Adversarially Regularized Autoencoders

Jake Zhao* [1,3] Yoon Kim* [2] Kelly Zhang [1] Alexander Rush [2]
Yann LeCun [1,3]

[1]NYU CILVR Lab

[2]Harvard NLP

[3]Facebook AI Research

Training Deep Latent Variable Models

Two dominant approaches

- **Variational inference**: bound $\log p_\theta(\mathbf{x})$ with the evidence lower bound (ELBO) and find a variational distribution that approximates the posterior $\implies$ Variational Autoencoders (VAE)
- **Implicit density methods**: Avoid dealing with the likelihood directly and learn a discriminator that distinguishes between real/fake samples $\implies$ Generative Adversarial Networks (GAN)

Training Deep Latent Variable Models

Two dominant approaches

- **Variational inference**: bound $\log p_\theta(\mathbf{x})$ with the evidence lower bound (ELBO) and find a variational distribution that approximates the posterior $\Longrightarrow$ Variational Autoencoders (VAE)
- **Implicit density methods**: Avoid dealing with the likelihood directly and learn a discriminator that distinguishes between real/fake samples $\Longrightarrow$ Generative Adversarial Networks (GAN)

Training Deep Latent Variable Models

Two dominant approaches

- **Variational inference**: bound $\log p_\theta(\mathbf{x})$ with the evidence lower bound (ELBO) and find a variational distribution that approximates the posterior $\implies$ Variational Autoencoders (VAE)
- **Implicit density methods**: Avoid dealing with the likelihood directly and learn a discriminator that distinguishes between real/fake samples $\implies$ Generative Adversarial Networks (GAN)

Training GANs for natural language is hard because the loss is not differentiable with respect to the generator

GAN: Problem

Possible solutions

- Use policy gradient techniques from reinforcement learning (Yu et al. 2017, Lin et al. 2017)
  - unbiased but high variance gradients
  - need to pre-train with MLE
- Consider a "soft" approximation to the discrete space (Rajeswar et al. 2017, Shen et al. 2017):
  - e.g. with the Gumbel-Softmax distribution (Maddison et al. 2017, Jang et al. 2017)
  - hard to scale to longer sentences/larger vocabulary sizes

GAN: Problem

Possible solutions

- Use policy gradient techniques from reinforcement learning (Yu et al. 2017, Lin et al. 2017)
  - unbiased but high variance gradients
  - need to pre-train with MLE
- Consider a "soft" approximation to the discrete space (Rajeswar et al. 2017, Shen et al. 2017):
  - e.g. with the Gumbel-Softmax distribution (Maddison et al. 2017, Jang et al. 2017)
  - hard to scale to longer sentences/larger vocabulary sizes

Our Work

Adversarially Regularized Autoencoders (ARAE)

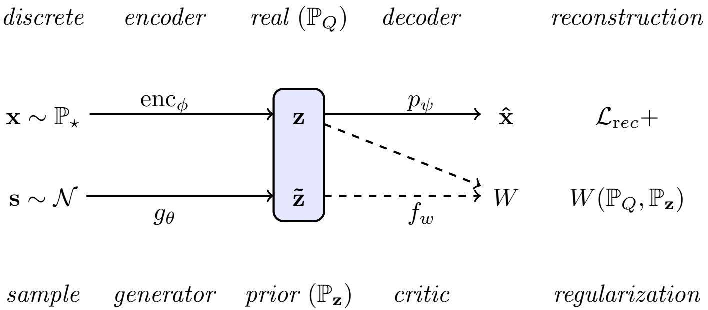
- Learns an autoencoder that encodes discrete input into a continuous space and decode from it.
- Adversarial training in the continuous space at the same time

Our Work

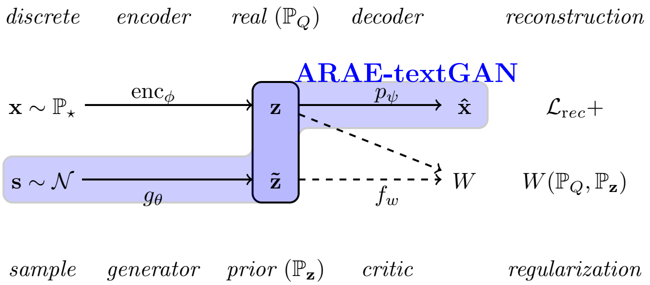
Adversarially Regularized Autoencoders (ARAE)

- Learns an autoencoder that encodes discrete input into a continuous space and decode from it.
- Adversarial training in the continuous space at the same time

Adversarially Regularized Autoencoders

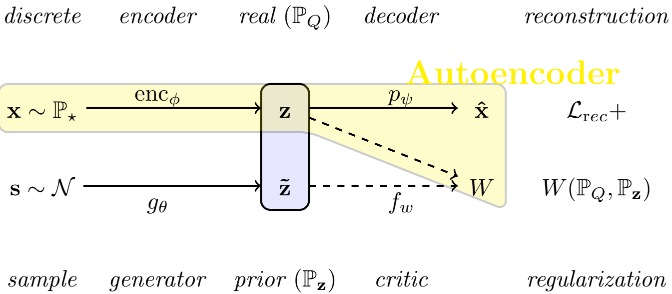


Adversarially Regularized Autoencoders



- In Corollary 1, we proved the equivalency of training ARAE and a latent variable model using the prior distribution, **in the discrete case**.
- Text generation
- Latent space manipulation: interpolation / vector arithmetic

Adversarially Regularized Autoencoders



- Semi-supervised learning
- Unaligned style transfer

Adversarially Regularized Autoencoders: experiments

New metric: **Reverse perplexity**, w/ normally used **Forward perplexity**

- Generate **synthetic** training data from generative model
- Train a RNN language model on generated data
- Evaluate perplexity, $PPL = \exp(-\frac{1}{N}\sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}))$ on **real** data
- Captures mode-collapse (vs regular PPL)
- Baselines
    - Autoregressive model: RNN language model
    - Autoencoder without adversarial regularization
    - Aversarial Autoencoders with no standalone generator (mode-collapse, Reverse PPL 980)
    - Unable to train VAEs on this dataset

Adversarially Regularized Autoencoders: experiments

New metric: **Reverse perplexity**, w/ normally used **Forward perplexity**

- Generate **synthetic** training data from generative model
- Train a RNN language model on generated data
- Evaluate perplexity, $PPL = \exp(-\frac{1}{N}\sum_{i=1}^{N}\log p(\mathbf{x}^{(i)}))$ on **real** data
- Captures mode-collapse (vs regular PPL)
- Baselines
  - Autoregressive model: RNN language model
  - Autoencoder without adversarial regularization
  - Aversarial Autoencoders with no standalone generator (mode-collapse, Reverse PPL 980)
  - Unable to train VAEs on this dataset

Adversarially Regularized Autoencoders: experiments

New metric: **Reverse perplexity**, w/ normally used **Forward perplexity**

- Generate **synthetic** training data from generative model
- Train a RNN language model on generated data
- Evaluate perplexity, $PPL = \exp(-\frac{1}{N}\sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}))$ on **real** data
- Captures mode-collapse (vs regular PPL)
- Baselines
  - Autoregressive model: RNN language model
  - Autoencoder without adversarial regularization
  - Aversarial Autoencoders with no standalone generator (mode-collapse, Reverse PPL 980)
  - Unable to train VAEs on this dataset

Adversarially Regularized Autoencoders: experiments

New metric: **Reverse perplexity**, w/ normally used **Forward perplexity**

- Generate **synthetic** training data from generative model
- Train a RNN language model on generated data
- Evaluate perplexity, $PPL = \exp(-\frac{1}{N}\sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}))$ on **real** data
- Captures mode-collapse (vs regular PPL)
- Baselines
  - Autoregressive model: RNN language model
  - Autoencoder without adversarial regularization
  - Aversarial Autoencoders with no standalone generator (mode-collapse, Reverse PPL 980)
  - Unable to train VAEs on this dataset

Adversarially Regularized Autoencoders: experiments

New metric: **Reverse perplexity**, w/ normally used **Forward perplexity**

- Generate **synthetic** training data from generative model
- Train a RNN language model on generated data
- Evaluate perplexity, $PPL = \exp(-\frac{1}{N}\sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}))$ on **real** data
- Captures mode-collapse (vs regular PPL)
- Baselines
  - Autoregressive model: RNN language model
  - Autoencoder without adversarial regularization
  - Aversarial Autoencoders with no standalone generator (mode-collapse, Reverse PPL 980)
  - Unable to train VAEs on this dataset

Adversarially Regularized Autoencoders

Data for Training LM	Reverse PPL
Real data	27.4
Language Model samples	90.6
Autoencoder samples	97.3
ARAE samples	82.2

(Lower perplexity means higher likelihood)

ARAE: Unaligned Style Transfer

Transfer Sentiment

- Train a classifier on top of the code space:
  $classifer(\mathbf{c}) =$ probability $\mathbf{c}$ is a positive sentiment sentence
- The encoder is trained to **fool** the classifier
- To transfer sentiment:
  - Encode sentence to get code $\mathbf{c}$
  - Switch the sentiment label, concatenate with $\mathbf{c}$
  - Generate using the concatenated vector

ARAE: Unaligned Style Transfer

Transfer Sentiment

- Train a classifier on top of the code space:
  $classifer(\mathbf{c}) =$ probability $\mathbf{c}$ is a positive sentiment sentence
- The encoder is trained to **fool** the classifier
- To transfer sentiment:
  - Encode sentence to get code $\mathbf{c}$
  - Switch the sentiment label, concatenate with $\mathbf{c}$
  - Generate using the concatenated vector

ARAE: Unaligned Style Transfer

Transfer Sentiment

- Train a classifier on top of the code space:
  $classifer(\mathbf{c}) =$ probability $\mathbf{c}$ is a positive sentiment sentence
- The encoder is trained to **fool** the classifier
- To transfer sentiment:
  - Encode sentence to get code $\mathbf{c}$
  - Switch the sentiment label, concatenate with $\mathbf{c}$
  - Generate using the concatenated vector

ARAE: Unaligned Style Transfer

Cross-AE: State-of-the-art model from Shen et al. 2017

	Positive ⇒ Negative
Original	great indoor mall .
ARAE	no smoking mall .
Cross-AE	terrible outdoor urine .
Original	it has a great atmosphere , with wonderful service .
ARAE	it has no taste , with a complete jerk .
Cross-AE	it has a great horrible food and run out service .
Original	we came on the recommendation of a bell boy and the food was amazing .
ARAE	we came on the recommendation and the food was a joke .
Cross-AE	we went on the car of the time and the chicken was awful .

ARAE: Unaligned Style Transfer

Cross-AE: State-of-the-art model from Shen et al. 2017

	Negative ⇒ Positive
Original	hell no !
ARAE	hell great !
Cross-AE	incredible pork !
Original	small , smokey , dark and rude management .
ARAE	small , intimate , and cozy friendly staff .
Cross-AE	great , , , chips and wine .
Original	the people who ordered off the menu did n't seem to do much better .
ARAE	the people who work there are super friendly and the menu is good .
Cross-AE	the place , one of the office is always worth you do a business .

ARAE: Unaligned Style Transfer

Automatic Evaluation

Model	Transfer	BLEU	PPL	Reverse PPL
Cross-Aligned AE	77.1%	17.75	65.9	124.2
ARAE	81.8%	20.18	27.7	77.0

Human Evaluation

Model	Transfer	Similarity	Naturalness
Cross-Aligned AE	57%	3.8	2.7
ARAE	74%	3.7	3.8

(Similarity/Naturalness scores are between [1,5], 5 being best)

ARAE: Unaligned Style Transfer

Topic Transfer from Yahoo! Answers Dataset

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 ?
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 ?
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 ?

ARAE: Conclusion

- Introduced a simple method for training a GAN for text by performing generation/discrimination in a continuous code space
- A (somewhat) successful text-GAN instatiation
- Can do unaligned style transfer through training an additional classifier (much exciting work in this area: Shen et al. 2017, Prabhumoye et al. 2018)

ARAE: Open source

All our code is available at: https://github.com/jakezhaojb/ARAE.

Poster: #58