A causal framework for explaining the predictions of black-box sequence-to-sequence models: Supplementary Material

David Alvarez-Melis and **Tommi S. Jaakkola**
CSAIL, MIT
{davidam, tommi}@csail.mit.edu

A Formulation of graph partitioning with uncertainty

The bipartite version of the graph partitioning problem with edge uncertainty considered by Fan et al. (2012) has the following form. Assume we want to partition U and V into K subsets each, say $\{U_i\}$ and $\{V_j\}$, with each U_i having cardinality in $[c^u_{min}, c^u_{max}]$ and each V_j in $[c^v_{min}, c^v_{max}]$. Let x^u_{ik} be the binary indicator of $u_i \in U_k$, and analogously for x^v_{jk} and v_j. In addition, we let y_{ij} be a binary variable which takes value 1 when u_i, v_j are in different corresponding subsets (i.e. $u_i \in U_k, v_j \in V_{k'}$ and $k \neq k'$). We can express the constraints of the problem as:

$$Y = \begin{cases} \sum_{k=1}^{K} x^v_{ik} = 1 & \forall i \quad (1) \\ \sum_{k=1}^{K} x^u_{jk} = 1 & \forall j \quad (2) \\ c^u_{min} \leq \sum_{i=1}^{N} x^v_{ik} \leq c^u_{max} & \forall i \quad (3) \\ c^v_{min} \leq \sum_{i=1}^{N} x^u_{jk} \leq c^v_{max} & \forall j \quad (4) \\ -y_{ij} - x^v_{ik} + x^u_{jk} \leq 0 & \forall i,j,k \quad (5) \\ -y_{ij} + x^v_{ik} - x^u_{jk} \leq 0 & \forall i,j,k \quad (6) \\ x^v_{ik}, x^u_{jk}, y_{ij} \in \{0,1\}, & \forall i,j,k \quad (7) \end{cases}$$

Constraints (1) and (2) enforce the fact that each s_i and t_j can belong to only one subset, (3) and (4) limit the size of the U_k and B_k to the specified ranges. On the other hand, (5) and (6) encode the definition of y_{ij}: if $y_{ij} = 0$ then $x^u_{ik} = x^v_{jk}$ for every k. A deterministic version of the bipartite graph partitioning problem which ignores edge uncertainty can be formulated as:

$$\min_{(x^u_{ik}, x^v_{ik}, y_{ij}) \in Y} \sum_{i=1}^{N} \sum_{j=1}^{M} w_{ij} y_{ij} \qquad (8)$$

The robust version of this problem proposed by Fan et al. (2012) incorporates edge uncertainty by adding the following term to the objective:

$$\max_{\substack{S: S \subseteq V, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{a}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{w}_{i_t, j_t} y_{i_t, j_t} \qquad (9)$$

where Γ is a parameter in $[0, |V|]$ which adjusts the robustness of the partition against the conservatism of the solution. This term essentially computes the maximal variance of a single cut $(S, V \setminus S)$ of size $|\Gamma|$. Thus, larger values of this parameter put more on the edge variance, at the cost of a more complex optimization problem. As shown by Fan et al. (2012) the objective can be brought back to a linear form by dualizing the term (9), resulting in the following formulation

$$\begin{aligned} \min \quad & \sum_{i=1}^{M} \sum_{j=1}^{M} w_{ij} y_{ij} + \Gamma p_0 + \sum_{(i,j) \in J} p_{ij} \\ \text{s.t.} \quad & p_0 + p_{ij} - \hat{a}_{ij} y_{ij} \geq 0, \quad (i,j) \in J \\ & p_{ij} \geq 0, \quad (i,j) \in J \\ & p_0 \geq 0 \\ & (x^u_{ik}, y^v_{jk}, y_{ij}) \in Y, \end{aligned} \qquad (10)$$

This is a mixed integer programming (MIP) problem, which can be solved with specialized packages, such as GUROBI.

B Details on optimization and training

Solving the mixed integer programming problem (10) to optimality can be prohibitive for large

Sampling temperature α ↓

Input:	Students said they looked forward to his class .	The part you play in making the news is very important .
Perturbations	Students said they looked forward to his class	The part with play in making the news is important .
	Students said they looked forward to his history .	The question you play in making the funding is a important .
	Students said they looked around to his class .	The part was created in making the news is very important .
	Some students said they really went to his class .	This part you play a place on it is very important .
	Students know they looked forward to his meal .	The one you play in making the news is very important .
	Students said they can go to that class .	These part also making newcomers taken at news is very important .
	You felt they looked forward to that class .	The terms you play in making the news is very important .
	Producers said they looked forward to his cities .	This part made play in making the band , is obvious .
	Note said they looked forward to his class .	The key you play in making the news is very important .
	Students said they tried thanks to the class ;	The part respect plans in making the pertinent survey is available .
	Why they said they looked out to his period .	In part were play in making the judgment , also important .
	Students said attended navigate to work as deep .	The issue met internationally in making the news is very important .
	What having they : visit to his language ?	In 50 interviews established in place the news is also important .
	Transition said they looked around the sense ."	The part to play in making and safe decision-making is necessary .
	What said they can miss them as too .	The order you play an making to not still unique .

Table 1: Samples generated by the English VAE perturbation model around two example input sentences for increasing scaling parameter α.

graphs. Since we are not interested in the exact value of the partition cost, we can settle for an approximate solution by relaxing the optimality gap tolerance. We observed that relaxing the absolute gap tolerance from the Gurobi default of 10^{-12} to 10^{-4} resulted in little minimal change in the solutions and a decrease in solve time of orders of magnitude. We added a run-time limit of 2 minutes for the optimization, though in all our experiments when never observed this limit being reached.

C Details on the variational autoencoder

For all experiments in Sections 5.3 through 5.5 we use the same variational autoencoder: a network with three layer-GRU encoder and decoder and a stacked three layer variational autoencoder connecting the last hidden state of the encoder and the first hidden state of the decoder. We use a dimension 500 for the hidden states of the GRUs and 400 for the latent states $\mathbf{z}$. We train it on a 10M sentence subset of the English side of the WMT14 translation task, with KLD and variance annealing, as described in the main text. We train for one full epoch with no KLD penalty and no noise term (i.e. decoding directly from the mean vector mu), and start variance annealing on the second epoch and KLD annealing on the 8th epoch. We train for 50 epochs, freezing the KLD annealing when the validation set perplexity deteriorates by more than a pre-specified threshold.

Algorithm 1 Variational autoencoder perturbation model for sequence-to-sequence prediction

```
procedure PERTURB($\mathbf{x}$)
    $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \text{ENCODE}(\mathbf{x})$
    for $i = 1$ to $N$ do
        $\tilde{\mathbf{z}}_i \sim \mathcal{N}(\boldsymbol{\mu}, \textbf{diag}(\alpha\boldsymbol{\sigma}))$
        $\tilde{\mathbf{x}}_i \leftarrow \text{DECODE}(\tilde{\mathbf{z}}_i)$
    end for
    return $\{(\tilde{\mathbf{x}}_i)\}_{i=1}^N$
end procedure
```

Once trained, the variational autoencoder is used as a subroutine of SOCRAT to generate perturbations as described in Algorithm 2. Given an input sentence $\mathbf{x}$, we use the encoder to obtain approximate posterior parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$, and then repeatedly sample latent representations from the a gaussian distribution with these parameters. The scaling parameter α constrains the locality of the space from which examples are drawn, by scaling the variance of the encoded representation's approximate posterior distribution. Larger values of α encourage samples to deviate further away from the mean encoding of the input $\boldsymbol{\mu}$, and thus more likely to result in diverse samples, at the cost of potentially less semantic coherence with the original input $\mathbf{x}$. In Table 1 we show example sentences generated by this perturbation model on two input sentences from the WMT14 dataset with increasing scaling value α.

D Black-box system specifications

The three systems used in the machine translation task in Section 5.3 are described below.

Azure's MT Service Via REST API calls to Microsoft's Translator Text service provided as part of Azure's cloud services.

Neural MT System A sequence-to-sequence model with attention trained with the OpenNMT library (Klein et al., 2017) on the WMT15 English-German translation task dataset. A pretrained model was obtained from http://www.opennmt.net/Models/. It has two layers, hidden state dimension 500 and was trained for 13 epochs.

A human A native German speaker, fluent in English, was given the perturbed English sentences and asked to translate them to German in one go. No additional instructions or context were provided, except that in cases where the source sentence is not directly translatable as is, it should be translated word-to-word to the extent possible. The human's German and English language models were trained for 28 and 16 years, respectively.

References

Neng Fan, Qipeng P. Zheng, and Panos M. Pardalos. 2012. Robust optimization of graph partitioning involving interval uncertainty. In *Theor. Comput. Sci.*, volume 447, pages 53–61.

G. Klein, Y. Kim, Y. Deng, J. Senellert, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.