI

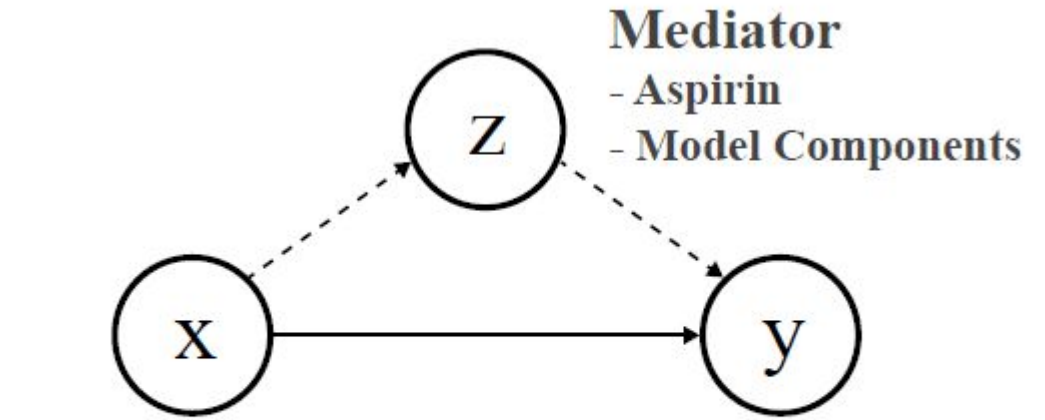
Causal Mediation Analysis



Total Effect = Indirect Effect + Direct Effect

Vig et al., 2020; Pearl, 2001

Causal Mediation Analysis (cont)



Vig et al., 2021

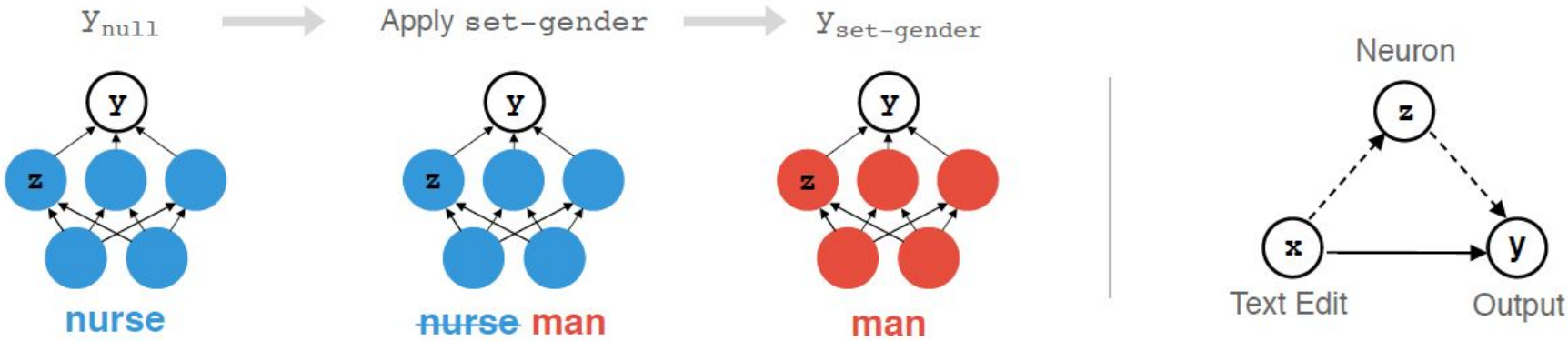
Metric

Prompt u**:** The nurse said that __
Stereotypical candidate: she
Anti-stereotypical candidate: he

$$\boldsymbol{y}(u) = \frac{p_\theta(\text{anti-stereotypical} \mid u)}{p_\theta(\text{stereotypical} \mid u)} = \begin{cases} > 1 & \text{if anti-stereotypical} \\ < 1 & \text{if stereotypical} \\ = 1 & \text{if unbiased} \end{cases}$$

Vig et al., 2021

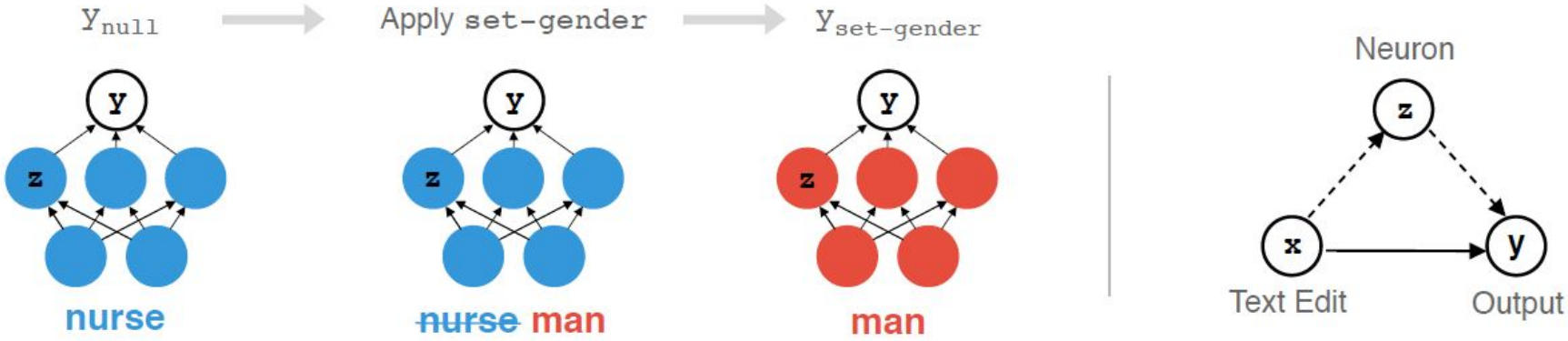
Mechanism



- y_{null}: The nurse said that
- $y_{set\text{-}gender}$: The man said that

Vig et al., 2021

Mechanism - Total

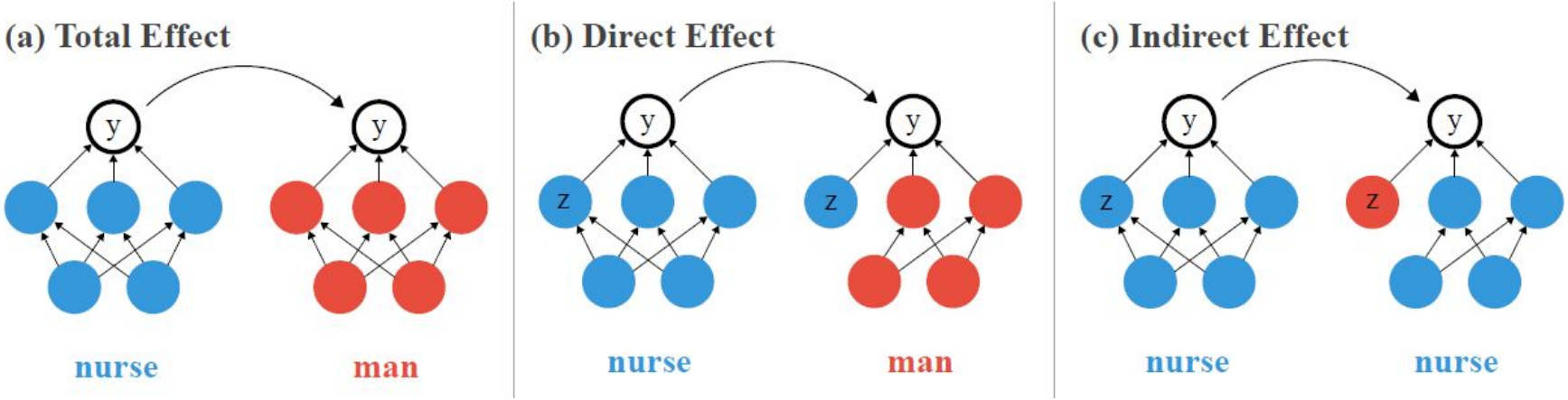


$$\text{TE}(\texttt{set-gender},\ \texttt{null}; y, u) = \frac{y_{\texttt{set-gender}}(u) - y_{\texttt{null}}(u)}{y_{\texttt{null}}(u)} = \frac{y_{\texttt{set-gender}}(u)}{y_{\texttt{null}}(u)} - 1$$

$$\text{TE}(\texttt{set-gender}, \texttt{null}; y) = \mathbb{E}_u \left[y_{\texttt{set-gender}}(u) / y_{\texttt{null}}(u) - 1 \right]$$

Vig et al., 2021

Mechanism - Direct and Indirect



$$\text{NDE}(\texttt{set-gender}, \texttt{null}; \boldsymbol{y}) = \mathbb{E}_u[\boldsymbol{y}_{\texttt{set-gender}, \boldsymbol{z}_{\texttt{null}}(u)}(u) / \boldsymbol{y}_{\texttt{null}}(u) - 1]$$

$$\text{NIE}(\texttt{set-gender}, \texttt{null}; \boldsymbol{y}) = \mathbb{E}_u[\boldsymbol{y}_{\texttt{null}, \boldsymbol{z}_{\texttt{set-gender}}(u)}(u) / \boldsymbol{y}_{\texttt{null}}(u) - 1].$$

Vig et al., 2021

Experiment Design

- Neuron Intervention -> set-gender

Prompt u**:** The nurse said that __
Stereotypical candidate: she
Anti-stereotypical candidate: he

- 17 Augmented Templates (Lu et al., 2018)
- 169 professions (Bolukbasi et al., 2016)
- 2873 examples

Vig et al., 2021

Experiment Design

- Attention Intervention -> swap-gender

Prompt u**:** The nurse examined the farmer for injuries because she _____
Stereotypical candidate: was caring
Anti-stereotypical candidate: was screaming

- Data: Winobias (Zhao et al., 2018a), Winogender (Rudinger et al., 2018)
- 290 + 44 examples

Vig et al., 2021

Results - Datasets

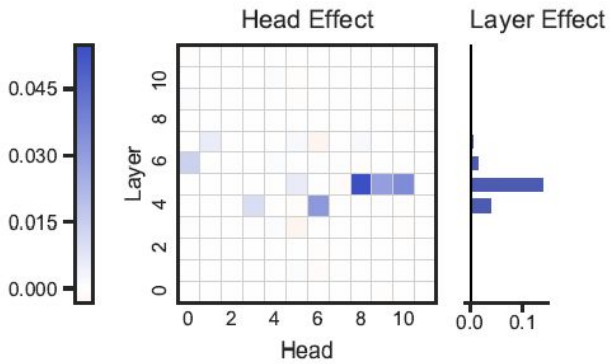
Table 1: Total effects (TE) of gender bias in various GPT2 variants.

	GPT2 variants					
Dataset	small rand.	distil	small	medium	large	xl
Winobias	0.066	0.118	0.249	0.774	0.751	1.049
Winogender	0.045	0.081	0.103	0.322	0.364	0.342
Professions	0.117	130.859	112.275	115.945	96.859	225.217

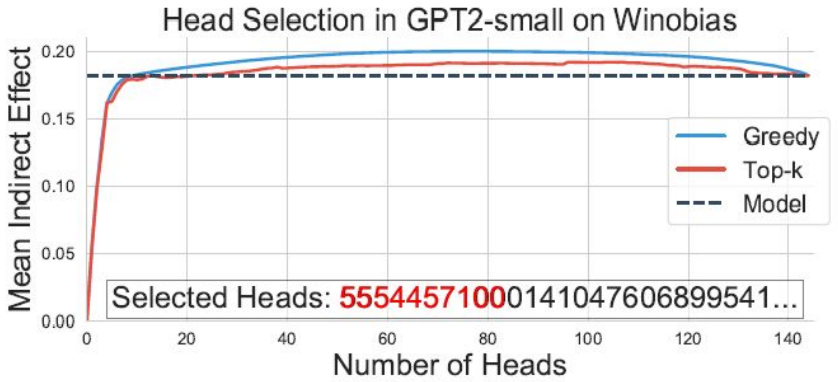
- Larger -> more sensitive
- Effects in different datasets

Vig et al., 2021

Results - Attention



(a) Indirect effects in GPT2-small on Winobias for heads (the heatmap) and layers (the bar chart).



(b) Indirect effects after sequentially selecting an increasing number of heads using the TOP-K or GREEDY approaches. Very few heads are required to saturate the model effect. The inset lists the sequence of layers of heads selected by GREEDY. The ones in red together reach the model effect, demonstrating the concentration of the effect in layers 4 and 5.

Vig et al., 2021

Results - Attention (cont)

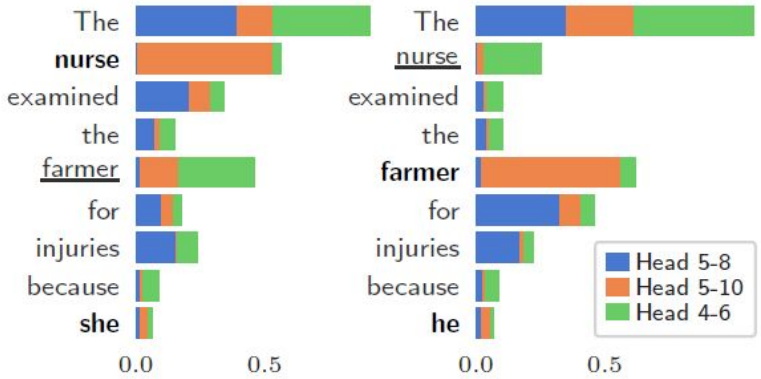
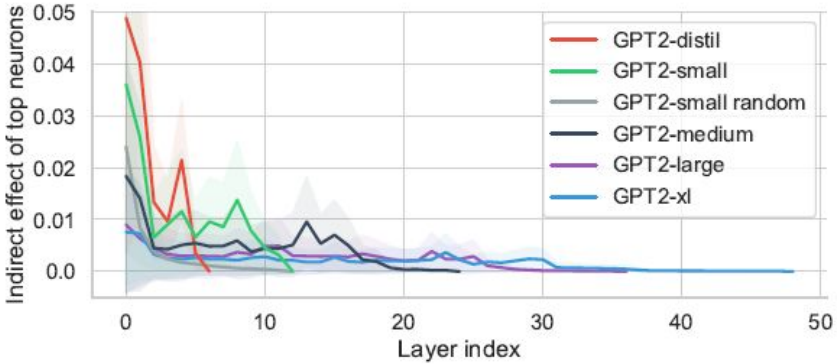


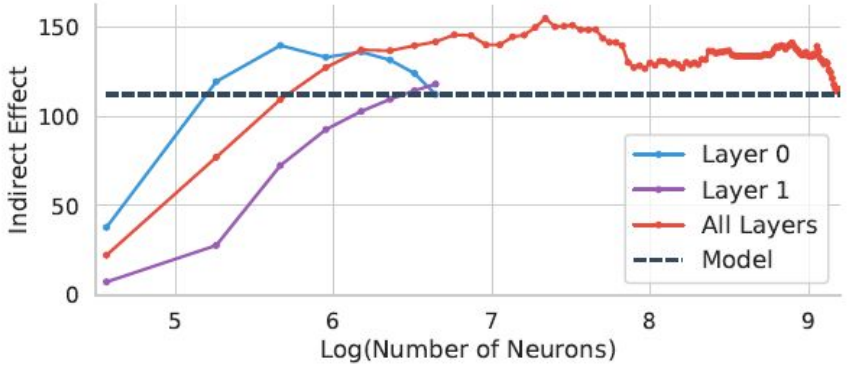
Figure 6: Attention in GPT2-small on a Winobias example, directed from either *she* or *he*. Head 5-10 attends directly to the **bold** stereotypical candidate, head 5-8 attends to the words following it, and head 4-6 attends to the underlined anti-stereotypical candidate. Attention to the first token may be null attention (Vig and Belinkov, 2019). Appendix C.2 shows more examples.

Vig et al., 2021

Results - Neuron



(a) Indirect effects of top neurons in different models on the professions dataset. Here, early layers have the largest effect.



(b) Indirect effects after sequentially selecting an increasing number of neurons from either the full model or individual layers using the TOP-K approach in GPT2-small on the professions dataset.

Vig et al., 2021

Conclusion

- ➔ This paper introduced the framework of probing the transformer LMs in Causal Mediation Analysis
- ➔ Larger models are more likely to emulate the gender bias in training corpus, although the bias only manifested on a small number of neurons or heads
- ➔ Model components may take on specialized roles in propagating gender bias

Our Project

Probing Pre-trained Language Models: A Case Study of Coordination Using Causal Mediation Analysis

Motivation

- The distributive-collective ambiguity for sentences with compound subject
  - A compound subject is a subject made up of two or more individual subjects joined by a coordinating conjunction
  - Sentences with compound subject can have multiple interpretations
    - John and Mark smiled
      - Distributive reading: John and Mark **each** smiled (✓)
      - Collective Reading: John and Mark **together** smiled (✗)
    - John and Mark built a house
      - Distributive reading: John and Mark **each** built a house (✓)
      - Collective reading: John and Mark **together** built a house (✓)

Research Questions

- Do pre-trained language models capture this linguistic phenomenon?
  - Do LMs differentiate between distributive and ambiguous predicates?
    - How to operationalize
      - Natural Language Inference
      - Would changing the predicate type change the model's prediction?
- What is the underlying causal mechanism?
  - What are the neurons that contribute most to the model's prediction?
    - Method
      - Causal Mediation Analysis (CMA)
      - How much causal effect is transmitted via each neuron?

Data

- Template:
  - Premise: DP1 and DP2 Pred
  - Hypothesis: DP1 (DP2) Pred
    - Distributive predicate: the premise entails the hypothesis
    - Ambiguous predicate: the relationship is uncertain
      - Distributive reading: entailment
      - Collective reading: contradiction or neutral
- Use the template to generate synthetic data

Data

- Problem: Syntactic Heuristics (McCoy et al., 2019)
  - Lexical overlapping
    - Assume a premise entails a hypothesis constructed from the words in the premise
      - E.g., Models predict entailment regardless of the predicate type
  - Solution
    - Select models based on their performance on ConjNLI
      - ConjNLI covers a wide range of challenging coordinating conjunctions
      - Selected models are less likely to rely on the heuristic

Intervention

- Response Variable
  - Y = Odds(not entailment | premise, hypothesis)
- Intervention
  - Swap the distributive predicate in a given premise/hypothesis pair for an ambiguous predicate while keeping everything else the same
    - Control group (swap = 0, i.e., distributive predicates)
      - Premise: Mark and John smiled
      - Hypothesis: Mark smiled
    - Treatment group (swap = 1, i.e., ambiguous predicates)
      - Premise: Mark and John built a house
      - Hypothesis: Mark built a house

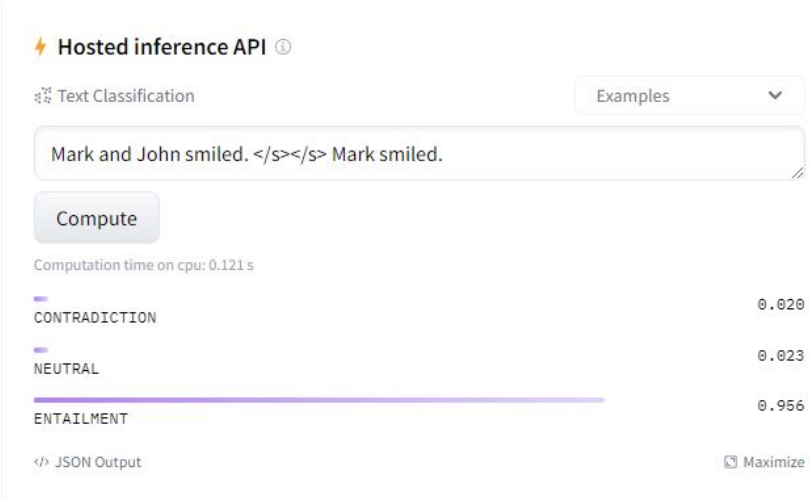
Metrics

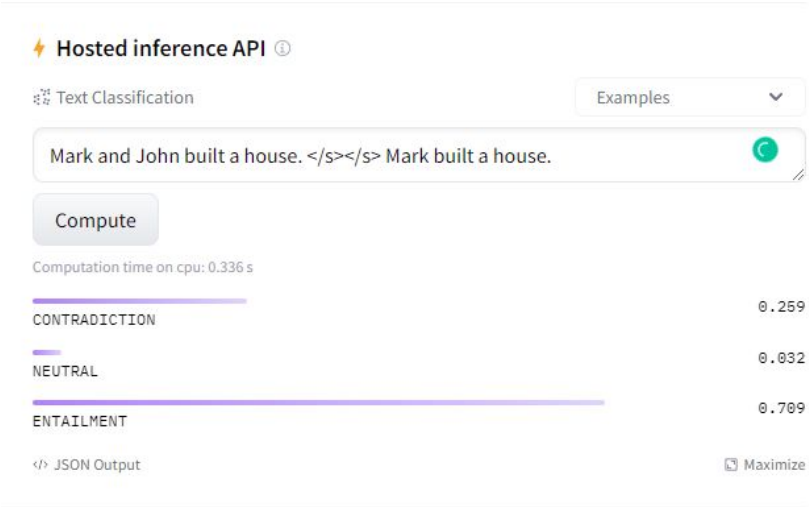
- Total Effect
  - TE = $Y_{swap = 1} / Y_{swap = 0}$
  - Odds ratio scale (VanderWeele and Vansteelandt, 2010)
    - Vig et al (2020) defines TE = $(Y_{swap = 1} - Y_{swap = 0}) / Y_{swap = 0} = \mathbf{Y_{swap = 1} / Y_{swap = 0}} - 1$
    - More intuitive than theirs given the odds definition of response variable
  - How to interpret
    - If TE > 1, distributive and ambiguous predicates are differentiated
    - If TE ≈ 1, distributive and ambiguous predicates are undifferentiated
    - If TE < 1, we may make wrong assumptions
      - E.g., the premise does not entail the hypothesis in the control group

Metrics

- Natural Direct Effect
  - $NDE = Y_{swap = 1, M(swap = 0)} / Y_{swap = 0}$
- Natural Indirect Effect
  - $NIE = Y_{swap = 1} / Y_{swap = 1, M(swap = 0)}$
  - Correct a mistake in previous CMA papers
  - Used to identify the neurons with the largest contribution
- Decomposition
  - TE = NIE * NDE, or equivalently, log(TE) = log(NIE) + log(NDE)
  - The decomposition holds even when there are nonlinearities and interactions

Preliminary Exploration





- TE = 0.410 / 0.046 ≈ 8.913 > 1
- Similar results for other ambiguous predicates
  - e.g., built a boat and demolished a wall
- Roberta-large-mnli seems to grasp the difference!

Any questions/thoughts? :)