Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking

Michael Schlichtkrull, Nicola De Cao, and Ivan Titov

University of Amsterdam, University of Edinburgh

Overview

We introduce GraphMask, a novel interpretation technique for GNNs:

- We learn an erasure function that predicts, for every edge $\langle u, v\rangle$ at every layer k, whether that connection influences predictions.
- To enable gradient-based optimization for the erasure function, we rely on sparse stochastic gates (Louizos et al., 2018).
- We show that many existing methods are susceptible to hindsight bias, a failure mode for faithfulness.
- We use GraphMask to analyse real-world GNN models for two NLP tasks.

The Technical Details

GNNs pass messages through an input graph to produce predictions. A GNN can be defined through a message function M and an aggregation function A such that for the k-th layer:

$$m_{u,v}^{(k)} = M^{(k)}\left(h_u^{(k-1)}, h_v^{(k-1)}, r_{u,v}\right) \quad (1)$$

$$h_v^{(k)} = A^{(k)}\left(\left\{m_{u,v}^{(k)} : u \in \mathcal{N}(v)\right\}\right) \quad (2)$$

We search for edges which can be replaced with a learned baseline $b^{(k)}$ through hard binary choices:

$$m_{u,v}^{(k)} = z_{u,v}^{(k)} \cdot m_{u,v}^{(k)} + b^{(k)} \cdot (1 - z_{u,v}^{(k)}) \quad (3)$$

To avoid *hindsight bias*, where the optimizer aggressively prunes even useful edges by exploiting access to predicted labels, we compute $z_{u,v}^{(k)}$ through a small probe g_π learned once for every task:

$$z_{u,v}^{(k)} = g_\pi(h_u^{(k)}, h_v^{(k)}, m_{u,v}^{(k)}) \;, \quad (4)$$

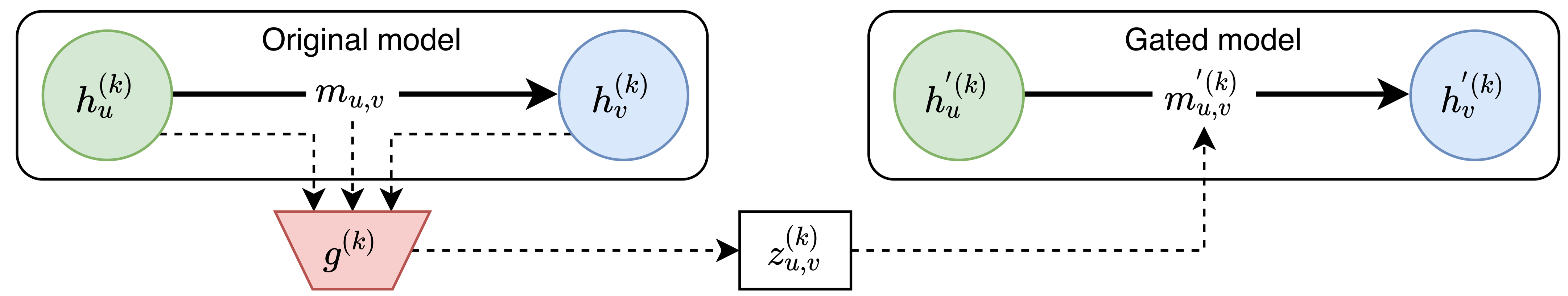

Figure 1: GraphMask uses vertex hidden states and messages at layer k (left) as input to a classifier g that predicts a mask $z^{(\ell)}$. We use this to mask the messages of the kth layer and re-compute the forward pass with modified node states (right). The classifier g is trained to mask as many hidden states as possible without changing the output of the gated model.

Synthetic Experiment

For real tasks, models and data are too complex for human gold standards (Jacovi & Goldberg, 2020). We illustrate GraphMask's resilience to hindsight bias on synthetic data:

- The input is star graphs with 6-12 edges randomly given one of 6 colours.
- Given an embedding of the central vertex and two colours x and y, predict whether there are more edges coloured x than y.
- The GNN is a single-layer R-GCN (Schlichtkrull et al., 2018), optimised to perfect performance.
- A faithful interpretability technique assigns positive attribution to all edges coloured x or y, and no attribution to other edges.
- Only GraphMask correctly analyses all test examples (see Table 1).

Method	Prec.	Recall	F_1
Erasure search*	100.0	16.7	28.6
Integrated Gradients	88.3	93.5	90.8
Information Bottleneck	55.3	51.5	52.6
GNNExplainer	100.0	16.8	28.7
Ours (non-amortized)	96.7	26.2	41.2
Ours (amortized)	98.8	100.0	**99.4**

Table 1: Comparison using the faithfulness gold standard on the synthetic task.

Analysing Real-world Models

- We analyse models for QA (De Cao et al., 2019) and SRL (Marcheggiani & Titov, 2017).
- GraphMask provides example-level *and* dataset-level analysis.
- Dropping edges marked superfluous by GraphMask does not significantly harm performance.

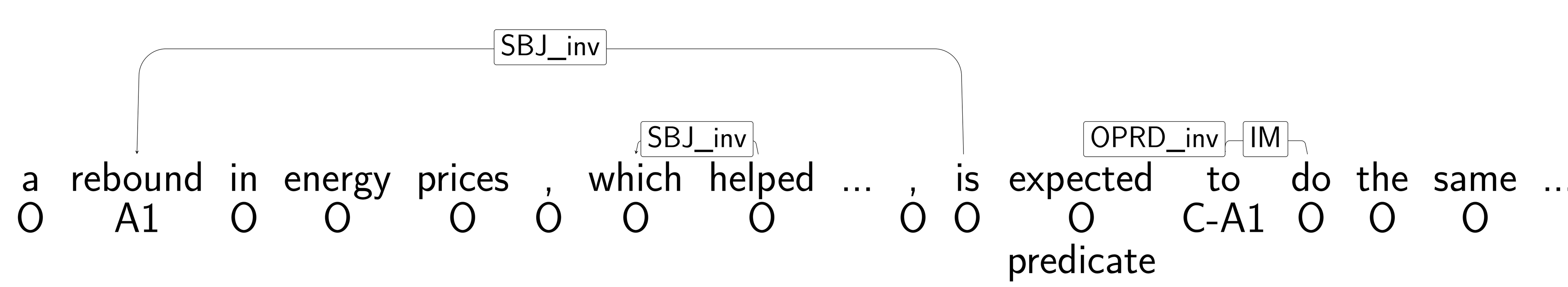

Figure 2: Example analysis on Marcheggiani & Titov's (2017) SRL system, using their GNN+LSTM model (superfluous arcs are excluded).

Edge Type	k = 0	k = 1	k = 2
MATCH (8.1%)	9.4%	11.1%	8.9%
DOC-BASED (13.2%)	5.9%	17.7%	10.7%
COREF (4.2%)	4.4%	0%	0%
COMPLEMENT (73.5%)	31.9%	0%	0%
Total (100%)	51.6%	28.8%	19.6%

Table 2: Dataset-level statistics of retained edges for De Cao et al.'s (2019) question answering GNN by layer (k) and type.

References

Nicola De Cao, Wilker Aziz, and Ivan Titov. Question answering by reasoning across documents with graph convolutional networks. In *NAACL*, pp. 2306–2317, 2019.

Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? *ACL*, 2020.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *ICLR*, 2018.

Diego Marcheggiani and Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In *EMNLP*, pp. 1507–1516, 2017.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC*, pp. 593–607, 2018.

Contact Information

: github.com/MichSchli/GraphMask

@: m.s.schlichtkrull@uva.nl

University of Amsterdam

The University of Edinburgh