Differentiable Perturb-and-Parse: Semi-Supervised Parsing with a Structured Variational Autoencoder

Corro & Titov, ICLR 2019

Tom Effland

April 24, 2019

Overview

- First, we'll discuss the core idea of the paper, relaxed perturb-and-MAP, abstracting over parsing-specific details - this is what can actually be of use to the class.

- Then we can discuss the idea's application to parsing, if people care. (But we still won't discuss the Eisner algorithm.)

General Treatment

Problem Statement

In NLP, we often want model some discrete structure given an input observation. (Let's call this inference)

- Observation $x \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots$,
  (e.g., $\mathcal{X}$ = set of variable length discrete sequences in vocab)
- Inferred structure $y \in \mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots$,
  (e.g., $\mathcal{Y}$ = set discrete label sequences, discrete segmentation, grammar derivation)
- Want to learn: $p_\phi(y|x) : \mathcal{X} \to \Delta_{\mathcal{Y}}$
- Want to predict: $\hat{y} \leftarrow \arg\max_{y' \in \mathcal{Y}} p_\phi(y'|x)$

Modeling Inference: Motivating Approaches

What are ways people approach this?

Why not just use a tractable generative model? (e.g., HMM or PCFG)
$p_\theta(y_{1:M}|x) \propto p_\theta(y_{1:M}, x)$

- They are too restrictive in the modeling assumptions
- $\Rightarrow$ They underperform, discriminative (conditional) models work better

Ok, use directed (locally normalized) conditional model: $p_\phi(y_{1:M}|x) = \prod_{i=1}^{M} p_\phi(y_i|y_{<i}, x)$

- No longer need independence assumptions on inputs (think naive bayes vs. logistic regression) or outputs for that matter
- But there's the problem when predicting structures:
  directed conditional models have a limited ability for later decisions to revise earlier ones, especially with beam-search

Modeling Inference: CRFs

Conditional Random Fields: Structure is influenced bidirectionally

- If your model decoding order doesn't reflect a causal process, undirected model is probably more appropriate

Instead of local normalization:

$$p_\phi(y_{1:M}|x) = \prod_{i=1}^{M} p_\phi(y_i|y_{<i}, x) = \prod_{i=1}^{M} \frac{\exp\{\phi(y_i|y_{<i}, x)\}}{\sum_{y_i'} \exp\{\phi(y_i'|y_{<i}, x)\}}$$

Global normalization:

$$p_\phi(y_{1:M}|x) = \frac{\prod_{i=1}^{M} \exp\{\phi(y_i|y_{<i}, x)\}}{\underbrace{\sum_{y' \in \mathcal{Y}} \prod_{i=1}^{M} \exp\{\phi(y_i'|y_{<i}, x)\}}_{Z_\phi(x)}}$$

When ϕ factor graph for y is a tree, $Z_\phi(x)$ is computable in polynomial time with dynamic programming (e.g., forward-backward, **sum-product**)

Semi-Supervised Learning

For semi-supervised learning, generative models are an attractive solution for learning on additional unsupervised data

- Principled: optimize marginal likelihood
- Prior can impose regularization
- Appropriate generative model can provide useful signal for inference

Embed our CRF inference model as the amortized approximate posterior in an VI setup! New setup, unsupervised case:

$$p_\theta(x)p_\theta(y|x), \quad q_\phi(y|x) \leftarrow p_\phi(y|x)$$
$$\log p_\theta(x) \geq \mathbb{E}_{y\sim q_\phi}[\log p_\theta(x|y)] - KL(q_\phi||p_\theta(y))$$

One MAJOR problem though, the usual one:

- What about $\nabla_\phi \mathbb{E}_{y\sim q_\phi}[\log p_\theta(x|y)]$?
- REINFORCE is often very poorly behaved in these situations

Remember Gumbel-Softmax?

Can draw a sample from a categorical with

$$\tilde{y} = \arg\max_{y'} \{\log \pi_{y'} + \gamma_{y'}\}, \quad \gamma_{y'} \sim \mathcal{G}(0,1)$$

and can draw a "relaxed" sample with

$$\tilde{y}_r = \frac{\exp\{\log \pi_{y'} + \gamma_{y'}\}}{\sum_{y'} \exp\{\log \pi_{y'} + \gamma_{y'}\}}, \quad \gamma_{y'} \sim \mathcal{G}(0,1)$$

Relaxed Perturb-and-MAP

(Gumbel-Softmax for tractable CRFs)

Can draw a sample from a CRF using Perturb-and-MAP [Papandreou and Yuille '11]

$$\tilde{y} = \arg\max_{y \in \mathcal{Y}} q_{\phi+\tilde{\gamma}}(y|x)$$

Gradient of log partition function is the joint distribution [Eisner '16, Mencsh and Blondel '18]

$$\nabla \log Z_\phi(x) = q_\phi(y|x)$$

and it's zero temperature limit is the MAP estimate (as one-hots)

$$\nabla \log Z_\phi(x;\tau) = q_\phi(y|x;\tau) \xrightarrow{\tau \to 0} \arg\max_{y \in \mathcal{Y}} q_\phi(y|x)$$

So we have that the gradient of the perturbed partition function converges to a sample as the temp approaches zero

$$\nabla \log Z_{\phi+\tilde{\gamma}}(x;\tau) = q_{\phi+\tilde{\gamma}}(y|x;\tau) \xrightarrow{\tau \to 0} \arg\max_{y \in \mathcal{Y}} q_{\phi+\tilde{\gamma}}(y|x) = \tilde{y}$$

Takeaway: Perturb and temper potentials, then run inference
$\Rightarrow$ Marginals are a relaxed sample from the CRF

Application to Dependency Parsing

Dependency Parsing

Dependency grammar is a formalism of *syntax* for how words modify each other in a sentence

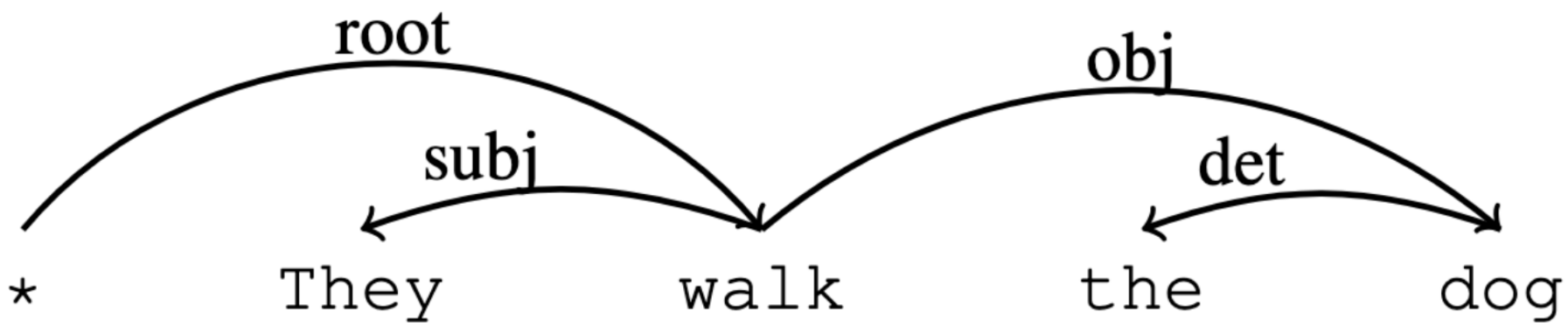


Figure: Example dependency structure

It can be represented as an adjacency matrix A (ignoring labels) where columns $A_{\cdot,j}$ sum to 1.

An entry at $A_{i,j} = 1$ if the edge $x_i \to x_j$ exists.

Trees are also *projective* – no crossing edges.

Inference Model

Model can be viewed as a CRF, with a potential for each cell in the matrix plus a special fully connected factor that ensures the tree constraints.

The model potential scores for some valid tree T are

$$q_\phi(T|x) = \frac{\exp\{\phi(T, W(x))\}}{\sum_{T' \in \mathcal{T}} \exp\{\phi(T', W(x)\}}, \qquad \phi(T, W) = \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} T_{ij} W_{ij}(x)$$

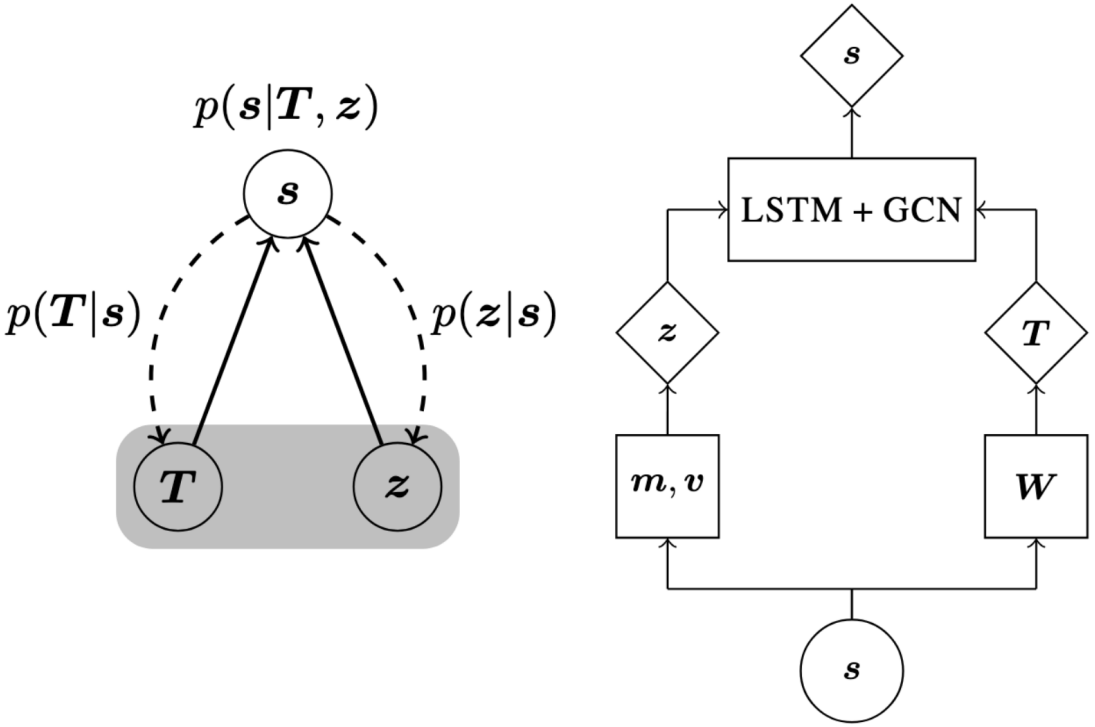
Projectivity of the tree implies that the argmax and marginals can be inferred in $O(n^3)$ (Eisner's Algorithm)

Also have a latent sentence vector $q_\phi(z|x) \sim \mathcal{N}(\mu(x), \sigma^2(x))$ from sentence encoding

Generation

Assume a generative model, for known sentence length n :

- $z \sim p(z|n) = \mathcal{N}(0, I_d)$
- $T \sim p(T|n)$ ▷ Uniform distribution of rooted projective tree matrices
- $x_{1:n} \sim p_\theta(x_{1:n}|z, T, n) = \prod_{i=1}^{n} p_\theta(x_i|x_{<i}, z, T_{\leq i, \leq i})$



(a) Probabilistic model (b) Computation Graph

Semi-Supervised Learning

They use the standard Semi-supervised VAE objective:

$$\mathcal{J}_L(\theta, \phi; x, T) = \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|T, z)]}_{\mathbb{E}_\epsilon[\log p_\theta(x|T, z_\phi(x,\epsilon))]} - \alpha_z KL(q_\phi(z|x)||p(z)) + \log q_\phi(T|x)$$

$$\mathcal{J}_U(\theta, \phi; x) = \underbrace{\mathbb{E}_{q_\phi(z,T|x)}[\log p_\theta(x|T, z)]}_{\mathbb{E}_{P,\epsilon}[\log p_\theta(x|T_\phi(x,P;\tau), z_\phi(x,\epsilon))]} - \alpha_z KL(q_\phi(z|x)||p(z)) - \alpha_T KL(q_\phi(T|x)||p(T))$$

$$\mathcal{L}(\theta, \phi; \mathcal{D}_L, \mathcal{D}_U) = \mathbb{E}_{(x,T)\sim\mathcal{D}_L}[\mathcal{J}_L] + \mathbb{E}_{(x)\sim\mathcal{D}_U}[\mathcal{J}_U]$$

Note: Strange balancing of objectives – OK, due to a combo of the datasets not being too heavily imbalanced towards $\mathcal{D}_U$ and the small KL weights reducing the impact of unsupervised regularization

Experiments

Test the state-of-the-art parsing architecture on three standard datatsets:

	Labeled	Unlabeled
English	3984	35848
French	1476	13280
Swedish	4880	5331

Figure: Dataset info.

	English	French	Swedish
Supervised	88.79 / 84.74	84.09 / 77.58	86.59 / 78.95
VAE w. z	89.39 / 85.44	84.43 / 77.89	86.92 / 80.01
VAE w/o z	89.50 / 85.48	84.69 / 78.49	86.97 / 79.80
Kipperwasser & Goldberg	89.88 / 86.49	84.30 / 77.83	86.93 / 80.12

Figure: Results: Edge Precision / Recall. Considerable improvement from unlabeled data, approaches fully supervised performance w/ 10% of the data

Worth noting: have to set KL weight for T to 0 and z to $.1$

Deep Generative Models - Natural Langauge Processing

Adversarially Regularized Autoencoders

Motivation

"learning deep latent variable models for text sequences has been a significantly more challenging empirical problem than for images."

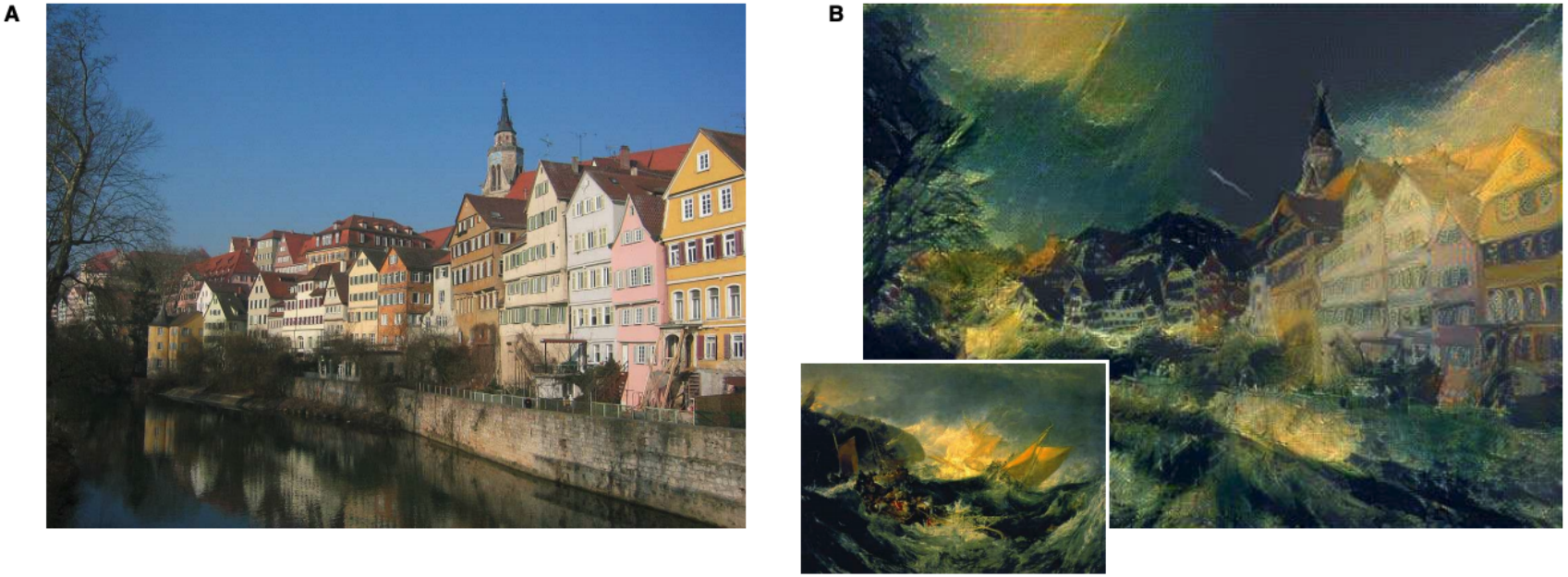
- The discrete nature, the multi-mode distribution.
- The optimization issue, the posterior collapse problem.
- The multi-task learning nature.

Yes it's hard, but first, what do we want from a generative model for text?

Goals of Text Generation

- The goals we want to achieve:
  - A generative model with good reconstruction.
    - current: not as good in text compared to image, even with the most SOTA architecture and engineering.
  - A conditional model allowing us to utilize explicit semantic signal (linguistic attributes like sentiment.)
    - current: the condition signal only affects word choice.
  - A latent model allowing us to manipulate generation. (for diversified generated texts).
    - current: challenging to learn a meaningful latent representation.
- This paper:
  - An ensamble of all techniques we have discussed: disentangling, VAEs, WGANs
  - The close-to-real task: text style transfer.

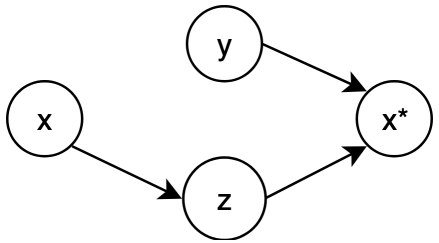
Recall: Image style transfer



- Given a image, transfer its style, maintain its content.

The Text Style Transfer Task

- Similar to image: given a sentence of a certain style, transfer the style, perserve the *content*.
- Given a restaurant review of negative sentiment, transfer the sentiment to positive. Preserve the restaurant.
- E.g. I hate the *food in this restaurant* $\rightarrow$ I love the *food in this restaurant*
- "style": sentiment, topic, gender, authorship, .etc
- Intuition - disentangling the style from the content:



  - $x \rightarrow y, z$
  - y: the style label - supervision signal
  - z: the latent content - adversarial regularization, style free
- A VAE model with disentangling. But, the posterior collapse ...

The Posterior Collapse Problem

- Back to the ELBO objective:

$$L(\phi, \psi) = \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\psi(x|z)]}_{\text{hard to fit, 20+ epochs}} - \underbrace{KL(q_\phi(z|x)||p(z))}_{\text{easy to fit, 1 epoch}} \tag{1}$$

- Auto-regressive decoder:

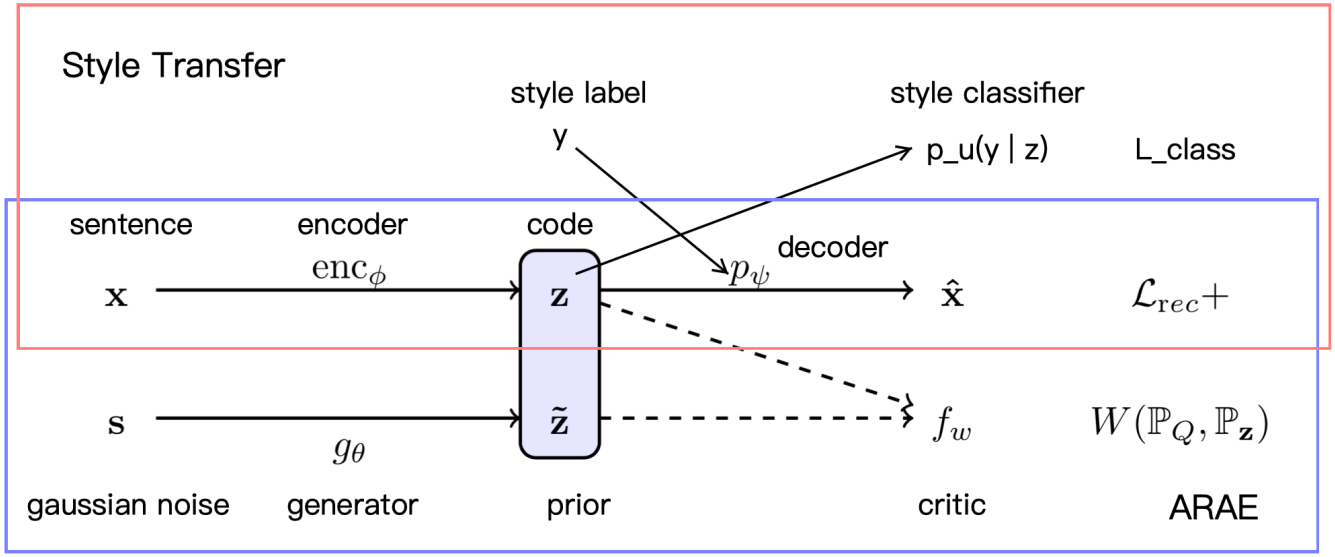
$$p_\psi(x|z) = \prod_{i=1}^{m} p(x_i|x_{1,\ldots,i-1}, z) \tag{2}$$

- When $KL(q_\phi(z|x)||p(z)) = 0$:
  - $q_\phi(z|x) = N(0, I)$
  - Sample from $q_\phi(z|x) =$ random gaussian noise. $\rightarrow$ Encoder meaningless.
  - VAE model $\rightarrow$ MLE model with the decoder. $\rightarrow$ An RNN language model.
  - $p_\psi(x|z) = \prod_{i=1}^{m} p(x_i|x_{1,\ldots,i-1}, z) \rightarrow \prod_{i=1}^{m} p(x_i|x_{1,\ldots,i-1})$
- Suggestions?
- This paper: a learned prior

An Adversarially Learned Prior

- Empirical evidence: posterior collapse is sensitive to the choice of prior
  - Von Mises-Fisher better than Gaussian.
  - Sidenote - vMF: places mass on the surface of the unit hypersphere; reparameterizable
- Want: learn a flexiable implicit prior (that might help mitigate the posterior collapse)
  - Generate the prior: $\mathbb{P}_z(\tilde{z})$, $\tilde{z} = g_\theta(s)$, $s \sim N(0, I)$
  - Encode the sentence: $\mathbb{P}_Q(z)$, $z = \text{enc}_\phi(x)$
  - Adversarially learn the EMD: $W(\mathbb{P}_z, \mathbb{P}_Q)$
- "Intuitively: this method aims to provide smoother hidden encoding for discrete sequences with a flexible prior."
- Empiricially: lead to better performance.
- Theoretically: seems no explanation. Suggestions?

Ensemble the Full ARAE



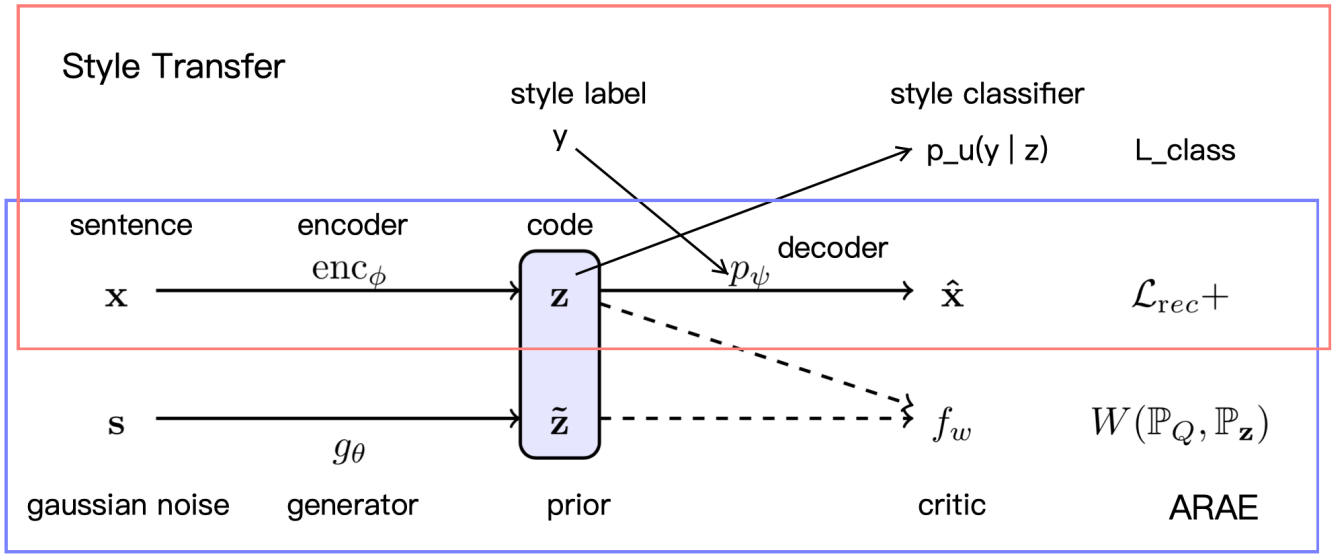
The disentangling model for style transfer:

- The latent content: $z = \text{enc}_\phi(x)$. (the posterior collapse)
- The style classifier: $p_u(y|z)$

The ARAE - adversarially learned content prior:

- The prior generator: $\mathbb{P}_z(\tilde{z})$, $\tilde{z} = g_\theta(s)$
- The encoder: $\mathbb{P}_Q(z)$, $z = \text{enc}_\phi(x)$. Shared with the disentangling.
- The critic: $f_w(z), f_w(\tilde{z})$
- The decoder: $p_\psi(x|z)$

How the Model Align with Our Goal?



- Look back to our original goal:
  - A generative model with good reconstruction. $\rightarrow \mathcal{L}_{\text{rec}}$
  - A conditional model allowing us to utilize explicit semantic signal. $\rightarrow \mathcal{L}_{\text{class}}$
  - A latent model allowing us to manipulate generation. $\rightarrow W(\mathbb{P}_z, \mathbb{P}_Q)$
- ... Messy model, any theoretical framework / support?

Theoretical Framework

Theorem

Let $G_\psi : \mathcal{Z} \to \mathcal{X}$ be a deterministic function (parameterized by ψ) from the latent space $\mathcal{Z}$ to data space $\mathcal{X}$ that induces a dirac distribution $\mathbb{P}_\psi(x|z)$ on $\mathcal{X}$, i.e. $P_\psi(x|z) = \mathbb{1}\{x = G_\psi(z)\}$. Let $Q(z|x)$ be any conditional distribution on $\mathcal{Z}$ with density $p_Q(z|x)$. Define its marginal to be $\mathbb{P}_Q$, which has density $p_Q(x) = \int_x p_Q(z|x)p_\star(x)dx$, then:

$$W_c(\mathbb{P}_\star, \mathbb{P}_\psi) = \inf_{Q(z|x):\mathbb{P}_Q=\mathbb{P}_z} \mathbb{E}_{\mathbb{P}_\star}\mathbb{E}_{Q(z|x)}[c(x, G_\phi(z))] \tag{3}$$

- Theoretical justification for adversarial autoencoders for Wasserstain Autoencoder.
- Interpretation: "learning an autoencoder can be interpreted as learning a generative model with latent variables (enc-dec), as long as we ensure that the marginalized encoded space is the same as the prior (critic)"
- But, multiple approximations happen here.
- The term $c(x, G_\phi(z))$ is further specified into a discrete form.
- No (deeper) justification on: (1). classifier regularization and (2). implicit prior.

Experiment Results

Recall our Original Goal

- A generative model with good reconstruction.
- A conditional model allowing us to utilize explicit semantic signal.
- A latent model allowing us to manipulate generation.
- From the experiments, are they achieved?

Experiments - Topic Transfer Samples

Science	what is an event horizon with regards to black holes ?
⇒ Music	what is your favorite sitcom with adam sandler ?
⇒ Politics	what is an event with black people ?
Science	take 1ml of hcl (concentrated) and dilute it to 50ml .
⇒ Music	take em to you and shout it to me
⇒ Politics	take bribes to islam and it will be punished .
Science	just multiply the numerator of one fraction by that of the other .
⇒ Music	just multiply the fraction of the other one that 's just like it .
⇒ Politics	just multiply the same fraction of other countries .
Music	do you know a website that you can find people who want to join bands ?
⇒ Science	do you know a website that can help me with science ?
⇒ Politics	do you think that you can find a person who is in prison ?
Music	all three are fabulous artists , with just incredible talent ! !
⇒ Science	all three are genetically bonded with water , but just as many substances , are capable of producing a special case .
⇒ Politics	all three are competing with the government , just as far as i can .
Music	but there are so many more i can 't think of !
⇒ Science	but there are so many more of the number of questions .
⇒ Politics	but there are so many more of the can i think of today .
Politics	republicans : would you vote for a cheney / satan ticket in 2008 ?
⇒ Science	guys : how would you solve this question ?
⇒ Music	guys : would you rather be a good movie ?
Politics	4 years of an idiot in office + electing the idiot again = ?
⇒ Science	4 years of an idiot in the office of science ?
⇒ Music	4) <unk> in an idiot , the idiot is the best of the two points ever !

Experiments - Latent Code Manipulation Samples

A man in a tie is sleeping and clapping on balloons . A man in a tie is clapping and walking dogs .	$\Rightarrow_{\text{walking}}$
The jewish boy is trying to stay out of his skateboard . The jewish man is trying to stay out of his horse .	$\Rightarrow_{\text{man}}$
Some child head a playing plastic with drink . Two children playing a head with plastic drink .	$\Rightarrow_{\text{Two}}$
The people shine or looks into an area . The dog arrives or looks into an area .	$\Rightarrow_{\text{dog}}$
A women are walking outside near a man . Three women are standing near a man walking .	$\Rightarrow_{\text{standing}}$
A side child listening to a piece with steps playing on a table . Several child playing a guitar on side with a table .	$\Rightarrow_{\text{Several}}$

Experiments - Numerical Metrics

Data	Reverse PPL	Forward PPL
Real data	27.4	-
LM samples	90.6	18.8
AE samples	97.3	87.8
ARAE samples	82.2	44.3

Table 1: Reverse PPL: Perplexity of language models trained on the synthetic samples from a ARAE/AE/LM, and evaluated on real data. Forward PPL: Perplexity of a language model trained on real data and evaluated on synthetic samples.

	Automatic Evaluation			
Model	Transfer	BLEU	Forward	Reverse
Cross-Aligned AE	77.1%	17.75	65.9	124.2
AE	59.3%	37.28	31.9	68.9
ARAE, $\lambda_a^{(1)}$	73.4%	31.15	29.7	70.1
ARAE, $\lambda_b^{(1)}$	81.8%	20.18	27.7	77.0

	Human Evaluation		
Model	Transfer	Similarity	Naturalness
Cross-Aligned AE	57%	3.8	2.7
ARAE, $\lambda_b^{(1)}$	74%	3.7	3.8

Table 3: Sentiment transfer. (Top) Automatic metrics (Transfer/BLEU/Forward PPL/Reverse PPL), (Bottom) Human evaluation metrics (Transfer/Similarity/Naturalness). Cross-Aligned AE is from Shen et al. (2017)

- Forward PPL↓: roughly = exp(NLL),
- * Reverse PPL↓: generate a corpus, train a LM on this generated corpus, calculate PPL on a test set. Measure mode collapse.
- Transfer↑: percentage of transfer cases which the classifier thinks is a success
- BLEU↑: n-gram matching between generated sentences and reference setences.

Conclusion and Discussion

- Goal-oriented modeling procedure.
- Assemble a disentangling model with an adversarially ragularized autoencoder.
- Theoretically not very well-discussed.
- Influence remains at word-level, short sentences.
- "models are quite sensitive to their training setup, and that different models do well on different metrics"
- Many open issues remained.