Relationship between Attention, Complexity and Explainability

David Cerny, Oliviero Nardi, Liselore Borel Rinkes

Agenda

- Introduction
  - AI & Transparency
  - Attention as an Explanation?
  - “Attention is not Explanation”
- Project focus
- Methodology
- Experiments
  - Setup
  - Results
- Conclusion

Agenda

- **Introduction**
  - AI & Transparency
  - Attention as an Explanation?
  - “Attention is not Explanation”
- Project focus
- Methodology
- Experiments
  - Setup
  - Results
- Conclusion

Introduction: AI & Transparency

- **Transparency** is the property of being transparent or *interpretable*
  - ***Interpretability*** *is the degree to which a human can understand the cause of a decision*[1]
- Need for transparency increases as AI systems grow
- How can we make models interpretable?
  - Attention mechanism

[1] (Miller, 2019, Explanation in artificial intelligence: Insights from the social sciences, AI Journal)

Introduction: Attention as Explanation?

- **Attention** is a way of defining a weighted sum over input features
  - Weights are learned
- Oftentimes presented as an explanator
  - Weights seem to resemble the “focus” of the model
- However, not everyone agrees
  - *Attention is not Explanation*, Jain & Wallace, 2019

Introduction: “Attention is not Explanation”

- Questions whether attention provides **meaningful explanations**
- Experiments across various datasets and domains
  - Binary Classification (BC), Question Answering (QA), Natural Language Inference (NLI)
- Results showed:
  - Attention does not correlate well with measures of **feature importance**
    - Gradient and leave-one-out measures
  - Counterfactual attention distributions exist
    - Permutation of weights
    - Adversarial attention
- However, this stance has been criticised
  - *Attention is Not Not Explanation*, Wiegreffe & Pinter, 2019
  - *Attention Interpretability Across NLP Tasks*, Vashishth et al, 2019
- Are the tasks complex enough?

Agenda

- Introduction
  - AI & Transparency
  - Attention as an Explanation?
  - “Attention is not Explanation”
- **Project focus**
- Methodology
- Experiments
  - Setup
  - Results
- Conclusion

Project focus

“Interpretable attention correlates with the complexity of the task”

- Seq2seq models
  - Neural Machine Translation (NMT) and autoencoder
- Our experiments
  - Do these models use attention?
    - Baseline performance
  - Given a more complex task, will we come to the same conclusion as Jain & Wallace?

Agenda

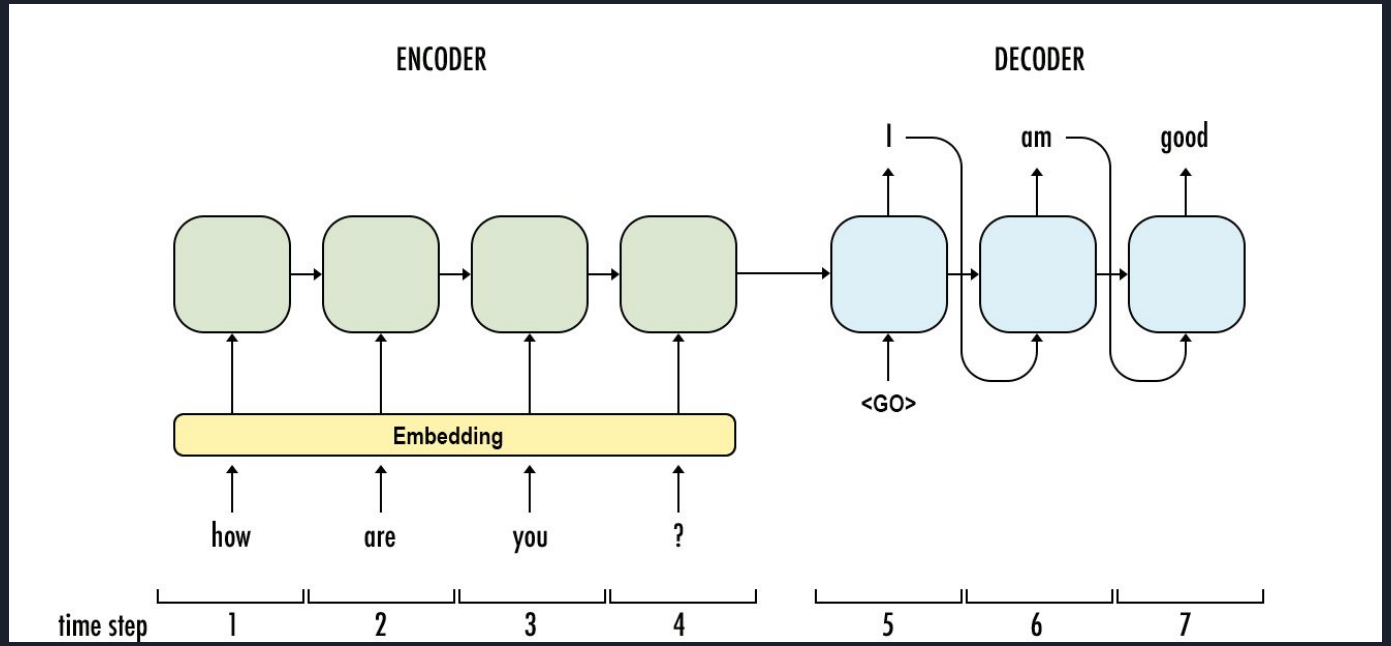
- Introduction
  - AI & Transparency
  - Attention as an Explanation?
  - “Attention is not Explanation”
- Project focus
- **Methodology**
- Experiments
  - Setup
  - Results
- Conclusion

Methodology

- “Attention is not Explanation” codebase[1]
  - Tasks: BC, QA, NLI
  - API expansion
- Seq2seq models
  - GRU Encoder/Decoder architecture
  - Attention defined over every decoding step
    - Each step can be seen as classification
    - Permutation is performed at each decoding step
- Two tasks: Sequence autoencoding and Neural Machine Translation (NMT)
  - Mirrored dataset for autoencoding
  - NMT: French to English
  - BLEU score

[1] https://successar.github.io/AttentionExplanation/

Methodology



https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d

Agenda

- Introduction
  - AI & Transparency
  - Attention as an Explanation?
  - “Attention is not Explanation”
- Project focus
- Methodology
- **Experiments**
  - Setup
  - Results
- Conclusion

Experiments: Setup

- “Attention is not explanation” setup
  - Focus on Bi-LSTM
  - Performance test
    - datasets → SST, AGNews, 20News, IMDB, babI 1, SNLI
    - attention → tanh, dot, none
    - accuracy
- Seq2seq
  - Performance test
    - linear vs. uniform attention
    - BLEU
  - Permutation test
    - 100 attention permutations per step/word
      - Total Variational Distance: sum of absolute distances
      - we take the median TVD
    - Maximum weight binning
    - Aggregate by average and. by maximum

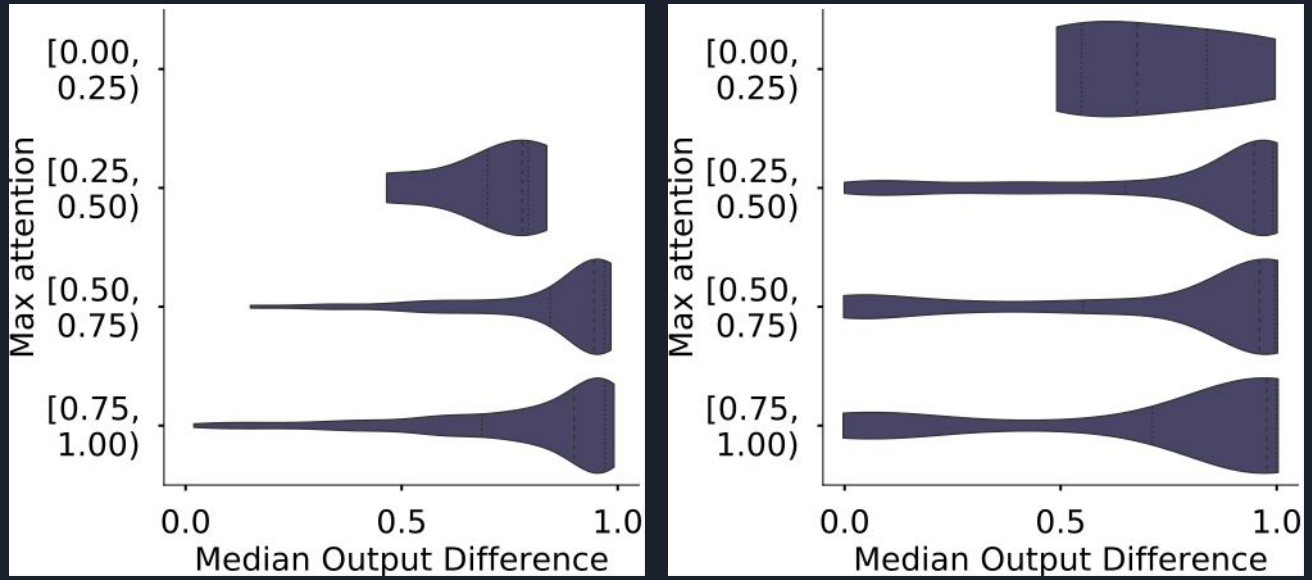
Experiments: Results

	BC				QA	NLI
	SST	**AG News**	**IMDB**	**20News**	**bAbI 1**	**SNLI**
tanh	0.804	0.955	0.886	0.931	0.913	0.762
dot	0.795	0.950	0.875	0.932	0.997	0.741
none	0.795	0.953	0.872	0.867	0.589	0.744
train size	6355	60000	25000	1426	10000	549367
test size	1725	3800	4356	334	1000	9824

Attention	**Autoencoder**	**NMT**
Yes	0.598	0.331
No	0.524	0.329

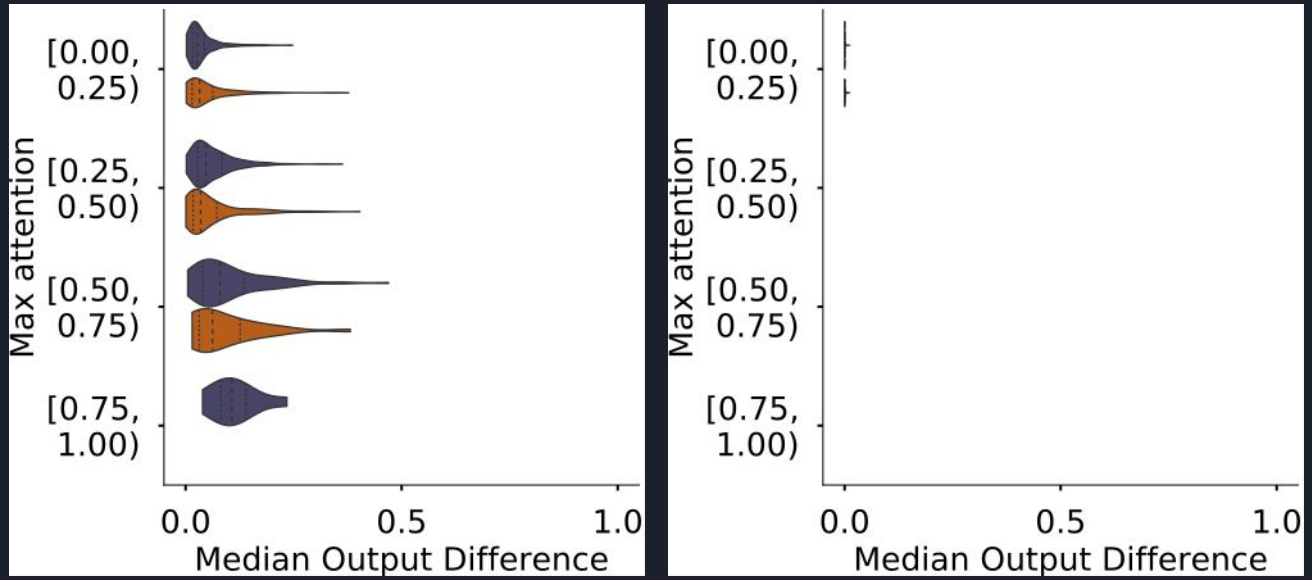
Performance results of all experiments with and without attention. (top): results of the original paper evaluated by accuracy. (bottom): results from the seq2seq domain evaluated by BLEU

Experiments: Results



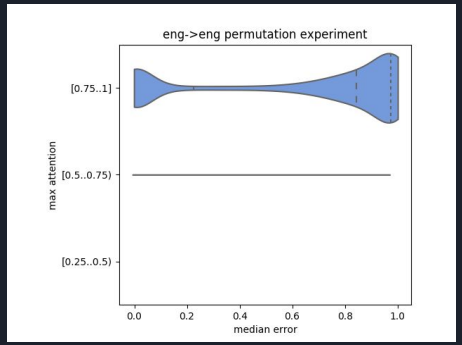
Permutation Experiment results on the bAbI 1 dataset. Left: TANH attention. Right: DOT attention. See https://successar.github.io/AttentionExplanation/docs/

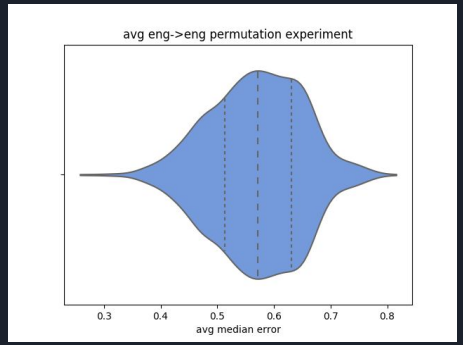
Experiments: Results

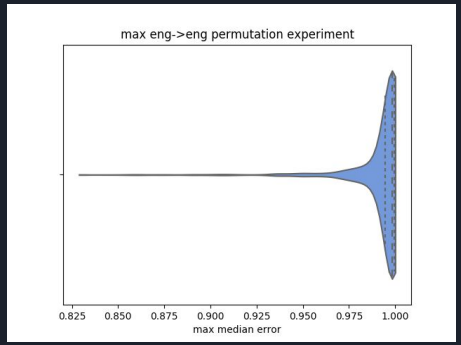


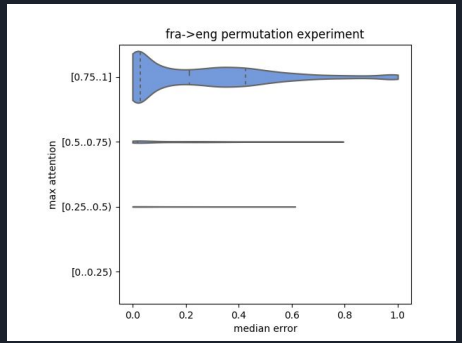
Permutation Experiment results on the SST dataset. Left: TANH attention. Right: DOT attention. See https://successar.github.io/AttentionExplanation/docs/

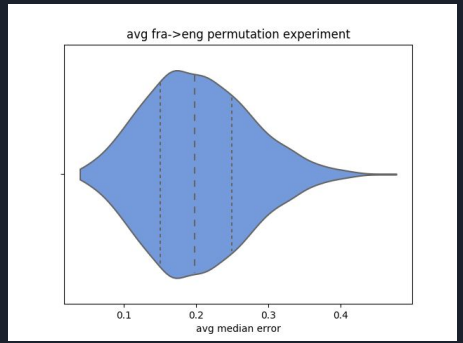
Experiments: Results

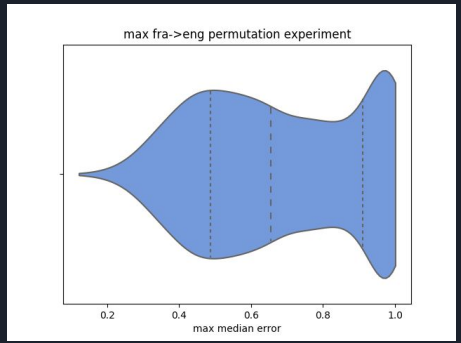












Violin plots displaying the results of the permutation experiments. The distributions show the median TVD error over 100 attention permutations.

Agenda

- Introduction
  - AI & Transparency
  - Attention as an Explanation?
  - “Attention is not Explanation”
- Project focus
- Methodology
- Experiments
  - Setup
  - Results
- **Conclusion**

Conclusion

“Interpretable attention correlates with the complexity of the task”

- Summary
  - not decisive; attention might be explanation
  - bias in original work

Thanks for the *Attention*!

David Cerny, Oliviero Nardi, Liselore Borel Rinkes