



HAL
open science

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

Duc Hau Nguyen, Cyrielle Mallart, Guillaume Gravier, Pascale Sébillot

► **To cite this version:**

Duc Hau Nguyen, Cyrielle Mallart, Guillaume Gravier, Pascale Sébillot. Regularization, Semi-supervision, and Supervision for a Plausible Attention-Based Explanation. 28th International Conference on Natural Language and Information Systems (NLDB), Jun 2023, Derby, United Kingdom. pp.285-298. hal-04132646

HAL Id: hal-04132646

<https://hal.science/hal-04132646v1>

Submitted on 19 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Regularization, Semi-supervision, and Supervision for a Plausible Attention-Based Explanation ^{*}

Duc Hau Nguyen¹[0000-0002-4061-3114], Cyrielle Mallart³[0009-0001-1420-9548],
Guillaume Gravier¹[0000-0002-2266-5682], and Pascale
Sébillot²[0000-0002-5429-4302]

¹ Univ Rennes, CNRS, Inria - IRISA, Rennes, France,
{duc-hau.nguyen,guig}@irisa.fr

² Univ Rennes, CNRS, Inria, INSA Rennes - IRISA, Rennes, France,
pascale.sebillot@irisa.fr

³ University of Rennes 2, Rennes, France, cyrielle.mallart@univ-rennes2.fr

Abstract. Attention mechanism is contributing to the majority of recent advances in machine learning for natural language processing. Additionally, it results in an attention map that shows the proportional influence of each input in its decision. Empirical studies postulate that attention maps can be provided as an explanation for model output. However, it is still questionable to ask whether this explanation helps regular people to understand and accept the model output (the plausibility of the explanation). Recent studies show that attention weights in RNN encoders are hardly plausible because they spread on input tokens. We thus propose three additional constraints to the learning objective function to improve the plausibility of the attention map: regularization to increase the attention weight sparsity, semi-supervision to supervise the map by a heuristic and supervision by human annotation. Results show that all techniques can improve the attention map plausibility at some level. We also observe that specific instructions for human annotation might have a negative effect on classification performance. Beyond the attention map, results on text classification tasks also show that the contextualization layer plays a crucial role in finding the right space for finding plausible tokens, no matter how constraints bring the gain.

Keywords: Attention mechanism · Explainability · Plausibility · Regularization · Semi-supervision · Supervision.

1 Introduction

Attention mechanisms [15] play a crucial role in recent success across many natural language processing (NLP) tasks and are present in most recent neural

^{*} Work partially funded by grant ANR-19-CE38-0011-03 from the French national research agency (ANR).

models. As a layer in a complex neural network, the attention mechanism attributes a weight to each input token and encodes each into a context vector through a weighted sum of the input vectors. When this context vector is used for prediction, the weight vector, also called *attention map* [11], can be considered as an explanation by showing the degree of influence of each input token to the prediction.

Despite the potential of attention maps as a form of explanation, there are concerns [11] about their validity on two properties that are not guaranteed: faithfulness and plausibility. Faithfulness, a widely discussed problem [19,3], focuses on whether the weight associated with a token reflects its influence on the prediction. Plausibility refers to the extent to which the attention map can resemble human reasoning [19,31].

While plausibility is an interesting feature that allows to present an easily comprehensible way to individuals with limited knowledge of neural models without additional computational costs, the contributions in this direction remain limited and rare. Given that multiple studies have suggested that raw attention weights lack plausibility (see, e.g., [20]), the issue of forcing their plausibility is an obvious one that calls for further exploration. As it is proven possible to incorporate constraints on attention while maintaining satisfactory performance [25,12,31], we propose three approaches for enforcing plausibility constraints on attention maps, namely, sparsity regularization, semi-supervised learning, and supervised learning.

The main contributions in this paper are : (1) we can to some extent force the model plausibility (as demonstrated by supervision) at no accuracy cost, (2) both regularization and semi-supervision can optimize the plausibility but the latter offers a solution without compromising performance and (3) the deep contextualization is harmful to attention plausibility. The last result provides insight into why this hardly transfers to transformers.

2 Related Works

The attention mechanism is widely used as a possible feature to explain the model decision [23,28]. However, the local explanation is facing an issue that the attention map can be manipulated while keeping the same prediction [12,31,29]. While this feature is considered as a weak faithful explanation [27,1], this enables the selection of a plausible map.

Among the few studies on attention supervision, [18] showed that supervision can harm classification performance in sentiment classification tasks. Regularization was considered to circumvent the issue of a rather flat distribution of attention weights as reported by [13]. [19] suggested an additional constraint in the learning objective to force this representation to be sparse.

To overcome the lack of human references in many datasets, many contributions offer task-specific solutions, such as [22] guiding attention based on topic-related vocabulary and [14] using a WordNet-based heuristic for evidence inference, while [20] provides an effective heuristic map that is closer to human

annotation but only for natural language inference (NLI). While the authors of existing techniques have not fully explored their effects and limitations in different tasks, this study aims to provide a comprehensive view of how different techniques improve attention plausibility.

Hard attention, also referred to as rationalized learning in the literature [5], is an alternative form of the attention mechanism that comprises two components: a generator function that masks irrelevant input tokens, and a predictor that is trained to make predictions on the remaining inputs. While hard-attention is advantageous with respect to soft-attention in robustness and faithfulness aspects, it introduced a trade-off between sparsity and accuracy because the full context is inaccessible [24].

Other post-hoc explanation techniques can provide faithful explanations (such as gradient-based methods or feature suppression), however with two main drawbacks: (i) incurring additional computational costs during each inference and (ii) offering benefits only to the model developer, without the flexibility to impose constraints for plausible explanations [2] while its explanation cannot be guaranteed to be plausible for end-users [3,21].

To the best of our knowledge, no existing study has brought a broad and comprehensive overview of how different techniques improve attention plausibility. Although regularization techniques are independent of human annotation and heuristics can overcome the lack of human annotation, it is still unclear how they improve plausibility compared to supervision. Furthermore, the authors of the existing techniques suggest improvement without questioning their implications and limitations in different tasks, especially in soft-attention models. This study focuses on addressing these fundamental issues and does not include a comparative analysis of hard-attention techniques and post-hoc explanation methods but they are promising for future works.

3 Tasks and Datasets

To ensure the generalization of our findings across different tasks, we investigate three different datasets from the ERASER benchmark [6] and [30] designed for plausibility studies.

The e-SNLI corpus [4], a reference dataset in NLI, consists of pairs of sentences, a premise and a hypothesis with a label stating whether the hypothesis entails, contradicts, or is unrelated to the premise. The annotators also answered the question *Why is a pair of sentences in a relation of entailment, neutrality, or contradiction?* by highlighting the relevant words in both the premise and hypothesis and providing a short explanatory text. The corpus consists of 549,367 sentence pairs for training and 9,842 pairs for the validation and test sets respectively. Note that the SNLI corpus is known to have artifacts [10], where some lexical fields appear mostly in one class. Also, the annotation instruction in e-SNLI leads to some particularities, such as not highlighting the common words between premise and hypothesis, thus making the annotation not convincing in some cases.

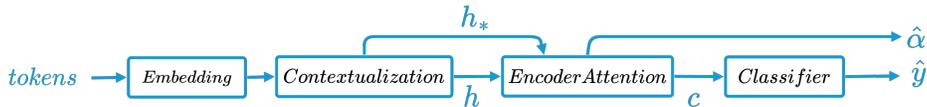


Fig. 1: Generic architecture of a RNN-based attention model for classification.

The HateXPlain dataset [17] was conceived by gathering posts from social networks that were labeled for the detection of hate speech. Each post belongs to either one of three labels: offensive, hateful, and normal speech. Annotators were also instructed to highlight the relevant part of the post to justify their choice of a label. Overall, the corpus consists of 15,383 posts for training, 1,922 for validation, and 1,924 for testing.

The Yelp-Hat dataset [26] was obtained by gathering reviews on restaurants from a website and by asking reviewers to highlight parts of the text to justify their choice. The corpus consists of 3,482 reviews for training and validation⁴. Yelp-Hat was split randomly into 2,436 sentences for training and 1,046 for validation.

4 Attention Mechanism on RNN Encoders

Being one of the most studied in NLP yet the most controversial in the explainability debate [3], we employ the attention model with RNN encoders. Preliminary experiments on BERT-like self-attention models have shown little hope in finding a single layer or head to provide plausible explanations.

The model, illustrated in Figure 1, consists of an embedding layer and a bi-LSTM layer, which produce contextualized token representations h_i , for $i \in [1, L]$ in a sentence of length L , as well as a sentence embedding h_* , which is the concatenation of the forward and backward last state. The attention encoder assigns weights $\hat{\alpha}_i$ to each h_i and computes a context vector c through a weighted sum. Finally, a multilayer perceptron classifier is applied to c for prediction. We also consider attention weights on each input token in the loss function. To simplify notation, we use the notation $h = [h_1, \dots, h_L]^T = [h_i]_{i=1}^L$ to denote the sequence of bi-LSTM outputs and $\hat{\alpha} = [\hat{\alpha}_i]_{i=1}^L$ to denote the attention weights. In this paper, we distinguish the model attention map $\hat{\alpha}$ from α which refers to the human annotation binary map.

The attention layer is adapted differently for various tasks. We begin by describing the attention layer formally as a function that takes query q , key k , and value v as input and generates c and $\hat{\alpha}$ as output according to [15]:

$$c, \hat{\alpha} = \text{Attention}(q, k, v) . \quad (1)$$

In text classification, the attention layer queries the text embedding ($q = h_*$) and use the token contextualized vectors as both key and value ($k = v = h$). The prediction is made on c . In NLI, we use text embeddings of the premise or

⁴ 15 “incoherent” samples were excluded, such as incompatible annotation maps and number of tokens in reviews.

hypothesis as query ($q = \bar{h}_*$) and keys/values are bi-LSTM representations from the opposite sentence ($k = v = h$). As a result, we obtain two context vectors, c_p for the premise and c_h for the hypothesis, which are concatenated [$c_p \oplus c_h$] for prediction.

5 Constraints on the Objective Function

We propose to control the behavior of the attention layer to improve its plausibility by extending the loss function to include a term on the attention

$$\mathcal{L}(y, \hat{y}, \hat{\alpha}) = \mathcal{L}_c(y, \hat{y}) + \lambda \mathcal{L}_a(\hat{\alpha}) \quad (2)$$

where we combine the classification loss L_c (cross-entropy loss) with a constraint on attention map $\mathcal{L}_a(\hat{\alpha})$ weighted by $\lambda \in [0, 1]$. We detail hereunder the different forms for \mathcal{L}_a in three approaches.

5.1 Sparsity Regularization

The sparsity constraint can be expressed in many different ways, which have different but marginal effects on convergence speed, or on the resulting explanation [19,13]. Shannon entropy offers a straightforward yet effective method to measure sparsity, where high entropy values indicate uniform weight distributions and low values indicate sparse ones. We incorporate Shannon entropy as a loss function, defined as

$$\mathcal{L}_a(\hat{\alpha}) = - \sum_{i=1}^L \hat{\alpha}_i \log_L(\hat{\alpha}_i) \quad (3)$$

5.2 Supervision from Reference Annotation

A difficulty in supervising attention layers with a reference annotation is that attention weights and reference annotations are conceptually of different nature. The former are weights such that $\sum \hat{\alpha}_i = 1$ while the latter are binary indicators of whether a token is useful for a plausible explanation or not. Contrary to [22], supervision directly on the attention map $\hat{\alpha}_i$ did not work out in practice in the models and tasks that we consider. Thus, we propose instead to supervise on $\hat{\beta}_i = \text{sigmoid}(\hat{a}_i)$ (similar to logistic attention [16]), where \hat{a}_i are the attention weights taken before the softmax that is applied within the Attention() function of Eq. 1. Due to the sparsity in the human annotation α , the traditional loss function would put too much emphasis on non-annotated tokens so we rather use the Jaccard loss function from [7] to avoid this bias, i.e.,

$$\mathcal{L}_a(\hat{\beta}, \alpha) = \frac{\hat{\beta}^T \alpha}{\sum_i \hat{\beta}_i + \sum_i \alpha_i - \hat{\beta}^T \alpha} \quad (4)$$

Note that $\hat{\beta}$ is only used in the loss function, c is still computed based on softmax attention map $\hat{\alpha}$.

5.3 Semi-supervision from Heuristics

Supervising with the reference human annotation aims at demonstrating whether supervision can be used to improve plausibility or not in an ideal scenario. This is however not realistic as human annotations are seldom available for this task and are costly to obtain. We thus investigate semi-supervision with annotations generated by simple heuristic rules. Indeed, [20] show that a simple heuristic attention map exploiting part-of-speech (POS) tags offers decent plausibility in the NLI task. The heuristic builds on the observation that verbs, nouns and adjectives (save for those in a small shortlist, such as auxiliary verbs) account for a fair amount of the tokens deemed as informative by human annotators⁵. In the e-SNLI dataset, 73.42% of the tokens in the human annotation fall in this category. To a slightly lesser extent, this is also observed on the HateXPlain and Yelp-Hat datasets used of text classification with more than 53% of the annotated tokens in this category.

We construct the heuristic map $\tilde{\alpha} = [\tilde{\alpha}_i]_{i=1}^L$ such that $\tilde{\alpha}_i = 0$ for tokens that are not nouns, verbs, adjectives, or stop-words, and reweight the remaining tokens based on the task. For classification tasks, the weight is the frequency of the token in the reference annotation. For NLI task, the weight is the sum of cosine similarities between the token and all tokens in the other sentence, applied equally to premise and hypothesis. Finally in all tasks, the heuristic map $\tilde{\alpha}$ is renormalized on a per-sentence basis, which transformed it into a probability vector. The Kullback-Leibler divergence

$$\mathcal{L}_a(\hat{\alpha}, \tilde{\alpha}) = \tilde{\alpha} \times [\log(\tilde{\alpha}) - \log(\hat{\alpha})] \quad (5)$$

is used to measure the loss between two probability vectors, $\tilde{\alpha}$ and $\hat{\alpha}$:

Our proposed heuristic for text classification has a limit as it indirectly relies on human annotations to weight each token, but one could make use of semantic lexicons such as SentiWordNet or VerbNet to craft heuristic weights for noun, verb, and adjective tokens.

6 Implementation and Training Parameters

To ensure consistency, the text data are pre-processed by tokenizing, lemmatizing, and lowercasing using the *spaCy* library. A unique vocabulary was generated for each dataset using the training set. All models reported in this study were initialized with the same GloVe embeddings (glove.42B.300d) and utilized ReLU activation functions with a softmax at the output of the classifier. The training settings were kept at their default configurations, including a learning rate of $lr = 1e - 3$ and a stabilizer of $\epsilon = 1e - 8$, as per community standards. To account for model variability, all runs were repeated three times.

Regarding evaluation, assessing the plausibility of attention maps faces three challenges: (1) the attention weights are continuous, (2) the magnitude of its

⁵ POS tags were detected with spaCy using the *en_core_web_sm* pipeline, which claims an accuracy of 97.2% in POS tagging.

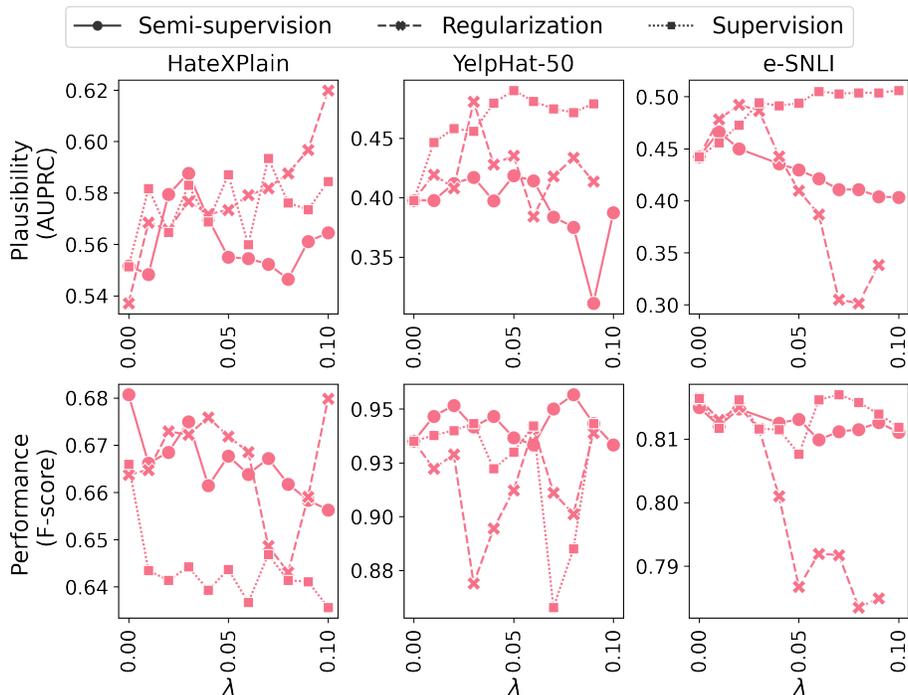


Fig. 2: Plausibility (AUPRC, top) and performance (F-score, bottom) on HateXPlain, YelpHat-50, and e-SNLI.

values depends on the sentence length and (3) only a few tokens are highlighted (class imbalance). To address these challenges, we apply a min-max scaler on the attention map and use the Area Under Recall/Precision curve (AURPC) as proposed in ERASER [6] to measure how close it is to human annotation. Additionally, we report Recall and Specificity by applying a threshold of 0.5 (the value is chosen following the ERASER benchmark [6]) for further insights.

Code to reproduce all experiments is available via [github](https://github.com/Kihansi95/Linkmedia_AttentionPlausibilityByConstraint)⁶.

7 Experimental Results

Firstly, the study investigates whether enhancing the plausibility of attention map is feasible without compromising classification performance. To answer this question, we evaluate three methods, namely, semi-supervision (solid line with round marker), regularization (dashed line with X marker), and supervision (dotted line with square marker) based on plausibility (AUPRC) and task performance (F-score) as reported in Figure 2. The figure showcases the evolution of the two metrics across three datasets under λ values ranging from $[0, 0.1]$ in a single bi-LSTM contextualization layer setting.

⁶ https://github.com/Kihansi95/Linkmedia_AttentionPlausibilityByConstraint

Across all three datasets, supervision and regularization show a consistent improvement in plausibility while preserving the classification performance. Although the semi-supervision shows little effect in HateXPlain, the technique shows improvement in YelpHat and e-SNLI. This suggests that effectiveness of the semi-supervision strategy in general depends on the specific characteristics of the input data, and users cannot always rely on it for improved performance. Poorer results on HateXplain can be explained by the fact that the heuristic is not good enough as explanation in HateXPlain depends highly on the context: In many cases, the same words could indicate either a hateful or a non-hateful meaning. This is not the case of YelpHat where sentiment words are rather unambiguous. The regularization approach turns out highly sensitive to λ , with performance getting hurt rapidly, while semi-supervision offers a more stable solution. Notice that supervision in e-SNLI leads to a loss of performance, due to the artifact in annotation instruction [10].

The impact of each constraint on attention maps in the NLI task is shown in Figure 3. When regularization is strengthened (λ increases), the attention maps progressively delete words that were initially highlighted in the baseline ($\lambda = 0$), which is the intended effect of regularization. However, when λ surpasses a certain threshold, attention maps become too concentrated on a few words, resulting in less plausible explanations. For instance, in Figure 3a, when $\lambda = 0.06$, the attention maps focus on only one word in each sentence ("with" in premise and "rug" in hypothesis), which renders the explanation implausible and negatively impacts performance (as seen in Figure 2, where the F-score drops from 0.815 to around 0.793).

In the case of supervision, attention maps gradually delete words from the baseline model, resulting in more plausible explanations that match the words highlighted by annotators. The constraint, however, does not ensure complete alignment with human annotations as shown in Figure 3b, where the attention map of $\lambda = 0.1$ does not select the words "two" and "on" to explain in hypothesis. In fact, with 10% of the loss devoted to making attention maps closer to human annotations, words selected by human annotation may not all be necessary for prediction. As can be seen from Figure 2, the attention maps cannot be constrained to be more similar to human annotations beyond $\lambda = 0.06$.

In semi-supervision, attention maps tend to retain the focus on words obtained from heuristic maps and do not impact performance. For instance, in the hypothesis attention map when $\lambda = 0.04$ (Figure 3c), the constraint deletes the words "two", "are", "on" and enforces attention values on "children" to match the heuristic map.

To confirm these observations, we report in Figure 4 recall and specificity of attention maps as a function of λ : while regularization encourages the selection of true positives (increase in recall), it tends to ignore some plausible words as indicated by the drop in specificity (we have more false negatives), as shown in e-SNLI and HateXPlain. This leads to a more conservative model that prefers to drop some words than highlight words that are not plausible. With supervision,

	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
$\lambda=0.01$	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
$\lambda=0.02$	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
$\lambda=0.06$	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

(a) regularization of attention

	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
$\lambda=0.03$	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
$\lambda=0.05$	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
$\lambda=0.1$	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

(b) supervision of attention

	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
$\lambda=0.01$	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
$\lambda=0.03$	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
$\lambda=0.04$	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

(c) semi-supervision of attention

Fig. 3: Examples of attention maps on one of the e-SNLI entailment pair.

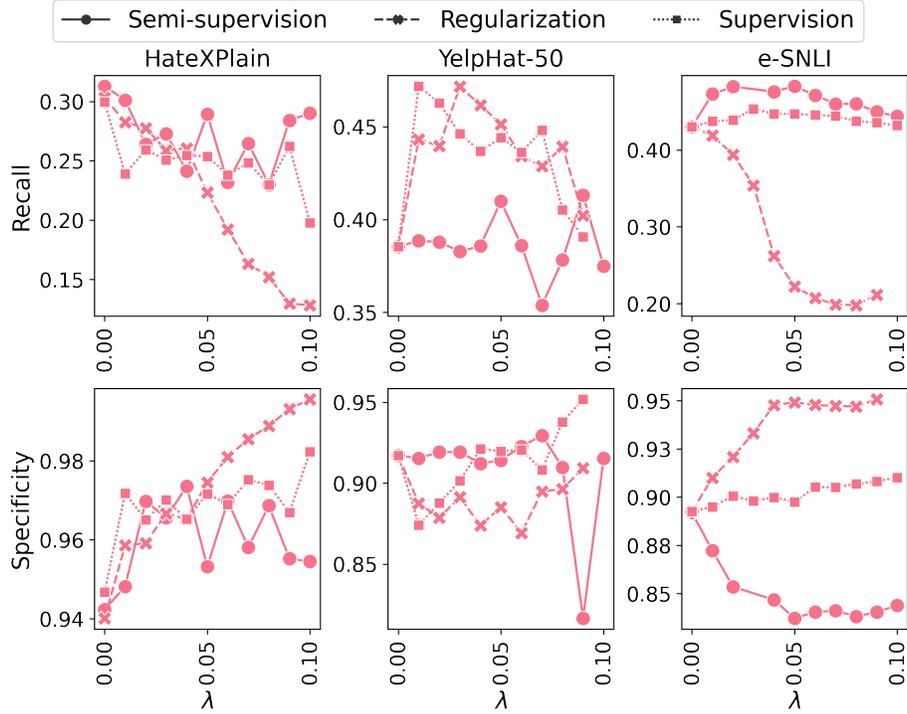


Fig. 4: Recall (top) and specificity (bottom) of attention map against annotation.

the model does the opposite and highlights more correct words by taking the risk of selecting more non-plausible words, thus increasing the false positive rate.

We further study the impact of the LSTM-based contextualization on plausibility. By stacking multiple layers of contextualization, a more semantically meaningful (or deeper) representation of each token can be obtained but also results in a uniform attention map across the entire sentence [8,9]. As regularization and semi-supervision can remove words from the attention map and make it sparse, we explore their potential to overcome the limitation of flat attention distribution in deeply contextualized models. Figure 5a reports results on the three tasks with one layer (in red (bullets)), three layers (green (crosses)), and five layers (blue (squares)) of bi-LSTMs contextualization, considering the three attention regularization strategies. Note that the scale of λ is different in each dataset. The effect of regularization depends on the task. In easy tasks (YelpHat), regularization does not yield improvement in plausibility. Surprisingly, the semi-supervision can actually improve in the case of the YelpHat corpus, overreaching supervision. In fact, the model’s plausibility converges to the AU-RPC of the heuristic map (0.6546 in YelpHat-50 and 0.5224 for HateXPlain). Although semi-supervision can offer a stable solution in classification tasks, its utility in complex tasks such as NLI and HateXPlain requires careful design. Finally, deeper contextualization with several bi-LSTM layers makes it harder to obtain a plausible attention map, no matter the technique. This suggests that

the contextualization by selectively keeping important features for classification suppressed other information that allow the attention layer to distinguish input tokens between them.

8 Conclusion

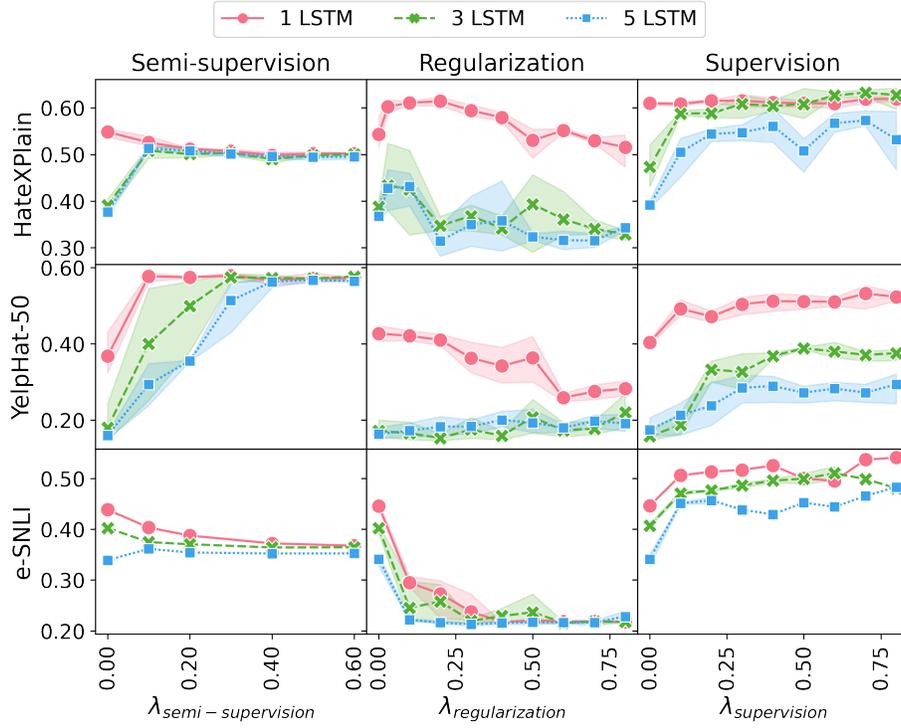
In this work, we compared three approaches to improve the plausibility of attention maps on top of RNN encoders at no extra cost, by adding an attention loss function to the classification loss. Regularization of the attention layer with an entropy criterion limits the words attended to in the model, marginally improving plausibility but risking the deletion of too many plausible tokens or focusing on the wrong ones. Supervision by human annotation encourages attention to focus on words it would not naturally attend to, but it may negatively impact the model’s performance depending on the quality and peculiarity of the annotation. Semi-supervision by a heuristic annotation of plausible tokens offers a valuable compromise by improving plausibility without sacrificing performance, but it is limited by the plausibility of the heuristic annotation. We show that the techniques for enforcing plausibility have a lesser impact than the depth of the contextualization with a bi-LSTM encoder. The plausibility of a model decreases with the number of bi-LSTM layers as model performance improves, regardless of attention regularization, suggesting that plausibility from attention in deep transformer-based models remains doubtful. This orients our future efforts to focus on creating an appropriate contextualized vector space that retains enough information to explain the model’s decision for humans through contextualization layers.

References

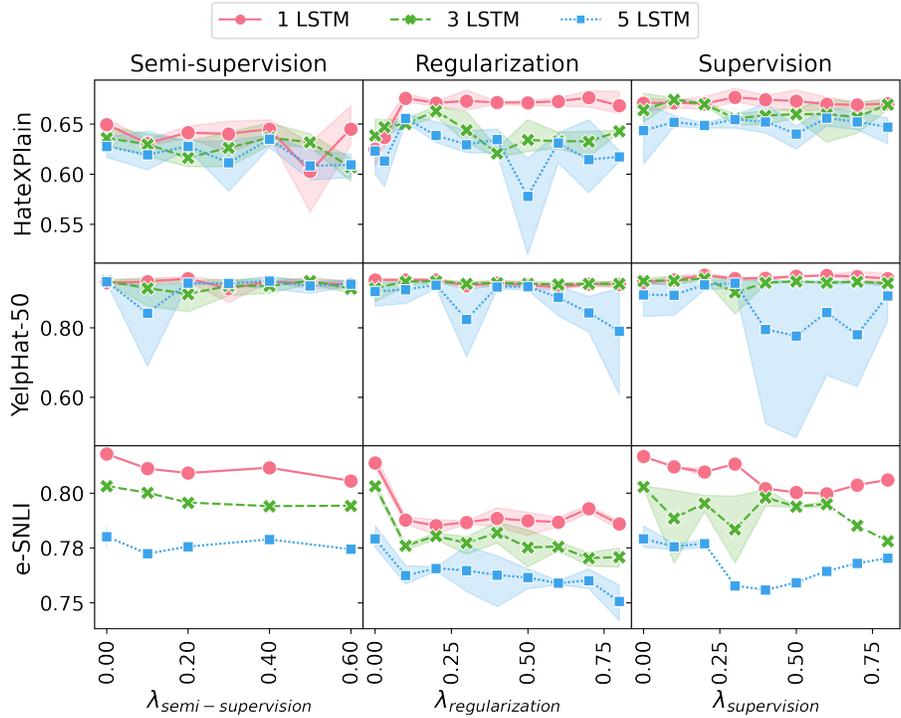
1. Bai, B., Liang, J., Zhang, G., Li, H., Bai, K., Wang, F.: Why attentions may not be interpretable? In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2021)
2. Bastings, J., Filippova, K.: The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (2020)
3. Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., Watrin, P.: Is Attention Explanation? An Introduction to the Debate. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (2022)
4. Camburu, O.M., Rocktäschel, T., Lukasiewicz, T., Blunsom, P.: E-SNLI: Natural Language Inference with Natural Language Explanations. In: Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (2018)
5. Chen, H., He, J., Narasimhan, K., Chen, D.: Can Rationalization Improve Robustness? In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2022)

6. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: A Benchmark to Evaluate Rationalized NLP Models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2020)
7. Duque-Arias, D., Velasco-Forero, S., Deschaud, J.E., Goulette, F., Serna, A., Decenci ere, E., Marcotegui, B.: On power Jaccard losses for semantic segmentation. In: 16th International Conference on Computer Vision Theory and Applications (2021)
8. Ethayarajh, K.: How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (2019)
9. Fosse, L., Nguyen, D.H., S ebillot, P., Gravier, G.: Une  tude statistique des plongements dans les mod eles transformers pour le fran ais. In: 29th Conference Traitement Automatique des Langues Naturelles (2022)
10. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., Smith, N.A.: Annotation Artifacts in Natural Language Inference Data. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (2018)
11. Jacovi, A., Goldberg, Y.: 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 (2020)
12. Jain, S., Wallace, B.C.: Attention is not Explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota (2019)
13. Jia, W., Dai, D., Xiao, X., Wu, H.: ARNOR: Attention Regularization based Noise Reduction for Distant Supervision Relation Classification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)
14. Lehman, E., DeYoung, J., Barzilay, R., Wallace, B.C.: Inferring Which Medical Treatments Work from Reports of Clinical Trials. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (2019)
15. Luong, T., Pham, H., Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2015)
16. Martins, A., Astudillo, R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In: Proceedings of the 33rd International Conference on Machine Learning (2016)
17. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)
18. McGuire, E.S., Tomuro, N.: Sentiment Analysis with Cognitive Attention Supervision. Proceedings of the Canadian Conference on Artificial Intelligence (2021)
19. Mohankumar, A.K., Nema, P., Narasimhan, S., Khapra, M.M., Srinivasan, B.V., Ravindran, B.: Towards Transparent and Explainable Attention Models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
20. Nguyen, D.H., Gravier, G., S ebillot, P.: A Study of the Plausibility of Attention between RNN Encoders in Natural Language Inference. In: 20th IEEE International Conference on Machine Learning and Applications (2021)

21. Nguyen, D.H., Gravier, G., Sébillot, P.: Filtrage et régularisation pour améliorer la plausibilité des poids d'attention dans la tâche d'inférence en langue naturelle. In: *Traitement Automatique Des Langues Naturelles* (2022)
22. Nguyen, M., Nguyen, T.H.: Who is Killed by Police: Introducing Supervised Attention for Hierarchical LSTMs. In: *Proceedings of the 27th International Conference on Computational Linguistics* (2018)
23. Ousidhoum, N., Zhao, X., Fang, T., Song, Y., Yeung, D.Y.: Probing Toxic Content in Large Pre-Trained Language Models. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (2021)
24. Paranjape, B., Joshi, M., Thickstun, J., Hajishirzi, H., Zettlemoyer, L.: An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (2020)
25. Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., Lipton, Z.C.: Learning to Deceive with Attention-Based Explanations. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020)
26. Sen, C., Hartvigsen, T., Yin, B., Kong, X., Rundensteiner, E.: Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020)
27. Serrano, S., Smith, N.A.: Is Attention Interpretable? In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019)
28. Sun, X., Lu, W.: Understanding Attention for Text Classification. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020)
29. Vashishth, S., Upadhyay, S., Tomar, G.S., Faruqui, M.: Attention Interpretability Across NLP Tasks. *CoRR* (2019)
30. Wiegrefe, S., Marasović, A.: Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmark*. vol. 1 (2021)
31. Wiegrefe, S., Pinter, Y.: Attention is not not Explanation. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (2019)



(a) The plausibility (in AUPRC) of the attention map.



(b) The task performance (F-Score).

Fig. 5: Plausibility and performance in 3 datasets, for 3 techniques, for 3 settings.