CSCE 689: Special Topics in Trustworthy NLP

Lecture 19: Model Explainability

Kuan-Hao Huang
khhuang@tamu.edu

Guest Lecture (Online)

- **Time:** November 24, 4:10pm-5:25pm
- **Title:** From Risk to Resilience: Addressing Misalignment in (Multimodal) Large Language Models
- **Speaker:** Fei Wang
- **Zoom Link:** https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPApqDbJnVtktdW2AE6nb.1

Guest Lecture (Online)

Abstract: As (multimodal) large language models (LLMs) become central to intelligent systems, their use is expanding from everyday applications to high-stakes domains. Alignment plays a crucial role in the successful development of LLMs, ensuring that model behavior matches our expectations and remains consistent with various objectives. However, misalignment persists as a significant challenge that undermines the trustworthiness and reliability of these models. This talk will explore methods to tackle the misalignment problem by addressing three key research questions: (1) How to mitigate the risk of a misaligned LLM with only limited model accessibility? (2) How to ensure a reliable alignment process in multimodal scenarios? (3) How to integrate missing or customized alignment objectives to achieve precise control over model behavior in diverse contexts? Particularly, this talk will systematically address these challenges with resilient retrieval-augmented generation, conditional alignment, and constraint integration. In addition, this talk will shed light on the responsible development of LLMs across various scenarios and interdisciplinary contexts.

Guest Lecture (Online)

Speaker Bio:

Fei Wang is a Research Scientist at Google. Previously, he obtained his Ph.D. from University of Southern California. Fei's research focuses on developing post-training methods for robust and reliable LLMs and multimodal LLMs. His work has been recognized with an Amazon ML PhD Fellowship and an Annenberg PhD Fellowship. Additionally, he has instructed tutorials and served as area chairs at top tier NLP and ML conferences, including EMNLP, NAACL, ACL, and NeurIPS.

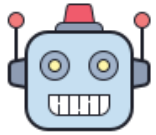
Course Project – Computations

- HPRC (https://hprc.tamu.edu/resources/)
  - FASTER: A100 GPUs, A10 GPUs, A30 GPUs, A40 GPUs and T4 GPUs
  - GRACE: A100 GPUs, RTX 6000 GPUs, T4 GPUs, and A40 GPUs

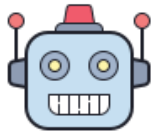
Model Explainability and Interpretability

Hello! Could you help me reserve a table at the *"The Best"* restaurant for tomorrow at 12pm?

Of course! I've reserved a table at the *"The Best"* restaurant for tomorrow at 12pm.

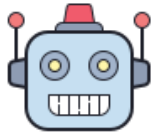


I generate this response is because I saw you mention reserve, one restaurant name, and one specific time. Therefore…

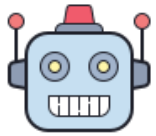


Hello! Could you help me reserve a table at the *"The Best"* restaurant for tomorrow at 12pm?

Of course! I've reserved a table at the *"The Best"* restaurant for tomorrow at 12pm.



I generate this response is because I saw you mention tomorrow. It is usually strongly related to restaurant reservation.



Provide additional information to decide if we should trust the answers

Good Explanations Should Be Faithful

- A faithful interpretation is one that accurately represents the reasoning process behind the model's prediction

Good Explanations Should Be Plausible

- An explanation is considered plausible if it is coherent with human reasoning and understanding
- Plausibility is also referred to as persuasiveness or understandability
- An explanation might be plausible but not faithful. Currently, many explanations are more plausible than faithful
- Example of faithful, but not plausible explanation: a copy of model weights

Good Explanations Should Be Informative

Hi prof, I have just finished this paper. Which venue do you think would best suit it?

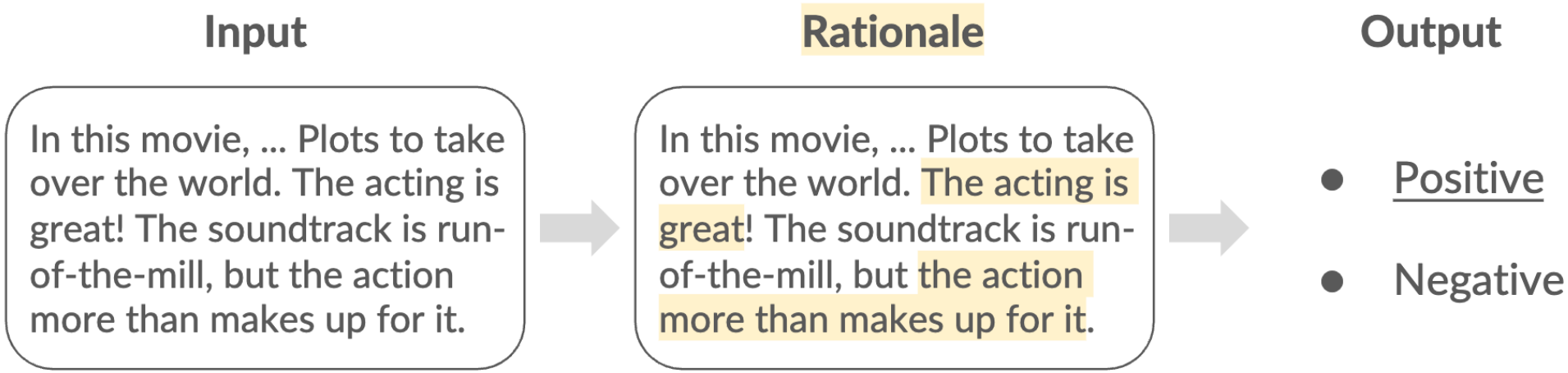
NAACL, because its deadline is just 3 days away, and it will be in Mexico, not far from here.

NAACL, because it is a top NLP conference.

Which explanation is more informative?

Extractive Rationales

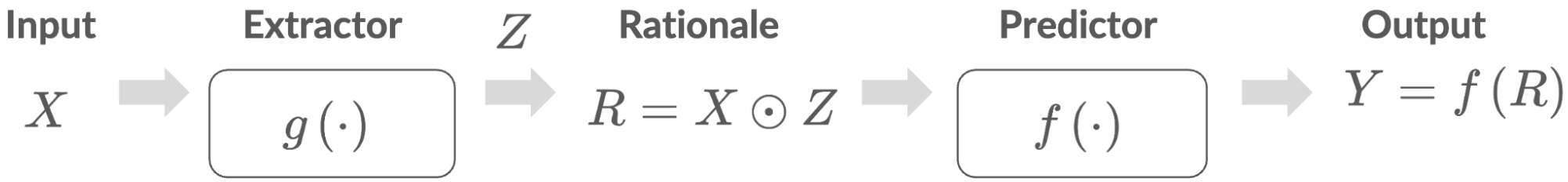
- Rationales: short snippets in inputs that support outputs



Rationalizing Neural Predictions, EMNLP 2016

Extractive Rationales

- Pipeline model



Model $P(z|x) = g(x)$

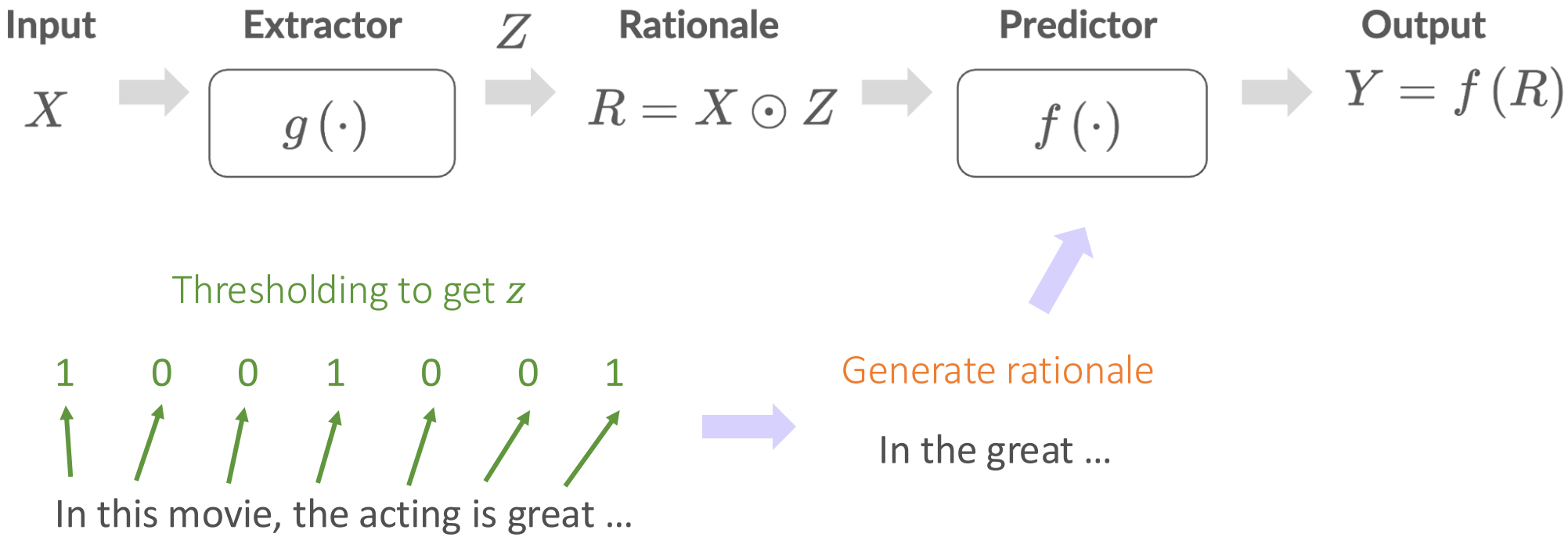
0.51 0.12 0.87 0.66 0.43 0.22 0.95

In this movie, the acting is great …

Rationalizing Neural Predictions, EMNLP 2016

Extractive Rationales

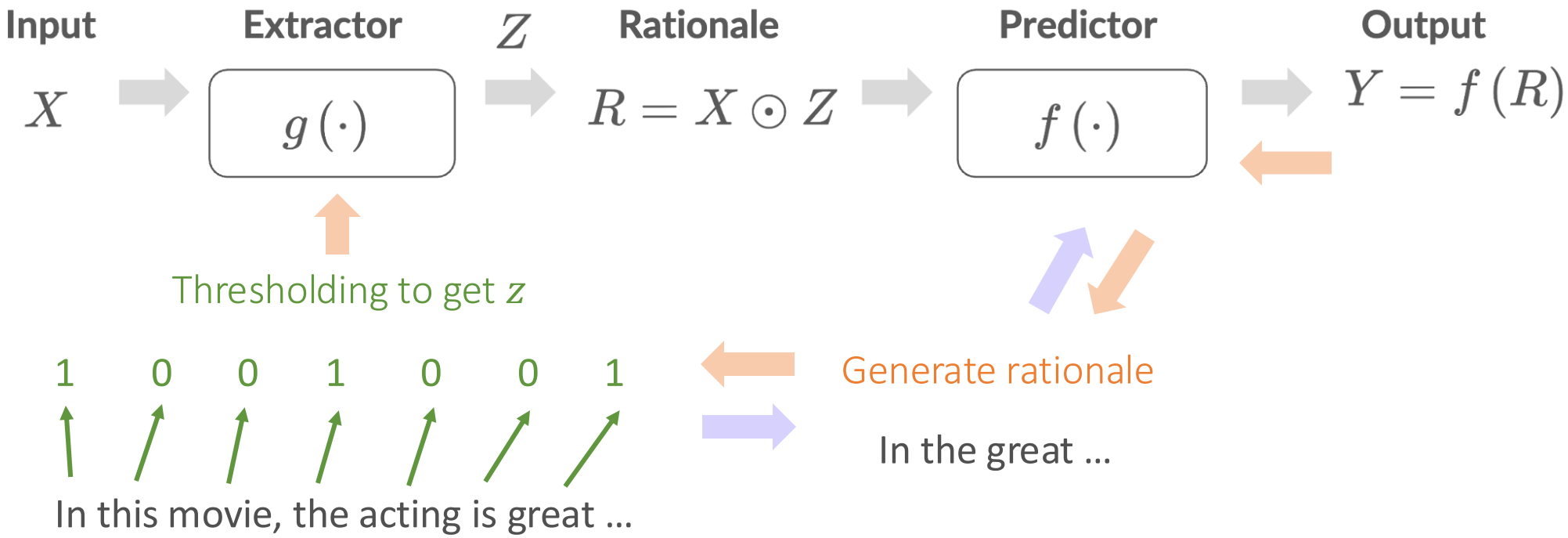
- Pipeline model



Rationalizing Neural Predictions, EMNLP 2016

Extractive Rationales

- Pipeline model



Rationalizing Neural Predictions, EMNLP 2016

Examples

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

---

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

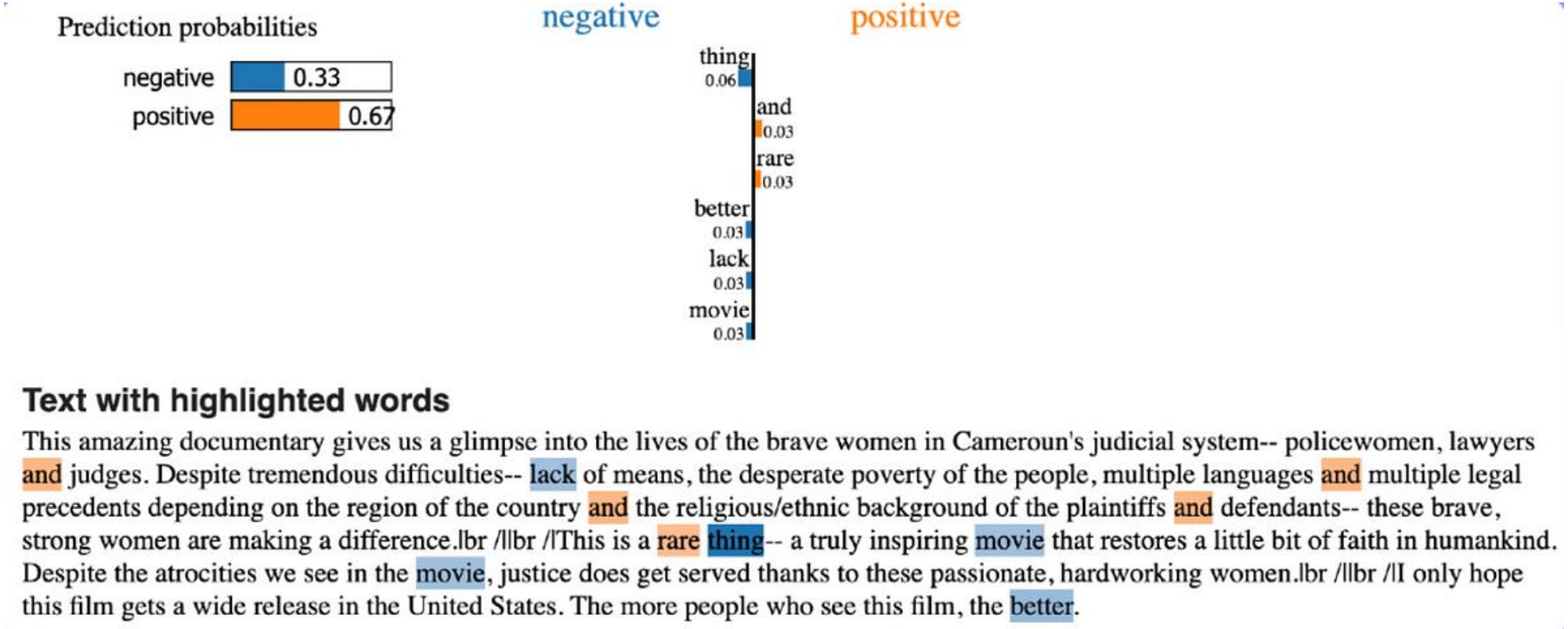
---

i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

---

a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate … … m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

LIME: Explanations for Any Black-Box Models



Text with highlighted words

This amazing documentary gives us a glimpse into the lives of the brave women in Cameroun's judicial system-- policewomen, lawyers and judges. Despite tremendous difficulties-- lack of means, the desperate poverty of the people, multiple languages and multiple legal precedents depending on the region of the country and the religious/ethnic background of the plaintiffs and defendants-- these brave, strong women are making a difference.lbr /llbr /lThis is a rare thing-- a truly inspiring movie that restores a little bit of faith in humankind. Despite the atrocities we see in the movie, justice does get served thanks to these passionate, hardworking women.lbr /llbr /lI only hope this film gets a wide release in the United States. The more people who see this film, the better.

"Why Should I Trust You?": Explaining the Predictions of Any Classifier, NAACL 2016

LIME: Local Interpretable Model-agnostic Explanations

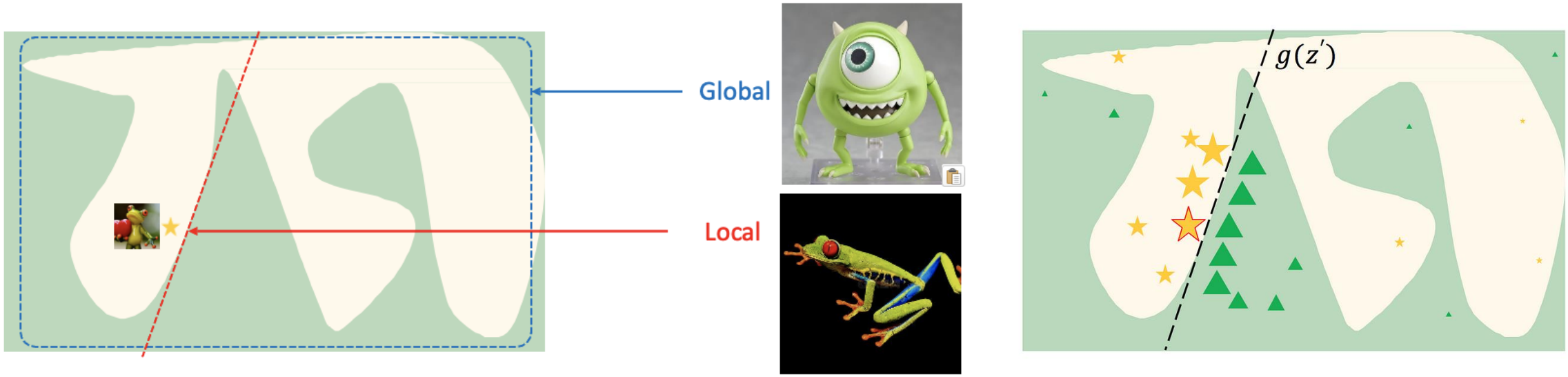
- Analysis model f
- Train a local interpretable model based on f and perturbed examples
- For one example, get prediction from f
  - "The storyline is boring, but the actors are great." → Positive (0.76)
- Perturb examples
  - "The storyline is boring, but the actors are [mask]." → Negative (0.35)
  - "The storyline is [mask], but the actors are great." → Positive (0.85)
  - "The storyline is boring, but the [mask] are great." → Positive (0.70)
  - "The [mask] is boring, but the actors are great." → Negative (0.48)

LIME

- New training examples for local interpretable model
  - “The storyline is boring, but the actors are great. → Positive (0.76)
  - “The storyline is boring, but the actors are [mask]. → Negative (0.35)
  - “The storyline is [mask], but the actors are great. → Positive (0.85)
  - “The storyline is boring, but the [mask] are great. → Positive (0.70)
  - “The [mask] is boring, but the actors are great. → Negative (0.48)
- Train a linear model to approximate the decision boundary
  - Text feature: bag-of-word, TF-IDF, n-gram, …
- The linear weights can be explanations
  - great (+2.7), boring (-3.6), but (+0.6), …

Local Faithfulness

- Train a surrogate model (interpretable model) to locally approximate the boundary



https://medium.com/sherry-ai/23898753bea5

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Experiments

- Correlation between attention-based and gradient-based/leave-one-out

		Gradient (BiLSTM) τ_g		Gradient (Average) τ_g		Leave-One-Out (BiLSTM) τ_{loo}	
Dataset	Class	Mean ± Std.	Sig. Frac.	Mean ± Std.	Sig. Frac.	Mean ± Std.	Sig. Frac.
SST	0	0.34 ± 0.21	0.48	0.61 ± 0.20	0.87	0.27 ± 0.19	0.33
	1	0.36 ± 0.21	0.49	0.60 ± 0.21	0.83	0.32 ± 0.19	0.40
IMDB	0	0.44 ± 0.06	1.00	0.67 ± 0.05	1.00	0.34 ± 0.07	1.00
	1	0.43 ± 0.06	1.00	0.68 ± 0.05	1.00	0.34 ± 0.07	0.99
ADR Tweets	0	0.47 ± 0.18	0.76	0.73 ± 0.13	0.96	0.29 ± 0.20	0.44
	1	0.49 ± 0.15	0.85	0.72 ± 0.12	0.97	0.44 ± 0.16	0.74
20News	0	0.07 ± 0.17	0.37	0.79 ± 0.07	1.00	0.06 ± 0.15	0.29
	1	0.21 ± 0.22	0.61	0.75 ± 0.08	1.00	0.20 ± 0.20	0.62
AG News	0	0.36 ± 0.13	0.82	0.78 ± 0.07	1.00	0.30 ± 0.13	0.69
	1	0.42 ± 0.13	0.90	0.76 ± 0.07	1.00	0.43 ± 0.14	0.91
Diabetes	0	0.42 ± 0.05	1.00	0.75 ± 0.02	1.00	0.41 ± 0.05	1.00
	1	0.40 ± 0.05	1.00	0.75 ± 0.02	1.00	0.45 ± 0.05	1.00
Anemia	0	0.47 ± 0.05	1.00	0.77 ± 0.02	1.00	0.46 ± 0.05	1.00
	1	0.46 ± 0.06	1.00	0.77 ± 0.03	1.00	0.47 ± 0.06	1.00
CNN	Overall	0.24 ± 0.07	0.99	0.50 ± 0.10	1.00	0.20 ± 0.07	0.98
bAbI 1	Overall	0.25 ± 0.16	0.55	0.72 ± 0.12	0.99	0.16 ± 0.14	0.28
bAbI 2	Overall	-0.02 ± 0.14	0.27	0.68 ± 0.06	1.00	-0.01 ± 0.13	0.27
bAbI 3	Overall	0.24 ± 0.11	0.87	0.61 ± 0.13	1.00	0.26 ± 0.10	0.89
SNLI	0	0.31 ± 0.23	0.36	0.59 ± 0.18	0.80	0.16 ± 0.26	0.20
	1	0.33 ± 0.21	0.38	0.58 ± 0.19	0.80	0.36 ± 0.19	0.44
	2	0.31 ± 0.21	0.36	0.57 ± 0.19	0.80	0.34 ± 0.20	0.40

Experiments

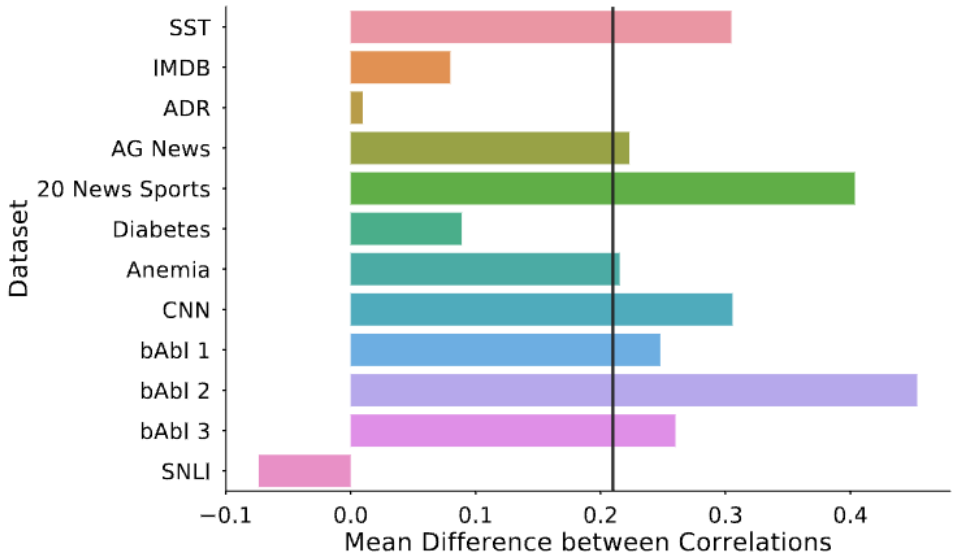


Figure 6: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. LOO scores using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by $>0.2\ \tau_{loo}$.

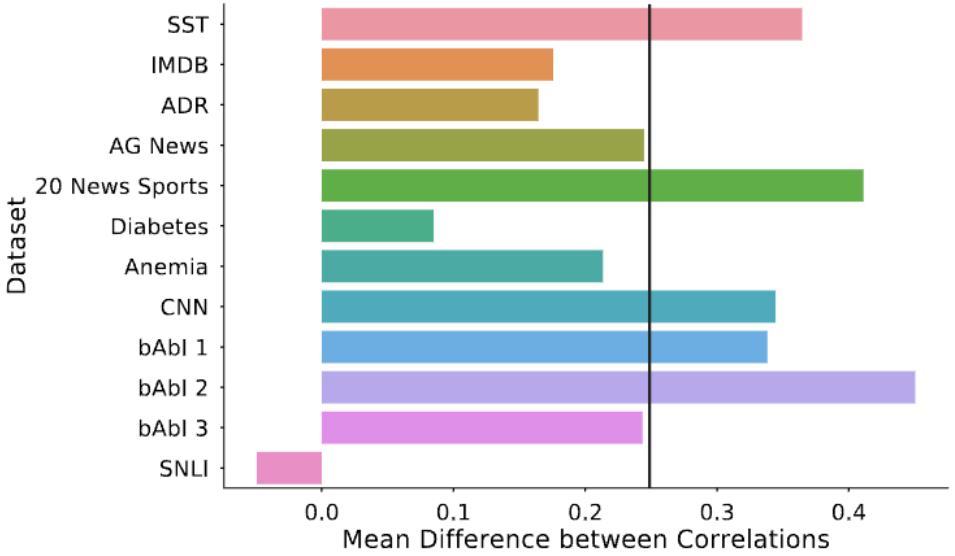
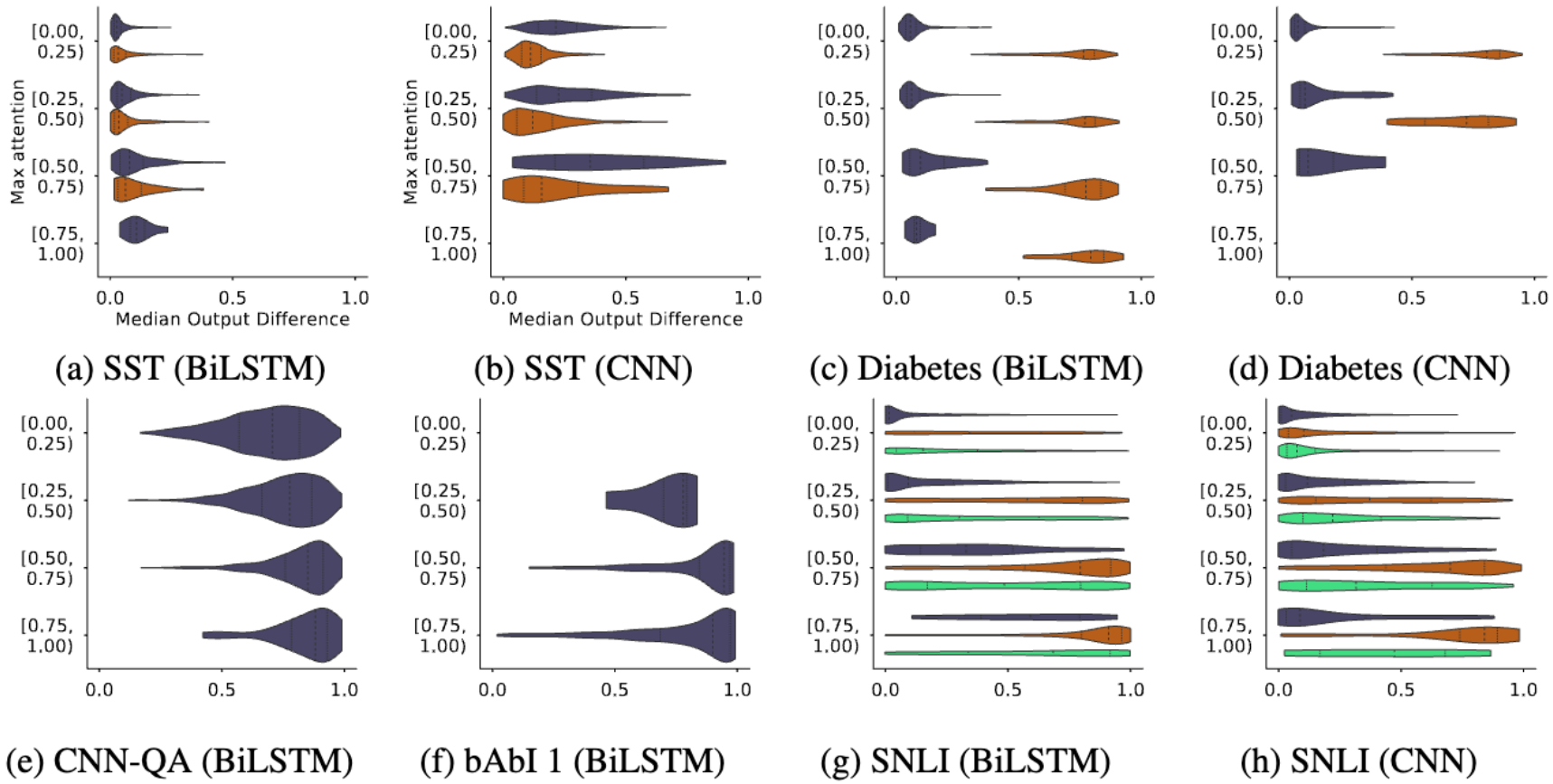


Figure 7: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. Gradients using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by $\sim 0.25\ \tau_g$.

Permutate Attention Weights



(a) SST (BiLSTM) (b) SST (CNN) (c) Diabetes (BiLSTM) (d) Diabetes (CNN)

(e) CNN-QA (BiLSTM) (f) bAbI 1 (BiLSTM) (g) SNLI (BiLSTM) (h) SNLI (CNN)

Adversarial Attention Weights

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

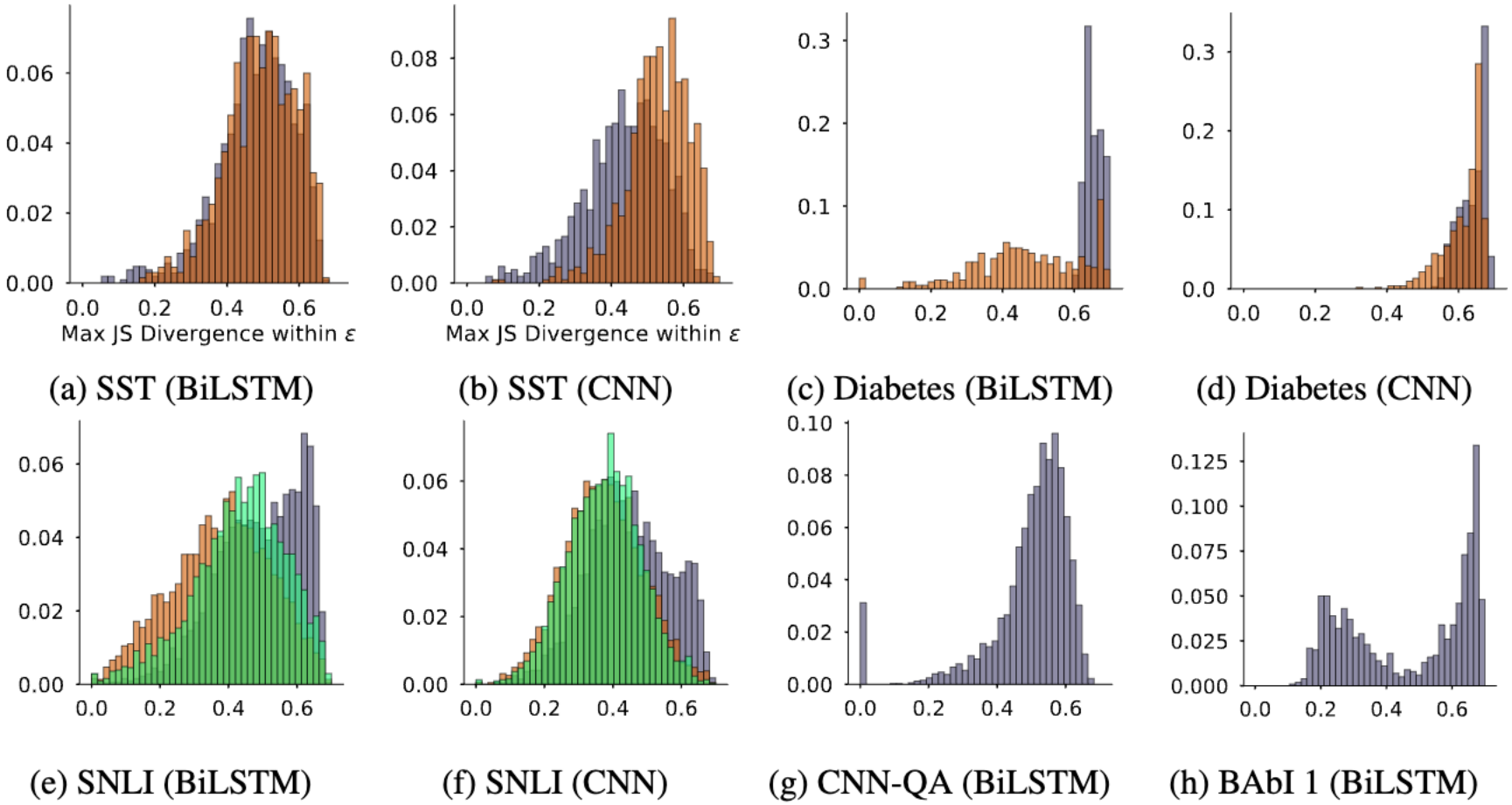
original α

$$f(x|\alpha, \theta) = 0.01$$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

adversarial $\tilde{\alpha}$

$$f(x|\tilde{\alpha}, \theta) = 0.01$$



(a) SST (BiLSTM) (b) SST (CNN) (c) Diabetes (BiLSTM) (d) Diabetes (CNN)

(e) SNLI (BiLSTM) (f) SNLI (CNN) (g) CNN-QA (BiLSTM) (h) BAbI 1 (BiLSTM)

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Uniform Attentions

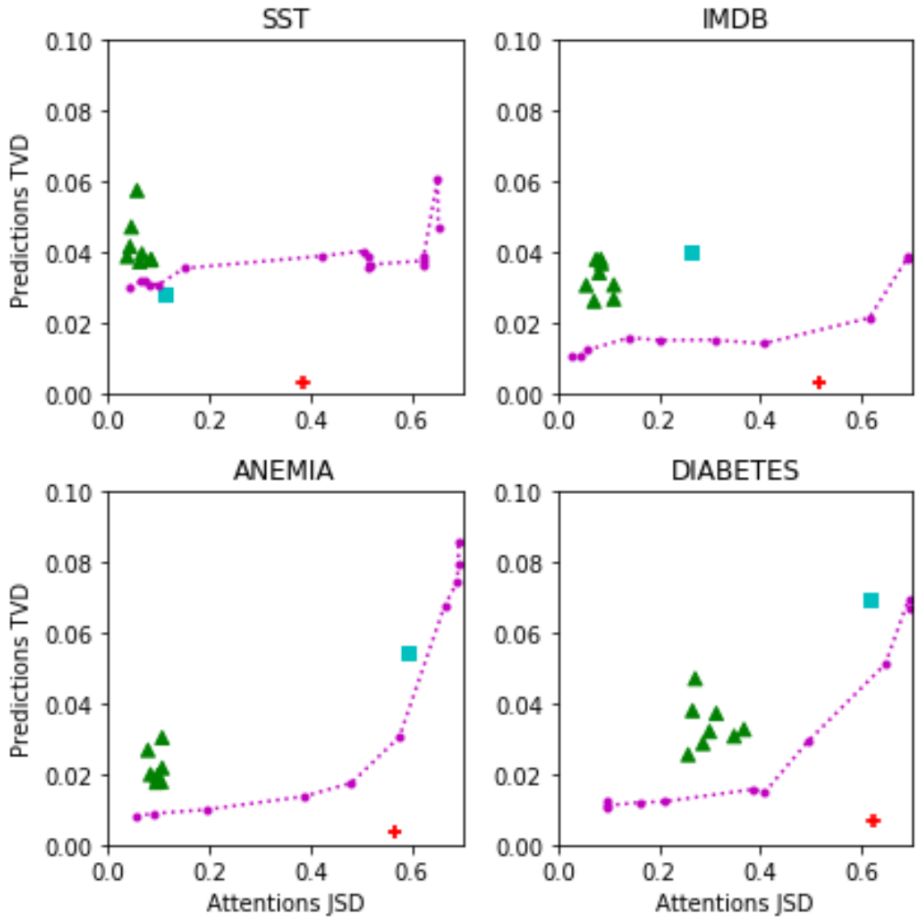
- If attention models are not useful compared to very simple baselines, there is no point in using their outcomes for any type of explanations

Dataset	Attention (Base)		Uniform
	Reported	Reproduced	
Diabetes	0.79	0.775	0.706
Anemia	0.92	0.938	0.899
IMDb	0.88	0.902	0.879
SST	0.81	0.831	0.822
AgNews	0.96	0.964	0.960
20News	0.94	0.942	0.934

Training an Adversary

- Attention distribution is not a primitive
  - We need to re-train for adversarial attention weights

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \text{TVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \lambda \ \text{KL}(\boldsymbol{\alpha}_a^{(\mathbf{i})} \parallel \boldsymbol{\alpha}_b^{(\mathbf{i})})$$



Takeaways

- Is attention good explanations?

Personal Thoughts