Adversarially Regularized Autoencoders

Jake (Junbo) Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun

Yao Fu, April 24th 2019

This is the paper I present in the Columbia DGM course.
The original paper skips many motivation of the task and the posterior collapse problem. In this presentation, I try to recover the missing points of the original paper

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?

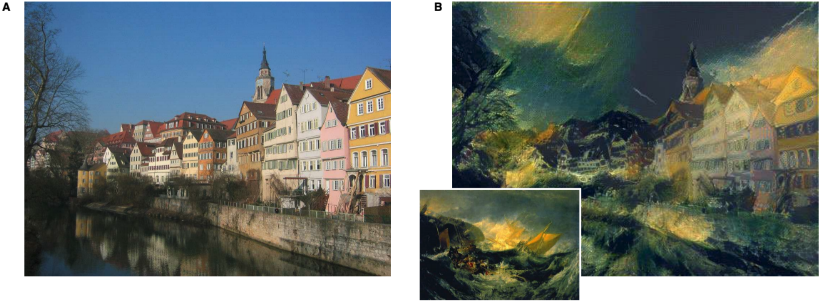
On multi-mode distribution: distribution of text usually exhibits multiple peaks. MLE tend to converge to fewer modes than the true distribution
On multi-task learning: we generally qualify natural language from different aspects (fluency, informativeness, style .etc). Different quality of language lead to different learning objectives

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.

On conditional signal: usually when conditioned on a label, only few words correlated with that label is influenced. But the syntactical structural usually remain unchanged.

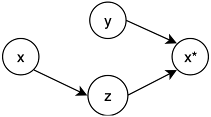
Recall: Image style transfer



- 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* → I love the *food in this restaurant*
- "style": sentiment, topic, gender, authorship, .etc
- Intuition - disentangling the style from the content:



Usually we think of modeling this with a Gaussian-VAE. But it will have posterior collapse

  - $x \to y, z$
  - 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:

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

This is to say, after the posterior collapse, the encoder will be useless, the decoder will be a unconditional LM

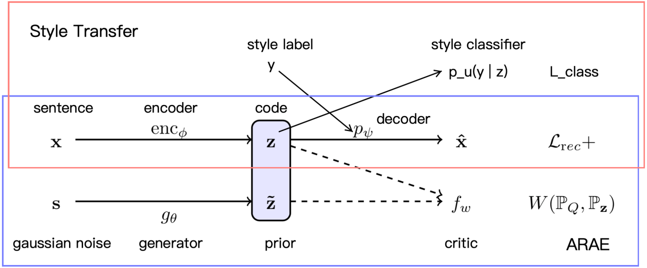
- When $KL(q_\phi(z|x)||p(z)) = 0$:
  - $q_\phi(z|x) = N(0, I)$
  - Sample from $q_\phi(z|x)$ = random gaussian noise. → Encoder meaningless.
  - VAE model → MLE model with the decoder. → An RNN language model.
  - $p_\psi(x|z) = \prod_{i=1}^{m} p(x_i|x_{1,\ldots,i-1}, z) \to \prod_{i=1}^{m} p(x_i|x_{1,\ldots,i-1})$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

During our seminar discussion, we still cannot find a solid way to interpret the choice of latent prior. The only thing we can say is, this approach will prevent the posterior from catching the prior during optimization

Ensemble the Full ARAE



The disentangling model for style transfer:

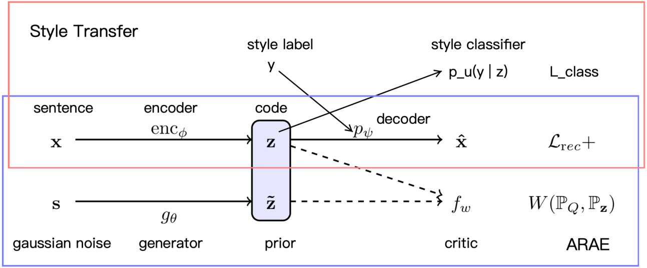
- The latent content: $z = \text{enc}_\phi(x)$. (the posterior collapse)
- The style classifier: $p_u(y|z)$

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: $p_\psi(x|z)$

Note that the upper part along is enough for style transfer. The lower part is only for learning a meaningful prior.

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 $G_\psi : \mathcal{Z} \rightarrow \mathcal{X}$ 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:

approximations happen at these three parts:

i.e. the infimum, the expectation, and the cost

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

the EMD between the prior and the posterior comes from here

the reconstruction loss

- 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.

i.e. The lower left part of the model graph

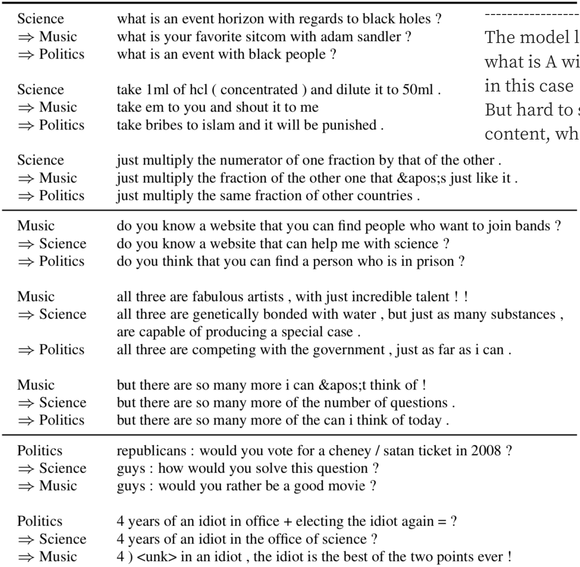
i.e. the upper part of the previous model graph

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



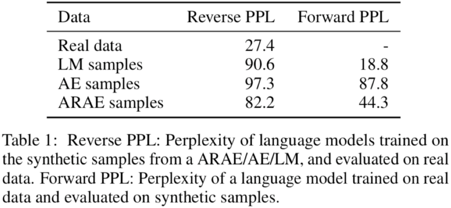
Science	what is an event horizon with regards to black holes ?
⇒ Music	what is your favorite sitcom with adam sandler ?
⇒ Politics	what is an event with black people ?
Science	take 1ml of hcl (concentrated) and dilute it to 50ml .
⇒ Music	take em to you and shout it to me
⇒ Politics	take bribes to islam 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 &apos;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 &apos;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 !

The model learns the pattern:
what is A with B?
in this case
But hard to say what is
content, what is style

Experiments - Latent Code Manipulation Samples

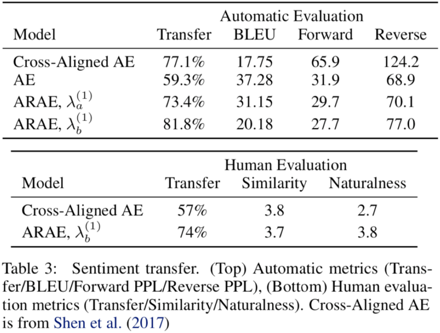
A man in a tie is sleeping and clapping on balloons . A man in a tie is clapping and walking dogs .	⇒walking
The jewish boy is trying to stay out of his skateboard . The jewish man is trying to stay out of his horse .	⇒man
Some child head a playing plastic with drink . Two children playing a head with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

What's interesting is that the reverse PPL of the ARAE is better than the LM, indicating that the ARAE covers more modes of the true distribution

- 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.