Interpretable Structure Induction Via Sparse Attention

Ben Peters[†] **Vlad Niculae**[†] and **André F.T. Martins**[†‡]
[†]Instituto de Telecomunicações, Lisbon, Portugal
[‡]Unbabel, Lisbon, Portugal
benzurdopeters@gmail.com, vlad@vene.ro, andre.martins@unbabel.com.

1 Introduction

Neural network methods are experiencing wide adoption in NLP, thanks to their empirical performance on many tasks. Modern neural architectures go way beyond simple feedforward and recurrent models: they are complex pipelines that perform soft, differentiable computation instead of discrete logic. Inspired by pioneering work by, e.g. Kohonen et al. (1981); Das et al. (1992); Schmidhuber (1992), such modern differentiable architectures include neural memories (Sukhbaatar et al., 2015) and attention mechanisms (Bahdanau et al., 2015). The price of such soft computing is the introduction of dense dependencies, which make it hard to disentangle the patterns that trigger a prediction. Our recent work on **sparse** and **structured** latent computation (Martins and Astudillo, 2016; Niculae and Blondel, 2017; Niculae et al., 2018; Malaviya et al., 2018) presents a promising avenue for enhancing interpretability of such neural pipelines. Through this extended abstract, we aim to discuss and explore the potential and impact of our methods.

The principle of *parsimony* suggests that simpler explanations are more plausible and interpretable. Our perspective is similar to prior work on regularizing model weights (Hastie et al., 2015), but with a twist: instead of model sparsity that tells us which "static" groups of variables are relevant for a task, we now have a "dynamic" form of sparsity that tells us, for a particular input object, where we should attend to produce a decision.

- **sparsity**: shrinking probabilities to zero to prune entire parts of the input when explaining a prediction (Martins and Astudillo, 2016);
- **regularization**: injecting prior assumptions, such as that neighbouring words should be fused together (Niculae and Blondel, 2017);
- **constraints**: constraining probabilities within lower and upper bounds, to prevent words from receiving too much or too little attention (Malaviya et al., 2018);
- **structure**: learning latent structure predictors (e.g. aligners or parsers), to induce a **compact representation** as a small, interpretable set of global structures (Niculae et al., 2018).

2 Attention Mechanisms

The key background for our work is the concept of attention. Attention mechanisms and memory networks are able to "point" to relevant items (e.g. words or pixels) that determine the final prediction, approximating a discrete choice (argmax) with a soft, differentiable one (softmax). Let $\mathbf{H} = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_L] \in \mathbb{R}^{D \times L}$ be a matrix whose columns are vectors encoding the L different choices (for example, words in a sentence). An attention mechanism maps a $\mathbf{H}$ and a control state $\boldsymbol{s}$ to a probability distribution $\boldsymbol{p} \in \triangle^{\mathrm{L}}$ over the L choices.[1] This can be split into (i) generating **scores** for each choice, e.g., $z_i = \boldsymbol{v}^\top \mathsf{tanh}(\mathbf{W}\boldsymbol{h}_i + \mathbf{U}\boldsymbol{s})$ for $i \in \{1, \dots, L\}$ and (ii) mapping the scores to a probability distribution. Common attention uses (Bahdanau et al., 2015; Luong et al., 2015) $\boldsymbol{p} = \mathsf{softmax}(\boldsymbol{z})$, i.e., $p_i = \exp(z_i)/\sum_j \exp(z_j)$. Since softmax is strictly positive, this leads to dense probability distributions. However, putting nonzero weight on every choice is not ideal for interpretability (Fig. 1, center); instead, we explore sparse selection, identifying a small set of choices responsible for a prediction. Niculae and Blondel (2017) proposed the general family

$$\Pi_\Omega(\boldsymbol{z}) = \operatorname*{argmax}_{\boldsymbol{p} \in \triangle^{\mathrm{L}}} \boldsymbol{z}^\top \boldsymbol{p} - \Omega(\boldsymbol{p}), \qquad (1)$$

recovering softmax for $\Omega(\boldsymbol{p}) = -\sum_j p_j \log p_j$.

[1] We denote by $\triangle^{\mathrm{L}} = \{\boldsymbol{p} \in \mathbb{R}^L \mid \sum_{i=1}^L p_i = 1, p_i \geq 0, \forall i\}$ the $(L-1)$-dimensional probability simplex.

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 365–367
Brussels, Belgium, November 1, 2018.

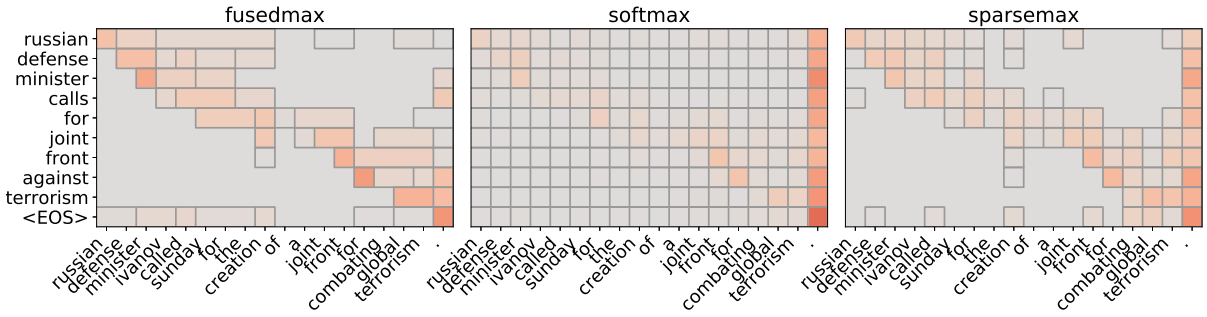

Figure 1: Attention weights for a sequence-to-sequence sentence compression instance. Traditional softmax attention (middle) yields dense weights, which are less interpretable than the sparse weights from sparsemax (right) or fusedmax (left); the latter further enhances interpretability by clustering probabilities of adjacent words. Image courtesy of Niculae and Blondel (2017).

Sparse attention. Martins and Astudillo (2016) proposed sparsemax, which replaces softmax with a Euclidean projection, remaining differentiable while also yielding sparse probabilities. This can be obtained by setting $\Omega = \frac{1}{2}\|\cdot\|_2^2$ in Eqn 1. The resulting probabilities are substantially more interpretable, as the contribution of irrelevant words is now shrunk to **exactly 0** (Fig. 1, right).

Regularized attention. Parsimony goes beyond sparsity: prior assumptions may encourage selecting groups or clusters with equal probability. Niculae and Blondel (2017) propose two linguistically-motivated regularized attention mechanisms: **fusedmax**, which tends to *group adjacent words together*, and **oscarmax**, which may *cluster non-adjacent words*, suitable for languages with flexible word order. Such mechanisms can select interpretable segments (Fig. 1, left).

Constrained attention. Some forms of parsimony must be strictly enforced using constraints, rather than simply encouraged via regularization. One such constraint is to add an upper bound to the cumulative attention an input variable may receive. This can be done using **constrained softmax** (Martins and Kreutzer, 2017) or its sparse analogue, **constrained sparsemax** (Malaviya et al., 2018). Constraining attention weights can be interpreted as specifying the *fertility* (Brown et al., 1993) of the alignments between the source and target, in machine translation.

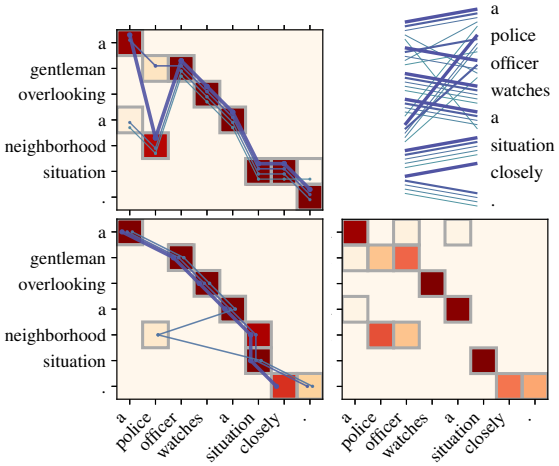

Figure 2: Structured alignment on SNLI (Niculae et al., 2018). The premise is on the y-axis, the hypothesis on the x-axis. Sequential alignment encourages monotonic alignments, matching induces a single symmetrical alignment.

3 Structured Attention

In this section, we consider *combinatorial* representations. Across application domains, but especially in NLP, many objects of interest can be represented by such structures: syntactic and dependency trees, sequential labellings, alignments. Allowing hidden layers to output structured representations can be valuable for modelling perspective but also for interpretability: discrete structures provide organized representations, in contrast to unstructured vectors of neuron activations.

SparseMAP (Niculae et al., 2018) allows handling discrete structures within end-to-end differentiable neural networks, able to automatically select only a few global structures. On natural language inference, for a word-to-word alignment joint attention mechanism, SparseMAP can induce structured alignments as illustrated in Fig. 2.

4 Conclusion

Building upon the principle of parsimony, we propose sparse, regularized, constrained and structured hidden layers. We seek to discuss the potentials of these strategies with an expert community on black-box interpretability.

Acknowledgments

This work was supported by the European Research Council (ERC StG DeepSPIN 758969) and by the Fundação para a Ciência e Tecnologia through contract UID/EEA/50008/2013.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Sreerupa Das, C Lee Giles, and Guo-Zheng Sun. 1992. Learning context-free grammars: Capabilities and limitations of a recurrent neural network with an external stack memory. In *Proceedings of The Fourteenth Annual Conference of Cognitive Science Society. Indiana University*, page 14.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Teuvo Kohonen, Erkki Oja, and Pekka Lehtio. 1981. Storage and processing of information in distributed associative memory systems. *Parallel Models of Associative Memory*, pages 129–167.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*.

Chaitanya Malaviya, Pedro Ferreira, and André F. T. Martins. 2018. Sparse and constrained attention for neural machine translation. In *Proc. of ACL*.

André F. T. Martins and Ramón Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proc. of ICML*.

André FT Martins and Julia Kreutzer. 2017. Learning what's easy: Fully differentiable neural easy-first taggers. In *Proc. of EMNLP*, pages 349–362.

Vlad Niculae and Mathieu Blondel. 2017. A regularized framework for sparse and structured neural attention. In *Proc. of NIPS*.

Vlad Niculae, André F. T. Martins, Mathieu Blondel, and Claire Cardie. 2018. SparseMAP: Differentiable sparse structured inference. In *Proc. of ICML*.

Jürgen Schmidhuber. 1992. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Proc. of NIPS*.