Stochastic SparseMAP

Mixed Sparse Structured Text Rationalization (SPECTRA)

Sophia Sklaviadis

July 27, 2024

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].

Latent Structure Models

Consider a text classification setting with

- a sentence of length L as input variable: $\mathbf{x} = \langle x_1, ..., x_L \rangle \in \mathbb{R}^{D \times L}$ 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: $\mathbf{z} \in \mathcal{Z} \subset \{0, 1\}^{L+\text{constraints}}$ 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:

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

Latent Structure Models

Deterministic

Identify an optimal $\hat{\mathbf{z}}(\mathbf{x}, \phi)$ and optimize

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

Probabilistic

Assume $Z \sim p(\mathbf{z} \mid \mathbf{x}, \phi)$ and optimize

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

Representation of structure $\mathbf{z}$ 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 $\mathbf{a}_{\mathbf{z}}$ 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...

Representation of structure $\mathbf{z}$ in a factor graph

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 a $d = 2L - 1$-dimensional bit vector, with $d <<| \mathcal{Z} |$,

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

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

Marginal Polytope

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

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

where $r_i \geq 0$ are constants encouraging contiguity, and $\boldsymbol{\eta} = [\mathbf{s}, \mathbf{r}]^{\mathsf{T}}$.

Note that a NN architecture maps the input to scores $s_i = s_i(\mathbf{x}; \phi)$, and ϕ denotes collectively the NN parameters.

---

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

Marginal Polytope

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

$$\mathbf{S} = \begin{pmatrix} \text{score}(\mathbf{a}_1) \\ \vdots \\ \text{score}(\mathbf{a}_z) \\ \vdots \\ \text{score}(\mathbf{a}_{|\mathcal{Z}|}) \end{pmatrix} = A^{\mathsf{T}}_{|\mathcal{Z}| \times d} \boldsymbol{\eta}_{d \times 1}$$ can be expressed in terms of the common low dimensional parameter $\boldsymbol{\eta}$.

Marginal Polytope

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

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 $\mathbf{z}$.

The $d = 2L - 1$-dimensional marginal polytope ($d <<| \mathcal{Z} |$) defined as the convex hull:

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

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:

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

$\mathbf{p}^*_{\mathbf{z}}$ 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 regularized optimization problem (i.e. the variational formulation of a CRF):

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

Deterministic Structured Oracles

Marginal Inference

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 $\boldsymbol{\mu}^* = A\mathbf{p}^*$:

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

where the maximization is over $\boldsymbol{\mu}$ but 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 (Niculae et al. [2018]).

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. the distribution induced by the (score) parameter $\boldsymbol{\eta} = \begin{bmatrix} \mathbf{s} \\ \mathbf{r} \end{bmatrix}$).

Deterministic Structured Oracles

MAP & SparseMAP: Regularizing by a squared l_2 penalty

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

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

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

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

Using as an optimal structure

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

in the loss function of the Categorical output Y,

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

we can differentiate wrt ϕ.[2]

[2] Mihaylova et al. [2020]

Stochastic Latent Structures

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

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

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 $\mathbb{E}_{\mathbf{z} \sim p_{\phi}}$ wrt the Gumbel r.v. $\mathbb{E}_{G \sim \text{Gumbel}}$, however, $\nabla_{\phi} \mathbf{z}$ is still not differentiable.

Stochastic Latent Structures

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.

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

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

  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.

Mixed Latent Structured Random Variables

Hence,

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

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.)

Experiments

Tuning on BeerAdvocate

(Force)Budget=10, Temperature=0.05, downstream MSE is < 0.02 across all experiments.

Transition	Spectra	Perturb 0.001*Gumbel(0,1)	Perturb 0.01*Gumbel(0,1)
0.001	0.61 (min=0.56/ max=0.68) Guerreiro and Martins [2021]	-	-
0.05	0.61	-	-
0.1	0.6117	-	-
0.5	0.635	-	-
1	0.6533	0.70718 (min=0.6729/ max=0.7209)	**0.7117** (min=0.6984/ max=0.728)
1.5	0.6364	?	?

Table: Aspect0, F1 scores based on human annotations.

Experiments

Tuning on BeerAdvocate

Transition	Spectra	Perturb N(0,1)	Diag Cov (learn scores for $\log(\sigma_i)_1^L$)	Hadamard (learned scores $\times$ distance toepliz $\rightarrow$ Normal cov)
1.5	**0.70866** (0.6978/0.7281)	0.63452 (0.5834/0.663)	0.68268 (0.6458/0.7413)	0.677625 (0.6612/0.6968)
1	0.6801 (0.5801/0.7186)	0.64132 (0.626/0.6639)	68638 (0.6556/0.7303)	0.6559825 (0.6348/0.7032)
0.5	0.61838 (0.4129/0.7158)	0.61836 (0.6087/0.629)	0.67396 (0.6424/0.6951)	0.63468 (0.5914/0.6721)
0.1	0.55624 (0.5614/0.6633)	0.63934 (0.6087/0.6646)	0.65266 (0.6154/0.6851)	0.63865 (0.6143/0.6634)
0.01	0.54496 (0.4888/0.6031)	0.64322 (0.6226/0.6629)	0.6572 (0.631/0.6797)	0.642525 (0.6218/0.6764)
0.001	0.51536 (0.4776/0.5479)	0.63962 (0.6098/0.6382)	0.63962 (0.6148/0.6745)	0.64352 (0.6222/0.6693)
0	0.52728 (0.4802/0.5744)	0.62842 (0.5989/0.6486)	0.63834 (0.5832/0.6638)	0.64584 (0.6113/0.6857)

Table: Aspect 1, F1 scores based on human annotations.

Experiments

Tuning on BeerAdvocate

Transition	Spectra	N(0,1)	0.001*G(0,1)	0.01*G(0,1)
1	0.6801 (0.5801/0.7186)	0.64132 (0.626/0.6639)	0.70952 (0.6729/0.7246)	0.7122 (0.728/0.7009)
0.5	0.61838 (0.4129/0.7158)	0.61836 (0.6087/0.629)	0.71314 (0.7226/0.7003)	**0.7151** (0.6885/0.7348)
0.001	0.51536 (0.4776/0.5479)	0.63962 (0.6098/0.6382)	0.64686 (0.6279/0.668)	0.63064 (0.6122/0.6513)

	0.1*G(0,1)	0.5*G(0,1)	1*G(0,1)	1.5*G(0,1)
1	0.68644 (0.6768/0.6902)	0.63594 (0.6157/0.6695)	0.59598 (0.6182/0.5444)	0.60566 (0.5772/0.6211)
0.001	0.65962 (0.6566/0.663)	0.60422 (0.5832/0.6386)	0.58788 (0.5672/0.5997)	0.57998 (0.5241/0.6081)

Table: Aspect 1, F1 scores based on human annotations.

Bibliography I

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.

Bibliography II

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.

V. Niculae, A. Martins, M. Blondel, and C. Cardie. Sparsemap: Differentiable sparse structured inference. In *International Conference on Machine Learning*, pages 3799–3808. PMLR, 2018.

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.