Explanations for Natural Language Processing

Feb 7, 2020 @ CogScale. By: Diego Garcia-Olano

diegoolano.com

1. Explainable AI (XAI)

2. XAI for NLP

3. Generating Black Box Counterfactuals using Reinforcement Learning (preliminary work)

1. Explainable AI (XAI)

The higher the interpretability/explainability of a model*, the easier it is for someone to comprehend why certain decisions or predictions have been made.

1. Explainable AI (XAI)

The higher the interpretability/explainability of a model*, the easier it is for someone to comprehend why certain decisions or predictions have been made.

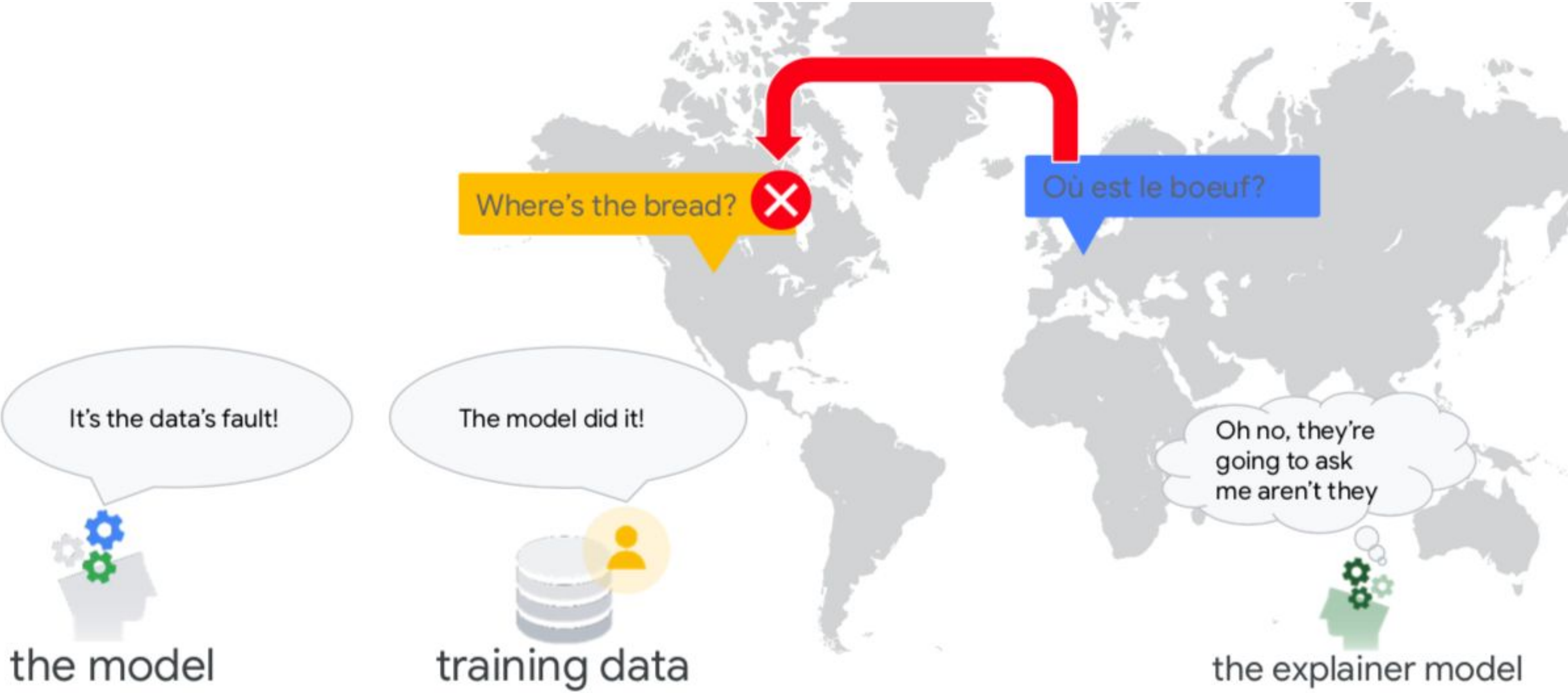
Implications for
fairness,
accountability,
transparency of AI systems

**Usually "model" here means something "deep"/non-linear where feature weights/coefficients are not immediately understandable to a human.*

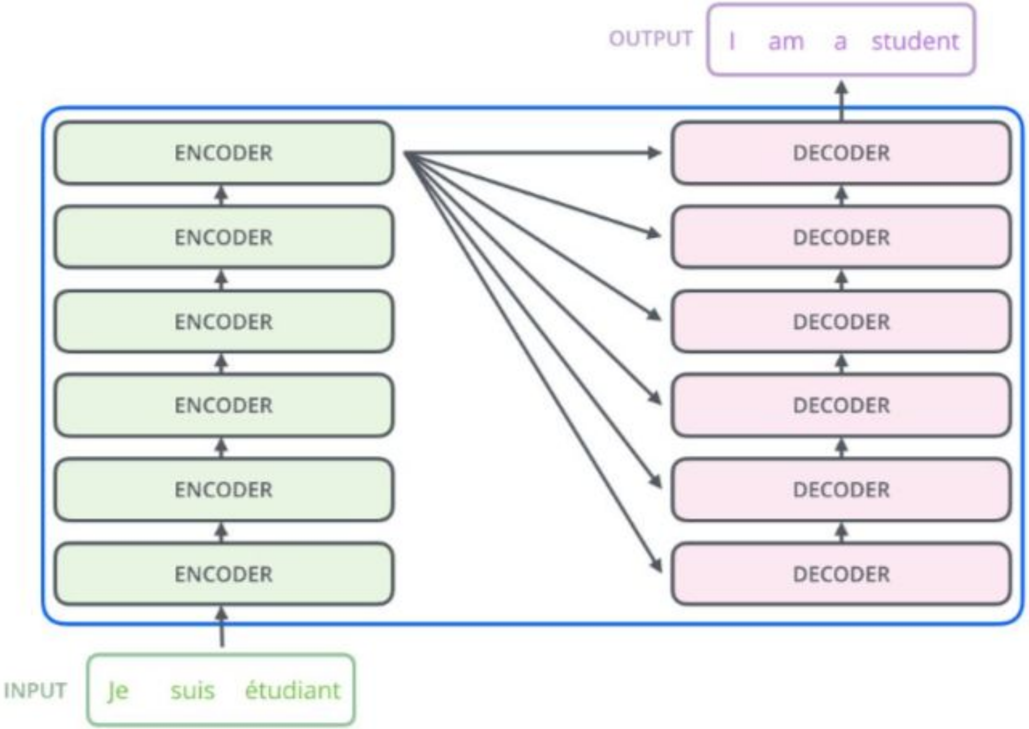
AAAI 2019 Tutorial: On Explainable AI
Interpretable ML book
Dr. Ghosh's Graduate Seminar on Responsible AI (Spring 2019)

Example: Language translation

Example: Language translation

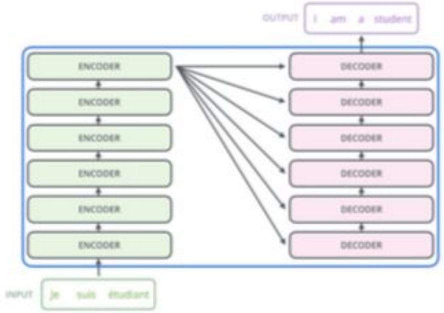


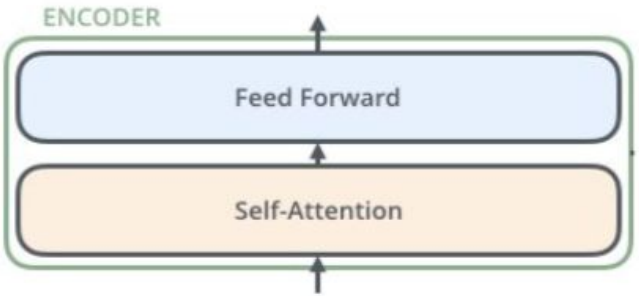
Let's take a look at that Transformer model

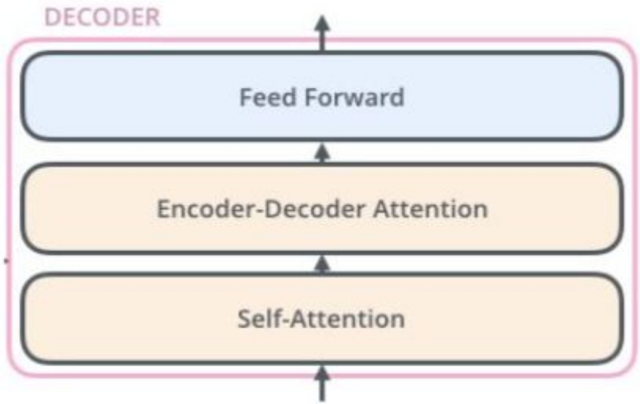


Vaswani, et al., 2017
http://jalammar.github.io/illustrated-transformer/

Let’s take a look closer...







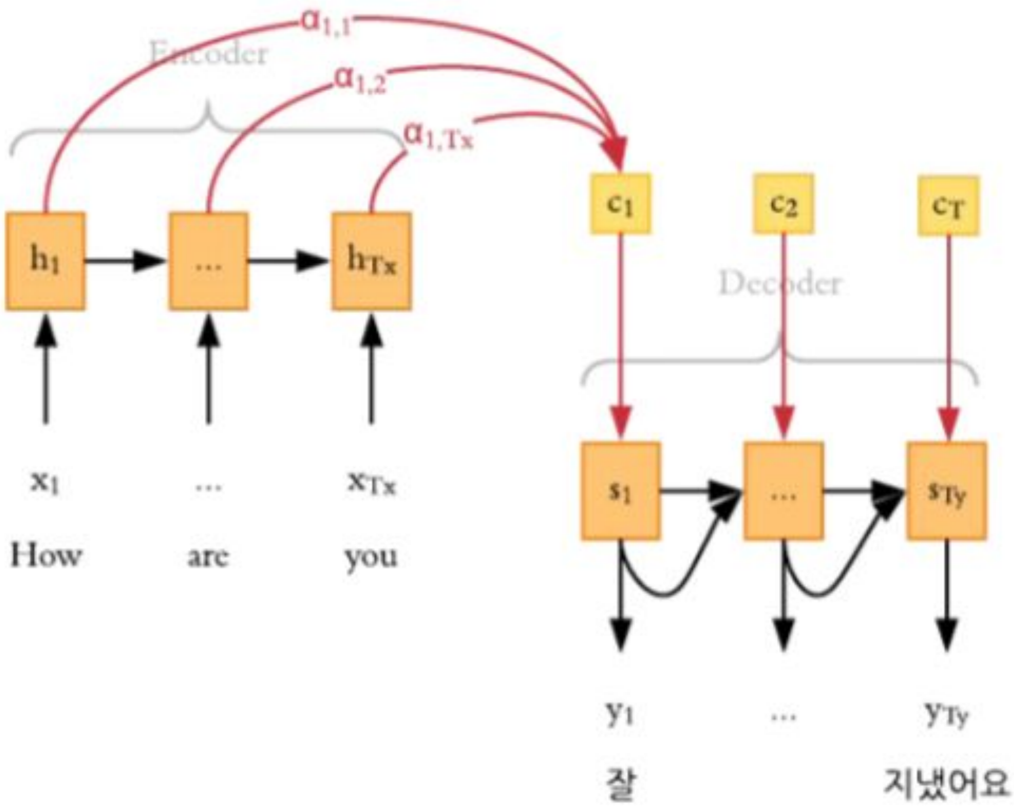
3 types of attention mechanisms

1. encoder self-attention
2. decoder self attention
3. encoder-decoder attention

Each of these is “Multi-headed” (ie, **8 attention heads** run independently in parallel whose outputs are concatenated and linearly transformed into the expected dimensions.

http://jalammar.github.io/illustrated-transformer/

Let's take a look at attention...



$$\begin{aligned} \mathbf{c}_t &= \sum_{i=1}^{n} \alpha_{t,i} \boldsymbol{h}_i \\ \alpha_{t,i} &= \text{align}(y_t, x_i) \\ &= \frac{\exp(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_i))}{\sum_{i'=1}^{n} \exp(\text{score}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_{i'}))} \end{aligned}$$

$$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$$

where both $\mathbf{v}_a$ and $\mathbf{W}_a$ are weight matrices to be learned

Bahdanau et al, 2015

https://medium.com/@joealato/attention-in-nlp-734c6fa9d983

https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html **

So how do we “explain” that?

- **Who** are we explaining to:

  An end user? Model developers?

- **White Box vs Black Box:**

  Do we have access to the model *and/or* the data it was trained on?

- **From where in process :** Pre-model, In-Model or Post Hoc explanations

  Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead (Rudin, et al, 2019 Nature)

- **Global** model vs **Individual** instance based explanations

Some Types of Explanations

Feature Attribution: which features contributed most for a model's output

- Path Integrated Gradients (IG)
- Shapley Additive Explanations (SHAP)
- Contrastive Explanations with Pertinent Negatives (link)

Influential examples: which training data most influenced a model's output

- Influence Functions (link)
- Representer Point Selection for Explaining Deep Neural Networks (link)

Counterfactuals: minimal change that would have led to a different output

Prototypes: find "prototypical" examples as a global summarization

- Deep Learning for Case-Based Reasoning through Prototypes (link)
- Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust DL (link)

Model Distillation:

- Auditing Black-Box Models Using Transparent Model Distillation (link)

2. XAI for NLP

Common tasks: Sentiment Analysis, QA, Text Generation, Style Transfer, Translation

XAI for NLP tends to be *very task dependent*

Considerations:

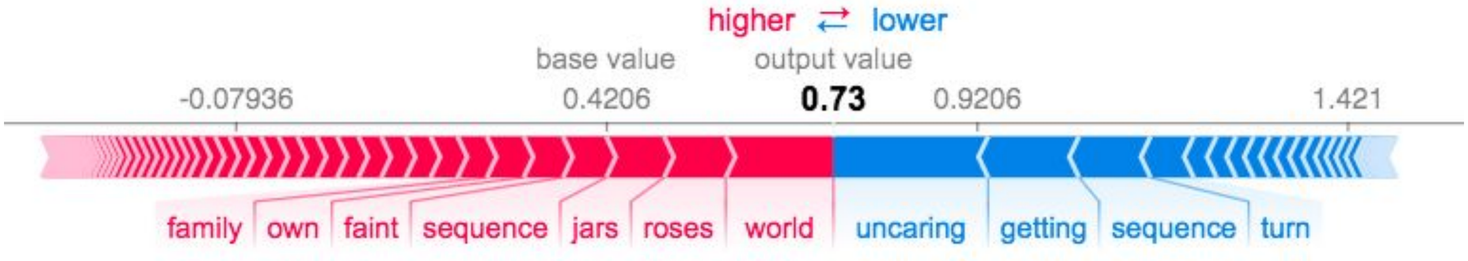
- Syntax, semantic meaning, factual correctness, coherence, etc
- Attention is/is not attribution
- Probing for linguistic meaning of embeddings and models
- Evaluation metrics (BLEU, ROUGE, BertScore, Human Eval)

Analyzing and interpreting neural networks for NLP (workshop at EMNLP)

2. XAI for NLP

- General XAI methods mostly used for classification tasks

SHAP for feature attribution (feature correlation can be an issue)



(ref1 & ref2)

Lai, et al 2019: explore attention/lime/shap over multiple models for text classification

2. XAI for NLP

- Integrated Gradients to guide learning and de-bias models.
  Requires users to specify the target attribution value for tokens of interest.

Method	Sentence	Probability
Baseline	I am *gay*	0.915
	I am straight	0.085
Our Method	*I am* gay	0.141
	I am straight	0.144

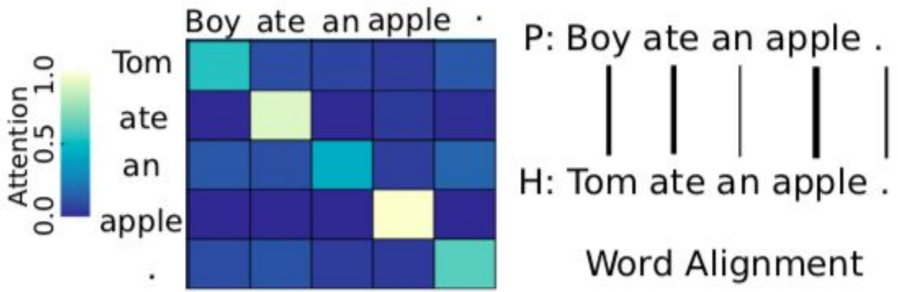
Table 1: Toxicity probabilities for samples of a baseline CNN model and our proposed method. Words are shaded based on their attribution and italicized if attribution is > 0.

$$\mathcal{L}^{joint} = \mathcal{L}(\boldsymbol{y}, \boldsymbol{p}) + \lambda \sum_{c}^{C} \mathcal{L}^{prior}(\boldsymbol{a}^{\boldsymbol{c}}, \boldsymbol{t}^{\boldsymbol{c}}) \quad (3)$$

where $\boldsymbol{a}^c$ and $\boldsymbol{t}^c$ are the attribution and attribution target for class c, λ is the hyperparameter that con-

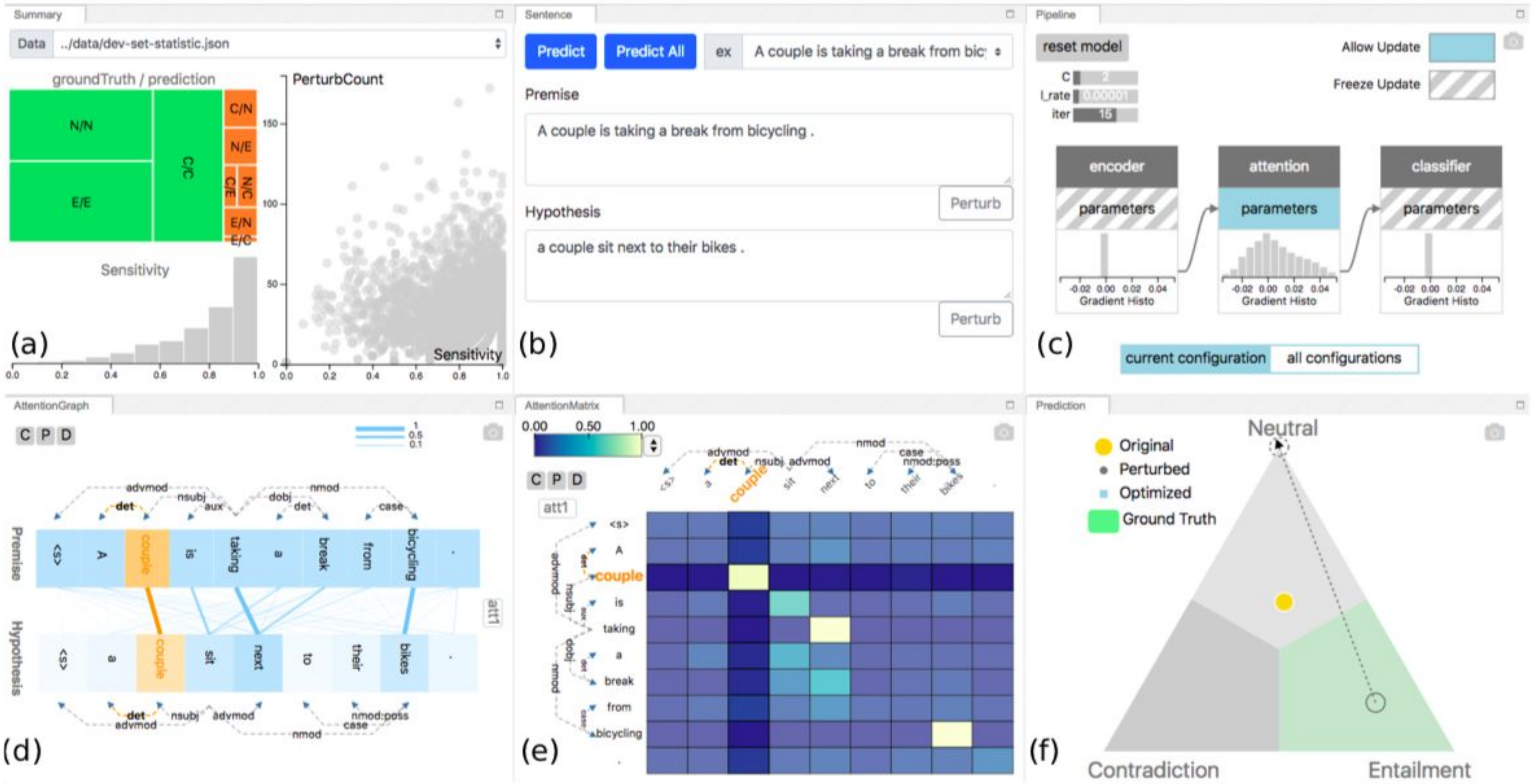
Liu & Avci 2019: Incorporating Priors with Feature Attribution on Text Classification

vying for Attention

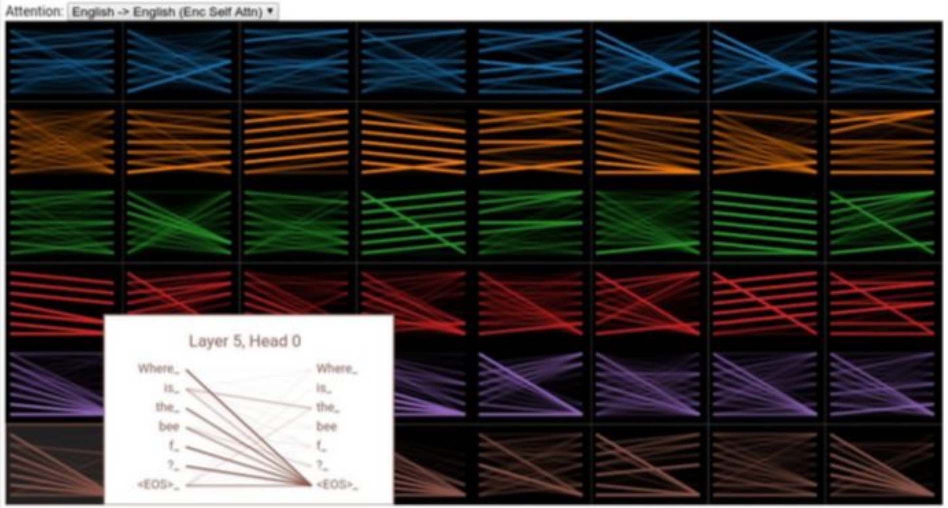


- Attention is All You Need (2017)
- Attention is Not Explanation (2019)
- Attention is Not Not Explanation (2019)
- What Does BERT Look At? An Analysis of BERT's Attention (2019)
- Analyzing the Structure of Attention in a Transformer Language Model (2019)
- Is Attention Interpretable? (2019)
- On the validity of Self-Attention as Explanation in Transformer Models? (2019)
- NLIZE: A Perturbation-Driven Visual Interrogation Tool for Analyzing and Interpreting Natural Language Inference Models (2019)

vying for Attention

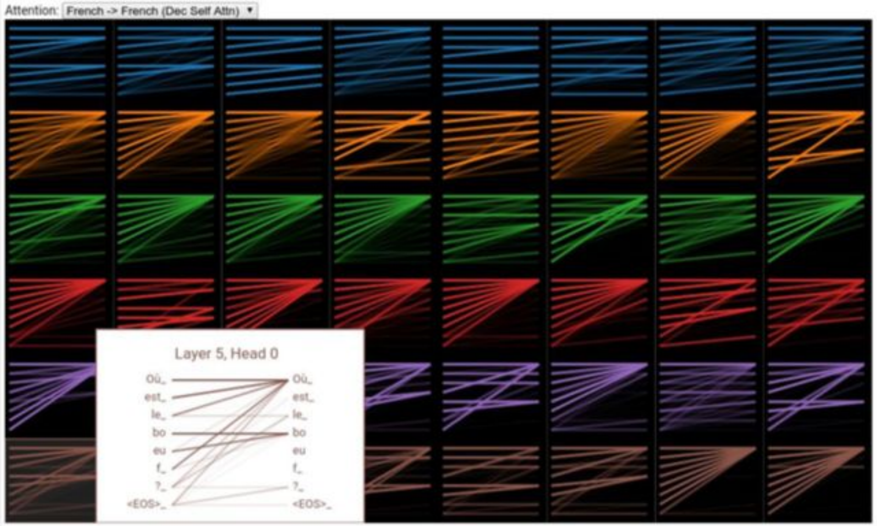


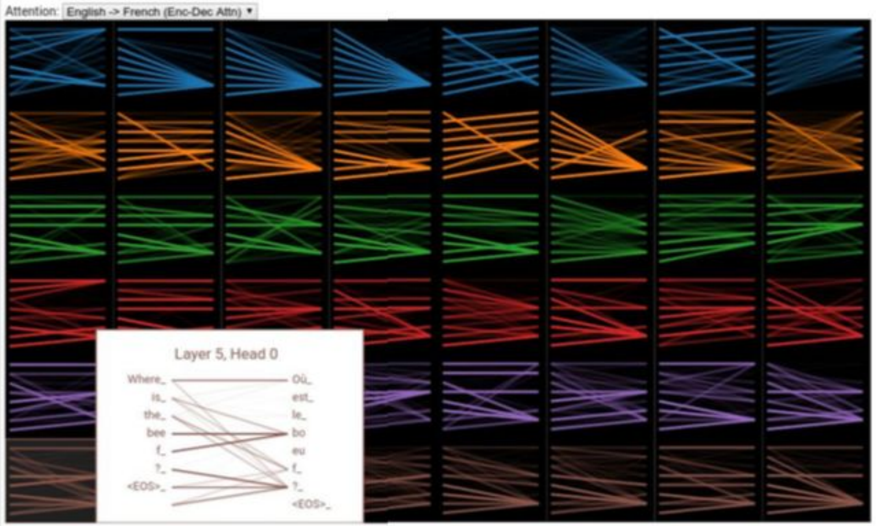
Explaining **Attention** to humans



6 layers / **8** attention heads

1. Encoder self attention
2. Decoder self attention
3. Encoder - Decoder attention





2. XAI for NLP

- For seq2seq tasks, XAI is less mature.
- Ongoing work on “explaining seq2seq models” for machine translation (looking at LSTMs / Transformers)*
- A lot of work on analyzing meaning of learned word embeddings, what phenomena models are actually learning & how to construct adversarial datasets from statistical cues for robustness purposes

- Learning Dense Representations for Entity Retrieval
- BERT Rediscovers the Classical NLP Pipeline
- Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in NLI
- Probing Neural Network Comprehension of Natural Language Arguments
- Learning The Difference That Makes A Difference: Counterfactually Augmented Data

Figure 3: A 2D projection of country embeddings (using t-SNE), color coded by continent.

link

Automated Evaluation still not there

- BLEU - Bilingual Evaluation Understudy
  - average of n-gram overlap (1-4) precision between a generated output and reference translations with a penalty for shorter outputs.
  - Good post on BLEU's limitations (*only use it for MT of documents*)
- ROUGE- Recall-Oriented Understudy for Gisting Evaluation
  - looks at how many n-grams in the reference translation show up in the output, rather than the reverse (focuses on recall rather than precision)
- Perplexity: if you don't have reference texts (pros/cons)
- BertScore (link): compare token embeddings for distance
- Human Eval (gold standard)

3. Generating Black Box Text Counterfactuals with RL

Negative Review

Long, boring, blasphemous. Never have I been so glad to see ending credits roll.

Human generated Positive Counterfactual Review:

Long, fascinating, soulful. Never have I been so sad to see ending credits roll.

** Super Preliminary Work !*

Setup: Dataset: 2.4k negative reviews / 2.4k positive human generated CF reviews

Initial input: **Long, boring, blasphemous. ...**

States: (current word, context, part of speech)

Actions: **Substitute** or **skip** word

Rewards: based on cosine_distance between initial and current sentence [0,1] and whether the sentiment of the review has changed.

If word is "skipped" -> **a reward of zero**

If its "substituted" -> **reward is a function of distance between new & initial review**

If counterfactual is reached we are **done**,

-> a reward of **100 - DM * cosine_distance** is given where DM is tunable param.

If max number of iterations or substitutions reached

-> a reward of **-100 + (1 / cosine_distance)**

Sentiment Model: BERT Uncased embeddings fine tuned on IMDB data

Substitution Mechanism:

1. Mask current word in the review
2. query Bert with sentence with masked word
3. Get top 5 candidates, filter based on part of speech and prior use
4. Sample from list based on probability weights
5. Replace current word in sentence with sampled word

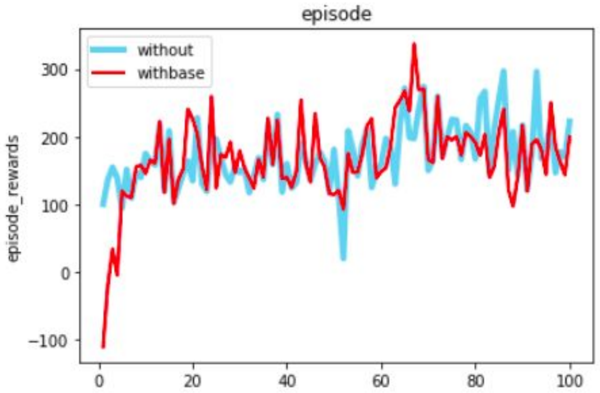
Each State: [Word, Current Sentence, POS] =
[768 dim embedding, 768 dim embedding , 20 dim one hot vec]

We feed this vector into our Policy & Value functions for our Actor Critic model

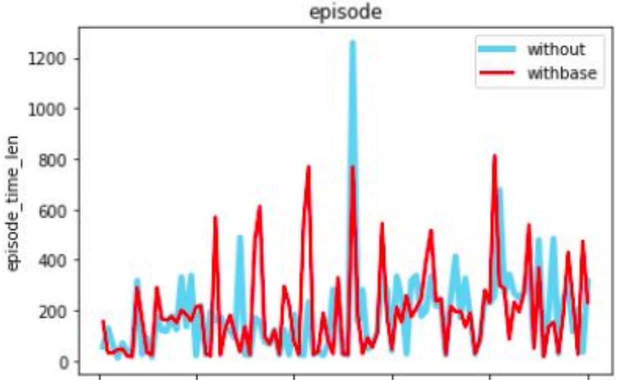
Actor learns to identify whether or not it’s beneficial to substitute a word

REINFORCE & REINFORCE + baseline

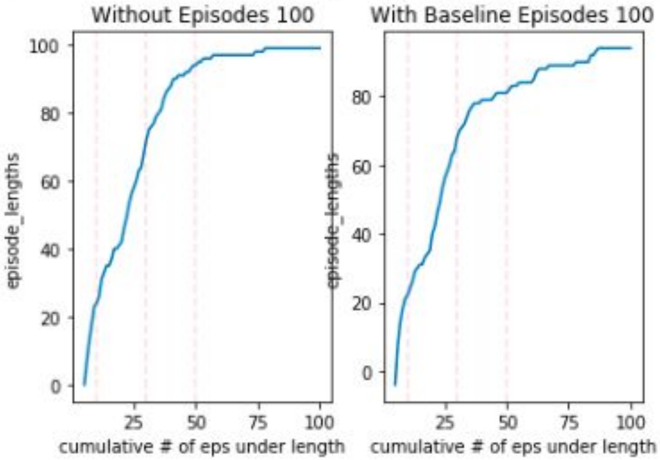
```
episode_rewards
Without mean 172.399 min 19.656 max 296.77 sum
Withbase mean 167.412 min -111.139 max 337.116
```




```
episode_time_len
Without mean 180.986 min 12.52 max 1258.696 sum
Withbase mean 194.023 min 15.673 max 811.491 sum
```




```
Pct under  6,  7,  8,  9,  10, 15, 20, 30, 40, 50, 60 steps
WithoutB:  [7, 13, 18, 23, 24, 35, 42, 72, 88, 94, 97]
Baseline: [7, 14, 18, 21, 22, 31, 40, 68, 79, 81, 84]
```



Initial findings and future considerations:

1) Automate analysis of change comparisons between my output and Lipton’s dataset

2) Importance of **context** and **attribution markers**
- Initial results are able to get CFs but change context words and meaning
  (ie “Nicolas Cage” -> Nicolas Castle)
- Compare against baseline Liang’s paper (debatable “black box”)
  Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer (2019)
- Versus simply preventing change of pronouns

3) Sampling vs Sequential, better pre-training of our actor model,
- Is Jittering enough to get where we want to go ?
- Guide with spans/external models (Perplexity / BertScore / Entailment / SpanBERT)?
- Do I need to distill to be fair?

4) Literature in Adversarial Attack and Style Transfer domains

Thanks!

Questions / Thoughts?

Donning his new canine decoder, Professor Schwartzman becomes the first human being on Earth to hear what barking dogs are actually saying.