Regularization, Semi-supervision, and Supervision for a Plausible Attention-Based Explanation

D.H. Nguyen, C. Mallart, G. Gravier, P. Sébillot

Attention-based explanation

Objective

Explain a predictive decision to a non-expert human.

Explanation techniques

- Post-hoc: LIME, SHAP, Gradient – based [1]
- Intrinsic: **Attention rationale** extraction [2]

Explanability qualities

- Faithfulness [3]
- **Plausibility** [4]

Classification tasks

Task	Dataset	Labels
Natural Language Inference	e-SNLI	Entailment / Contradiction / Neutral
Sentiment classification	YelpHat	Negative / Positive
Hate speech detection	HateXPlain	Hateful / Offensive / Neutral

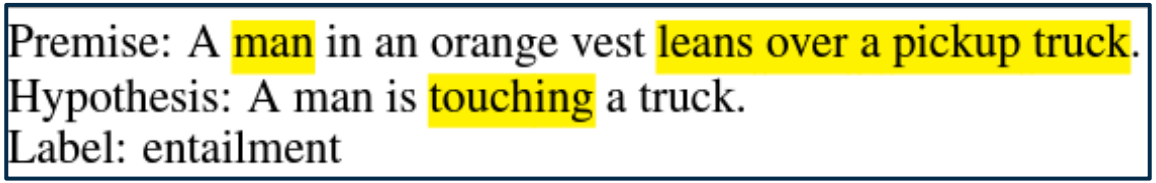


Fig 1. An example of human annotations in e-SNLI

Attention model Bi-LSTM encoders

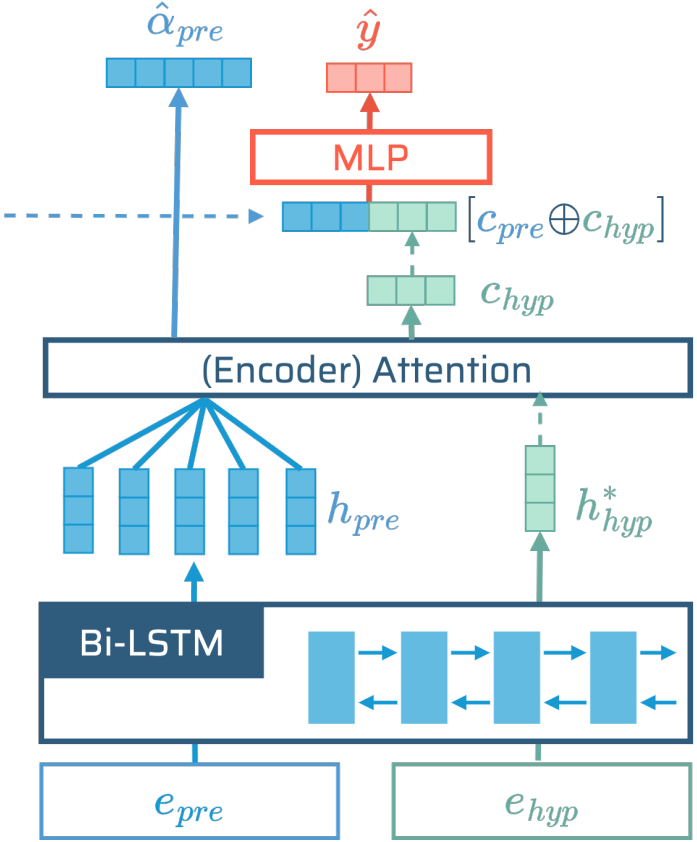


Fig 2. Bi-LSTM based Attention model

Attention encodes Bi-LSTM contextualized vectors h into context c.

Classification is done based on c

$$\hat{y} = MLP(c).$$

Explanation is given by attention map (or attention rationale) $\hat{\alpha}$.

Explanation from baseline attention

	Premise	Hypothesis
GROUNDTRUTH	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .
Baseline	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .

Fig 3. **GROUNDTRUTH** highlighted by human annotator and attention rationale from **Baseline** model.

Challenges

Attention rationale is spreading.
Human annotation is not always available.

Additional constraint to improve plausibility

$$\mathcal{L}(y, \hat{y}, \hat{\alpha}) = \mathcal{L}_c(y, \hat{y}) + \boldsymbol{\lambda}\, \mathcal{L}_\alpha(\hat{\alpha})$$

Supervision : Guide attention to explain closely to human annotation.

$$\mathcal{L}_{sup}(\hat{\beta}, \alpha) = \frac{\hat{\beta}^{\intercal}\alpha}{\sum_i^L \hat{\beta}_i + \sum_i^L \alpha_i - \hat{\beta}^{\intercal}\alpha}$$

Regularization : Make attention map to focus on few tokens [5].

$$\mathcal{L}_{reg}(\hat{\alpha}) = -\sum_i^L \hat{\alpha}_i \log_L(\hat{\alpha}_i)$$

Semi-supervision : Generate a heuristic maps $\tilde{\alpha}$ based on morpho-syntactic [5], then use it instead of human annotation.

$$\mathcal{L}_{semi}(\hat{\alpha}, \tilde{\alpha}) = \tilde{\alpha}\,[\log(\tilde{\alpha}) - \log(\hat{\alpha})]$$

Regularization by entropy

	Premise	Hypothesis	Label
GROUNDTRUTH	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
Baseline	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
λ=0.01	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
λ=0.02	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
λ=0.06	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment

Fig 4. Regularized attention rationale in one e-SNLI example.

Regularization by entropy

Supervision			Regularization				
λ	AUPRC	F-Score	λ	AUPRC	F-Score	Recall	Specificity
0.00	0.444 ± 0.006	0.815 ± 0.003	0.00	0.444 ± 0.006	0.815 ± 0.003	0.430 ± 0.003	0.893 ± 0.001
0.10*	0.506 ± 0.001	0.812 ± 0.000	0.02*	**0.492** ± 0.005	0.815 ± 0.003	0.394 ± 0.005	0.921 ± 0.004
1.00	**0.544** ± 0.000	0.798 ± 0.000	0.30	0.238 ± 0.050	0.787 ± 0.001	0.198 ± 0.006	0.911 ± 0.028

Regularization is sensible to λ .
As attention rationale shrinks, the model focus more on plausible tokens.

Semi-supervision by a heuristic map

	Premise	Hypothesis	Label
GROUNDTRUTH	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
HEURISTIC	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
Baseline	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
λ=0.01	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
λ=0.03	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment
λ=0.04	Two children re laying on a rug with some wooden bricks laid out in a square between them .	Two children are on a rug .	entailment

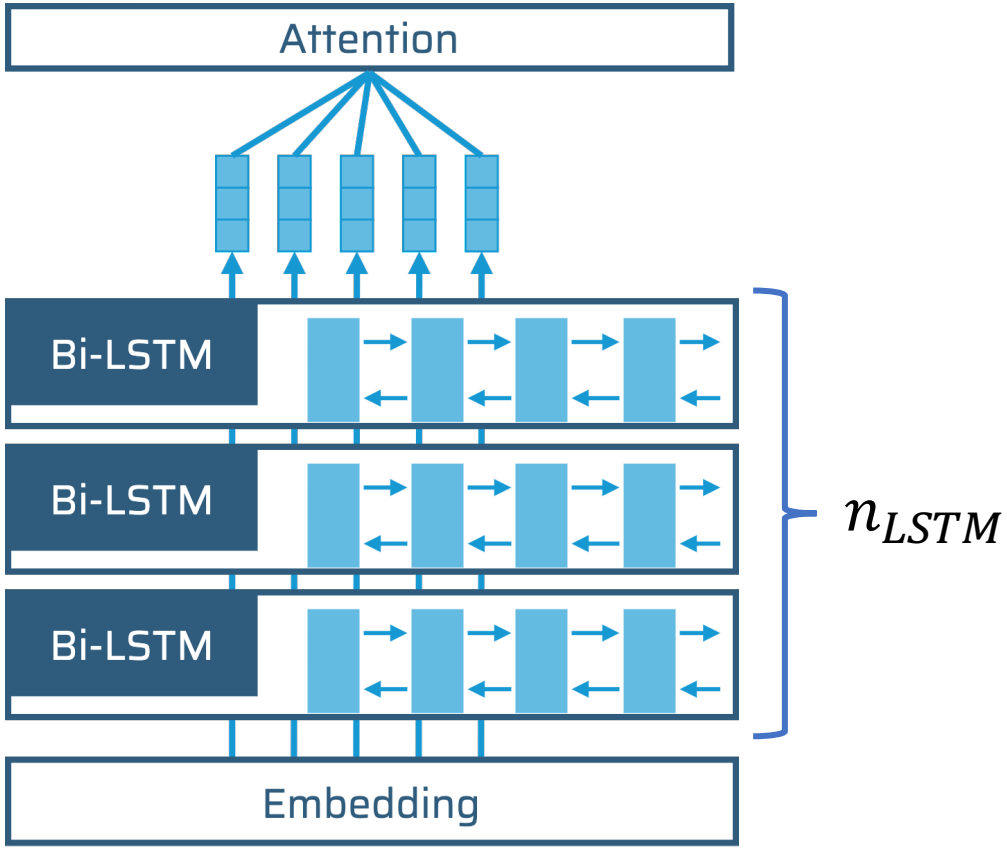
Fig 5. Semi-supervised attention rationale in one e-SNLI example.

Semi-supervision by a heuristic map

Supervision			Regularization				
λ	AUPRC	F-Score	λ	AUPRC	F-Score	Recall	Specificity
0.00	0.444 ± 0.006	0.815 ± 0.003	0.00	0.444 ± 0.006	0.815 ± 0.003	0.430 ± 0.002	0.893 ± 0.001
0.10*	0.506 ± 0.001	0.812 ± 0.000	0.02*	**0.460** ± 0.000	0.813 ± 0.000	**0.483** ± 0.000	0.854 ± 0.000
1.00	**0.544** ± 0.000	0.798 ± 0.000	0.30	0.437 ± 0.000	0.817 ± 0.000	0.431 ± 0.005	0.892 ± 0.003

Semi-supervision is less dependant to λ.
Attention rationale is searching for new plausible words.

Plausibility in deep contextualized representation



n_{LSTM}	AUPRC	F-Score
1	**0.444** ± 0.006	**0.815** ± 0.003
3	0.407 ± 0.004	0.803 ± 0.001
5	0.341 ± 0.014	0.779 ± 0.006

The deeper the model, the less plausible it is.

Can regularization improve plausibility in deeper models?

Supervision and regularization cannot alleviate this limitation.

Semi-supervision converges to heuristic rationale regardless to this limitation.

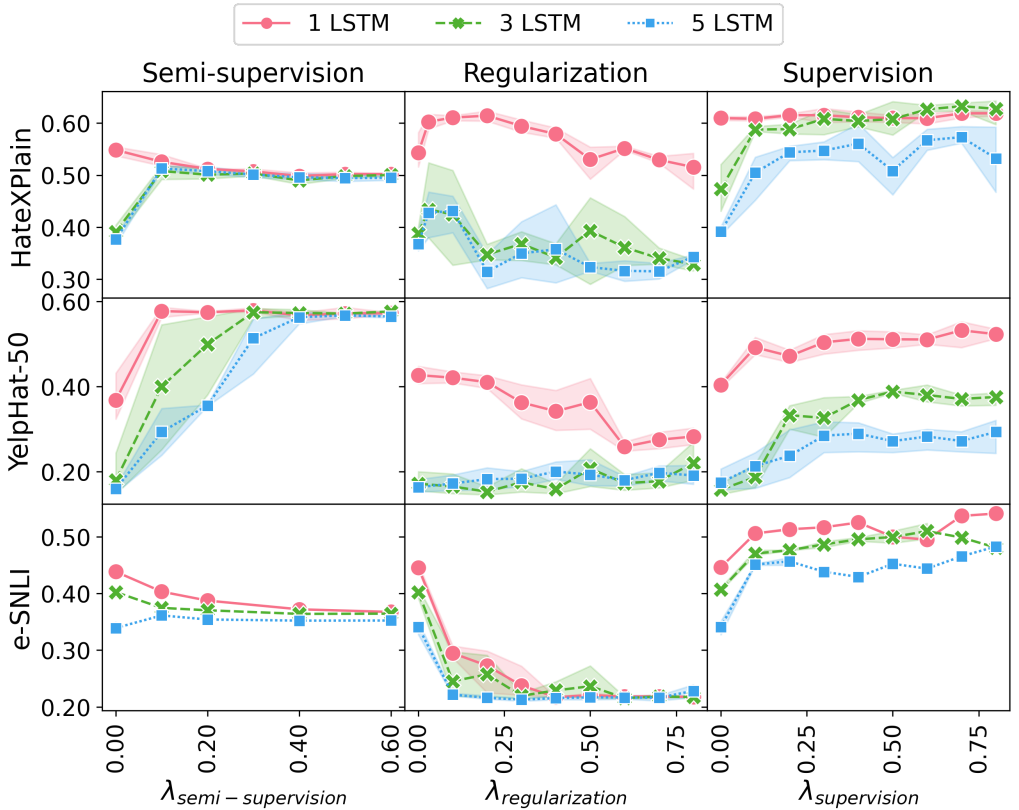


Fig 6. Plausibility (AUPRC) as a function of λ

Conclusion

- Regularization and semi-supervision can improve attention plausibility without losing performance
- Semi-supervision based on morpho-syntax is a more robust technique to improve plausiblity (on the tested tasks)
- Deeper contextualization poses a challenge to plausibility improvement

Future works:

- Impact of contextualisation on model plausibility
- Generalization in different architectures

Our code is available!

- duc-hau.nguyen@irisa.fr
- cyrielle.mallart@univ-rennes2.fr
- pascale.sebillot@irisa.fr
- guillaume.gravier@irisa.fr

GitHub/Kihansi95

www.irisa.fr @irisa_lab

cnrs UMR IRISA Université de Rennes INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

Institut de Recherche en Informatique et Systèmes Aléatoires

Bibliography

[1] Bastings, Jasmijn, and Katja Filippova. “The Elephant in the Interpretability Room: Why Use Attention as Explanation When We Have Saliency Methods?” In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. ACL 2020.

[2] Mohankumar, Akash Kumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. “Towards Transparent and Explainable Attention Models.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020.

[3] Jacovi, Alon, and Yoav Goldberg. “Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020.

[4] Nguyen, Duc Hau, Guillaume Gravier, and Pascale Sébillot. “A Study of the Plausibility of Attention between RNN Encoders in Natural Language Inference.” ICMLA 2021.

[5] Nguyen, Duc Hau, Guillaume Gravier, and Pascale Sébillot. “Filtrage et Régularisation Pour Améliorer La Plausibilité Des Poids d’attention Dans La Tâche d’inférence En Langue Naturelle.” TALN 2022.