Stochastic SparseMAP Notes

Sophia Sklaviadis

1 Stochastic Rationalizer Models

Problem statement

Explainability and model transparency: We are interested in making NN-based text classifiers interpretable by training jointly

i) a latent model that selects a rationale, i.e. a short and coherent extract of the input text, that serves as an explanation to the end user,

ii) and a classifier that learns from the words of the rationale alone.

Previous (most) related work: Lei et al. [2016], Bastings et al. [2019], Treviso and Martins [2020], Guerreiro and Martins [2021].

2 Latent Structure Models

Consider a text classification setting with

- input variable a sentence of length L: $\mathbf{x} = \langle x_1, ..., x_L \rangle \in \mathbb{R}^{D \times L}$ where D is the initial embedding size,
- a discrete structured latent variable $\mathbf{z}$ that consists of a combination of L binary parts that respect structural constraints and indicate which words are present in the rationale: $\mathbf{z} \in \mathcal{Z} \subset \{0, 1\}^L$ where $\mathcal{Z}$ is the set of feasible configurations $\mathbf{z}$ satisfying certain given constraints, and
- a categorical output variable Y, indicating the sentence's class:

$$Y|\mathbf{z}, \mathbf{x} \sim \mathrm{Cat}(\mathbf{x} \odot \mathbf{z}; \boldsymbol{\theta}).$$

Deterministic	**Probabilistic**
Identify an optimal $\hat{\mathbf{z}}(\mathbf{x}, \boldsymbol{\phi})$ and optimize	Assume $Z \sim p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\phi})$ and optimize
$\min_{\boldsymbol{\theta}, \boldsymbol{\phi}} - \log p(y \mid \mathbf{x}, \hat{\mathbf{z}}(\mathbf{x}, \boldsymbol{\phi}), \boldsymbol{\theta})$.	$\min_{\boldsymbol{\theta}, \boldsymbol{\phi}} - \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\phi})} \log p(y \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$.

3 Representation of structure z in a factor graph

Using a factor graph with

- variable nodes corresponding to tokens, and
- factor nodes encoding dependencies between the variables,

we can represent each structure $\mathbf{z}$ as a bit vector $\mathbf{a_z}$ that has

- one component per token indicating if it is part of $\mathbf{z}$, and
- additional components corresponding to factors that represent the instantiation of constraints.

Assume that the L components $\mathbf{z} = \langle z_1, ..., z_L \rangle$ that describe a rationale satisfy

- a global BUDGET constrain, i.e. a factor linked to all tokens imposing that at most B of them can be selected, and
- $L - 1$ pairwise factors for every pair of contiguous tokens.

The representation $\mathbf{a_z}$ is $d = 2L - 1$-dimensional bit vector, $d << |\mathcal{Z}|$,

$$\mathbf{a_z} \in \{0,1\}^{2L-1} \quad [\mathbf{a_z}]_i = \begin{cases} z_i & \text{for } i = 1, ..., L \\ z_{i-L} z_{i-L+1} & \text{for } L < i \leq 2L-1 \end{cases}$$

where $z_i = 1$ if token i is present in the rationale, else 0, and $\sum_{i=1}^{L} z_i \leq B$.

4 Marginal Polytope

Given a vector $\mathbf{s} = \langle s_i \rangle_{i=1}^{L}$ of scores for the unary parts $\langle z_i \rangle_{i=1}^{L}$ it is assumed that the score of the structure $\mathbf{z}$ is factored, so that structures with common parts share the corresponding scores

$$\begin{aligned} \text{score}(\mathbf{a_z}) &= \sum_{i=1}^{L} s_i z_i + \sum_{i=L+1}^{2L-1} r_i z_{i-L} z_{i-L+1} + \mathbb{1}_{\text{BUDGET}}{}^{1} \\ &= \boldsymbol{\eta}^{\mathrm{T}} \mathbf{a_z} \end{aligned}$$

where $r_i \geq 0$ are constants encouraging contiguity, and $\boldsymbol{\eta} = [\mathbf{s}, \mathbf{r}]^{\mathrm{T}}$. Note that a NN architecture maps the input to scores $s_i = s_i(\mathbf{x}; \boldsymbol{\phi})$, and $\boldsymbol{\phi}$ denotes collectively the NN parameters.

Denote by A the $d\times \mid \mathcal{Z} \mid$ matrix

- whose columns are the representations $\mathbf{a_z}$ of each possible $\mathbf{z}$,
- which specifies fully the structure of the problem.

Hence, the $\mid \mathcal{Z} \mid$-dim vector of all scores $\mathbf{S} = \begin{pmatrix} \mathrm{score}(\mathbf{a}_1) \\ \vdots \\ \mathrm{score}(\mathbf{a}_z) \\ \vdots \\ \mathrm{score}(\mathbf{a}_{|\mathcal{Z}|}) \end{pmatrix} = A^{\mathrm{T}}_{|\mathcal{Z}|\times d} \boldsymbol{\eta}_{d\times 1}$ can be expressed in terms of the common low dimensional parameter $\boldsymbol{\eta}$.

This factorization assumption provides a way to replace the simplex

$$\Delta^{|\mathcal{Z}|} = \{\mathbf{p} \ \in \mathbb{R}^{|\mathcal{Z}|}; \mathbf{1}^{\mathrm{T}}\mathbf{p} = 1, p \geq 0\}$$

where each component of $\mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_{\mathbf{z}} \\ \vdots \\ p_{|\mathcal{Z}|} \end{pmatrix}$ is the probability of a specific $\mathbf{z}$, with the $d = 2L - 1$-dimensional marginal polytope ($d << \mid \mathcal{Z} \mid$) defined as the convex hull:

$$\mathcal{M}_A = \mathrm{conv}\{a_1, ..., a_{|\mathcal{Z}|}\} = \{A_{d\times|\mathcal{Z}|}\mathbf{p}_{|\mathcal{Z}|\times 1}; \mathbf{p} \in \Delta^{|\mathcal{Z}|}\}.$$

5 Deterministic Structured Oracles

5.1 Marginal Inference

Any point $\boldsymbol{\mu} = A\mathbf{p}$ of the interior or $\mathcal{M}_A$ corresponds to a "canonical" parameter $\boldsymbol{\eta}$ that contains the scores and parametrizes the Gibbs distribution

$$\mathbf{p}^*_{\mathbf{z}} = P[Z = \mathbf{z}] \propto \exp(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{a}_{\mathbf{z}}).$$

$\mathbf{p}^*_{\mathbf{z}}$ is the structured equivalent of a component of the softmax that corresponds to the realization $\mathbf{z}$ of the random structure Z. So the full vector $\mathbf{p}^*$ is the solution of the Shannon negetropy

[1] For simplification of the exposition we do not include the Budget term in the subsequent notation.

regularized optimization problem

$$\mathbf{p}^* = \operatorname*{argmax}_{\mathbf{p}\in\Delta^{|\mathcal{Z}|}} \langle \boldsymbol{\eta}, A\mathbf{p}\rangle - \Omega(\mathbf{p}) \quad \text{where } \Omega(\mathbf{p}) = \sum_{\mathbf{z}=1}^{|\mathcal{Z}|} \mathbf{p}_\mathbf{z} \log \mathbf{p}_\mathbf{z}.$$

Denote

- by A_u the first L rows of A, and
- by $\boldsymbol{\mu}_u = A_u\mathbf{p}$ the first L elements of $\boldsymbol{\mu}$.

The marginal inference oracle is the $\boldsymbol{\mu}_u^*$ part of $\boldsymbol{\mu}^* = A\mathbf{p}^*$:

$$\begin{aligned}\boldsymbol{\mu}_u^* = \mathrm{Marginal}_A(\boldsymbol{\eta}) &= \operatorname*{argmax}_{\substack{\mathbf{p}\in\Delta^{|\mathcal{Z}|}\\ \boldsymbol{\mu}_u = A_u\mathbf{p}}} \boldsymbol{\eta}^{\mathrm{T}} A\mathbf{p} - \Omega(\mathbf{p}) \\ &= \operatorname*{argmax}_{\substack{\mathbf{p}\in\Delta^{|\mathcal{Z}|}\\ \boldsymbol{\mu}_u = A_u\mathbf{p}}} \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\mu} - \Omega_A(\boldsymbol{\mu})\end{aligned}$$

where the maximization is over $\boldsymbol{\mu}$ bu the unary part $\boldsymbol{\mu}_u$ is the return value of interest. Note that $\Omega_A(\boldsymbol{\mu}) = \Omega(\mathbf{p})$ does not have a closed form.

Hence the marginal inference oracle is $\boldsymbol{\mu}_u^* = \mathbb{E}_{\mathbf{p}^*} Z$ the unary part of the "mean" parameter of the Gibbs distribution, essentially the unique marginal distributions of the parts $\langle z_i\rangle_{i=1}^L$ that correspond to the Gibbs distribution (i.e. induced by its (score) parameter $\boldsymbol{\eta} = \begin{bmatrix}\mathbf{s}\\ \mathbf{r}\end{bmatrix}$).

5.2 SparseMAP

Regularizing by a squared l_2 penalty:

$$\begin{aligned}\mathrm{SparseMAP}(\boldsymbol{\eta}) &= \operatorname*{argmax}_{\mathbf{p}\in\Delta^{|\mathcal{Z}|}} \langle \boldsymbol{\eta}, A\mathbf{p}\rangle - \frac{1}{2} \| A_u\mathbf{p} \| \\ &= \operatorname*{argmax}_{\substack{\boldsymbol{\mu}\in\mathcal{M}_A\\ \boldsymbol{\mu}_u = A_u\mathbf{p}}} \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2} \| \boldsymbol{\mu}_u \|^2\end{aligned}$$

where again the return value of interest is the optimum $\boldsymbol{\mu}_u$.

5.3 MAP

$\mathrm{MAP}_A(\boldsymbol{\eta}) = \mathbf{z}^*$ where $\mathbf{z}^*$ is the first L components of

$$\mathbf{a}_\mathbf{z}^* = \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}} \boldsymbol{\eta}^{\mathrm{T}} \mathbf{a}_\mathbf{z}.$$

5.4 Surrogate gradients (Mihaylova et al. [2020])

Using as an optimal structure

$$\hat{\mathbf{z}}(\mathbf{x};\boldsymbol{\phi}) = \begin{cases} \text{SparseMAP}_A(\boldsymbol{\eta}) \\ \text{or} \\ \text{Marginal}_A(\boldsymbol{\eta}) \end{cases} \qquad \boldsymbol{\eta} = \begin{bmatrix} s(\mathbf{x};\boldsymbol{\phi}) \\ \mathbf{r} \end{bmatrix}$$

in the loss function of the Categorical output Y,

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} -\log p(y \mid \mathbf{x}, \hat{\mathbf{z}}(\mathbf{x},\boldsymbol{\phi}), \boldsymbol{\theta})$$

we can differentiate wrt $\boldsymbol{\phi}$.

6 Stochastic Latent Structures

Assuming a stochastic latent structure $Z \sim p(\cdot;\boldsymbol{\eta}(\mathbf{x},\boldsymbol{\phi}))$ we need to optimize the expected loss and compute

$$\nabla_{\boldsymbol{\phi}} \mathbb{E}_{\mathbf{z}\sim p_{\boldsymbol{\phi}}} -\log P(y \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\theta}).$$

Gumbel Max Trick: Motivation in the unstructured case: Let $Z \sim \text{Categorical}(\boldsymbol{\eta})$ then $Z = \text{argmax}_i(\boldsymbol{\eta} + G)_i$ where G is a Gumbel(0,1) r.v.

The Gumbel max trick provides an alternative representation of the Categorical r.v. Z as a transformation of a Gumbel r.v. G. Note that the Gumbel-max formulation enables rewriting $\mathbb{E}_{\mathbf{z}\sim p_{\boldsymbol{\phi}}}$ wrt the Gumbel r.v. $\mathbb{E}_{G\sim\text{Gumbel}}$, however, $\nabla_{\boldsymbol{\phi}}\mathbf{z}$ is still not differentiable.

Gumbel Softmax Trick: Approximate the discrete r.v. Z with the tempered softmax transformation of the Gumbel r.v. (Maddison et al. [2016], Jang et al. [2016]):

$$Z_\tau = \text{softmax}_\tau(\boldsymbol{\eta} + G) \xrightarrow{\tau\to 0} Z$$
$$Z_\tau \sim \text{Concrete}.$$

We can generalize the Gumbel Softmax trick to structured Z (Paulus et al. [2020]):

$$Z = \underset{\mathbf{p}\in\Delta^{|\mathcal{Z}|}}{\text{argmax}} \langle \boldsymbol{\eta} + G, A\mathbf{p}\rangle - \Omega(\mathbf{p})$$

where $\Omega(\mathbf{p})$ is the Shannon negetropy.

7 Mixed Latent Structured Random Variables

We assume that Z follows the Gaussian-SparseMAP distribution that can assign non-zero probability mass to the boundary of the marginal polytope $\mathcal{M}_A$ (Farinhas et al. [2021]). The distribution has the following generative story:

- generate an L-dim vector from the standard multivariate Normal $N \sim \mathrm{N}(\mathbf{0}, \mathrm{I}_{L\times L})$,
- perturb the scores of the L unary parts of the structure representation $\mathbf{a_z}$ so that its score is

$$\mathrm{score}(\mathbf{a_z}) = \begin{pmatrix} \mathbf{s} + \Sigma^{-1/2} N \\ \mathbf{r} \end{pmatrix}^{\mathrm{T}} \mathbf{a_z} = H^{\mathrm{T}} \mathbf{a_z}$$

  where Σ can capture possible correlation between the unary parts of the structure,
- $Z = \mathrm{SparseMAP}_A(H)$ is a sparse random vector that results from a transformation of the random variable H.

Hence,

$$\begin{aligned} Z = & \underset{\mathbf{p}\in\Delta^{|\mathcal{Z}|}}{\mathrm{argmax}} \ \langle H, A\mathbf{p}\rangle - \frac{1}{2} \| A_u \mathbf{p} \ \| \\ & \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_f \end{bmatrix} = A\mathbf{p} \\ & \boldsymbol{\mu}_u = A_u \mathbf{p} \\ = & \underset{\boldsymbol{\mu}\in\mathcal{M}_A}{\mathrm{argmax}} (\boldsymbol{s} + \Sigma^{-1/2} N)^{\mathrm{T}} \boldsymbol{\mu}_u + \boldsymbol{r}^{\mathrm{T}} \boldsymbol{\mu}_f - \frac{1}{2} \| \boldsymbol{\mu}_u \|^2 \end{aligned}$$

the random structure Z is the Euclidean projection on the marginal polytope $\mathcal{M}_A$ of the normally perturbed unary scores. (Recall that $\mathbf{s} = s(\mathbf{x}; \phi)$ depends on the input sentence $\mathbf{x}$ and the parameters ϕ.)

References

J. Bastings, W. Aziz, and I. Titov. Interpretable neural predictions with differentiable binary variables. *arXiv preprint arXiv:1905.08160*, 2019.

A. Farinhas, W. Aziz, V. Niculae, and A. F. Martins. Sparse communication via mixed distributions. *arXiv preprint arXiv:2108.02658*, 2021.

N. M. Guerreiro and A. F. Martins. Spectra: Sparse structured text rationalization. *arXiv preprint arXiv:2109.04552*, 2021.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

T. Mihaylova, V. Niculae, and A. F. Martins. Understanding the mechanics of spigot: Surrogate gradients for latent structure learning. *arXiv preprint arXiv:2010.02357*, 2020.

M. Paulus, D. Choi, D. Tarlow, A. Krause, and C. J. Maddison. Gradient estimation with stochastic softmax tricks. *Advances in Neural Information Processing Systems*, 33:5691–5704, 2020.

M. V. Treviso and A. F. Martins. The explanation game: Towards prediction explainability through sparse communication. *arXiv preprint arXiv:2004.13876*, 2020.