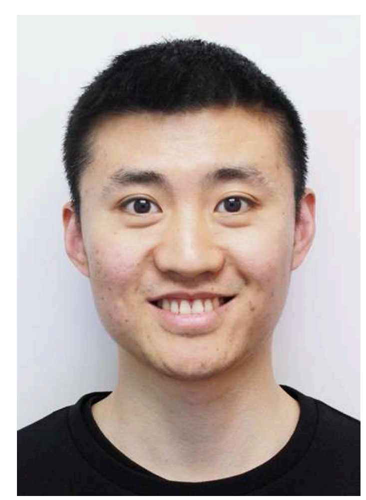
# Stochastic Wasserstein Autoencoder for Probabilistic Sentence Generation

Hareesh Bahuleyan[w], **Lili Mou[w]**, Hao Zhou[b], Olga Vechtomova[w]

University of Waterloo, ByteDance AI Lab

NAACL-HLT 2019



# Roadmap

- VAE
- WAE
- Stochastic WAE

UNIVERSITY OF WATERLOO

ByteDance 字节跳动

# Variational Autoencoder

$$x \to z \to x$$

$$p(z)$$

- VAE: Treating *z* as a random variable
  - Imposing prior $p(z) = \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - Variational posterior

$$q(z \mid x) = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{NN}}, \operatorname{diag} \sigma^2_{\mathrm{NN}})$$

  - Optimizing the variational lower bound

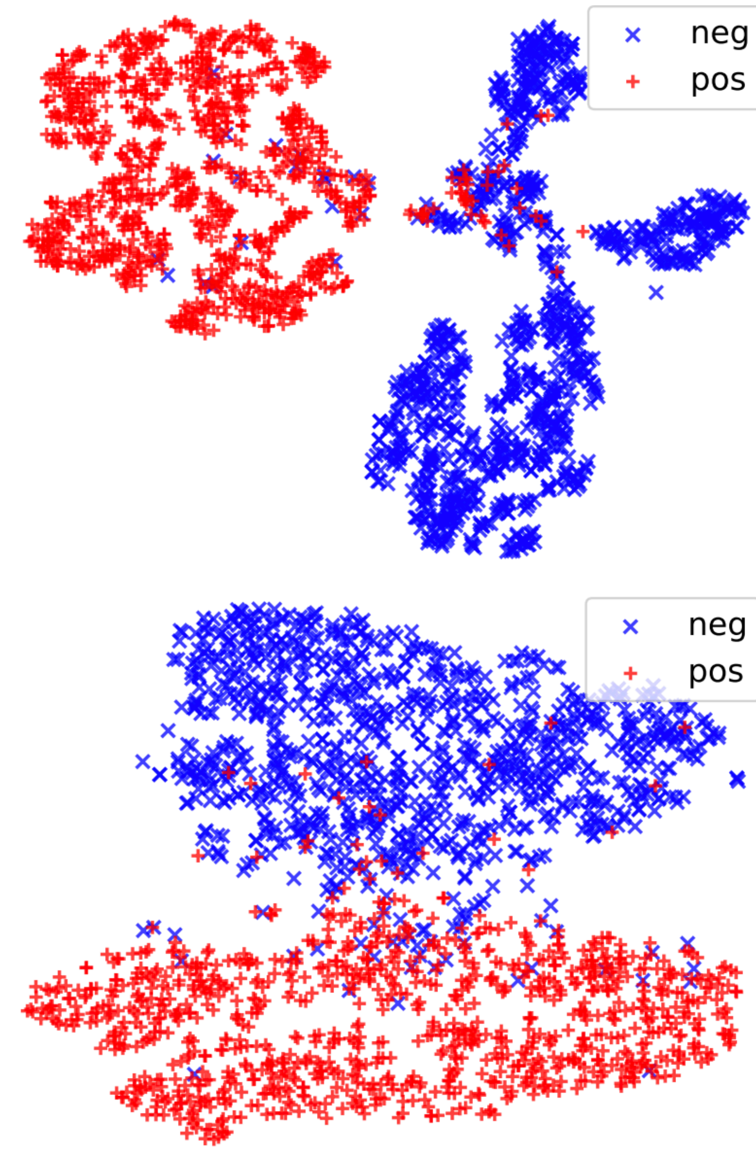
$$J = \mathop{\mathbb{E}}_{z \sim q(z|x)} [-\log p(x \mid z)] + \mathrm{KL}(q(z \mid x) \| p(z))$$

UNIVERSITY OF WATERLOO

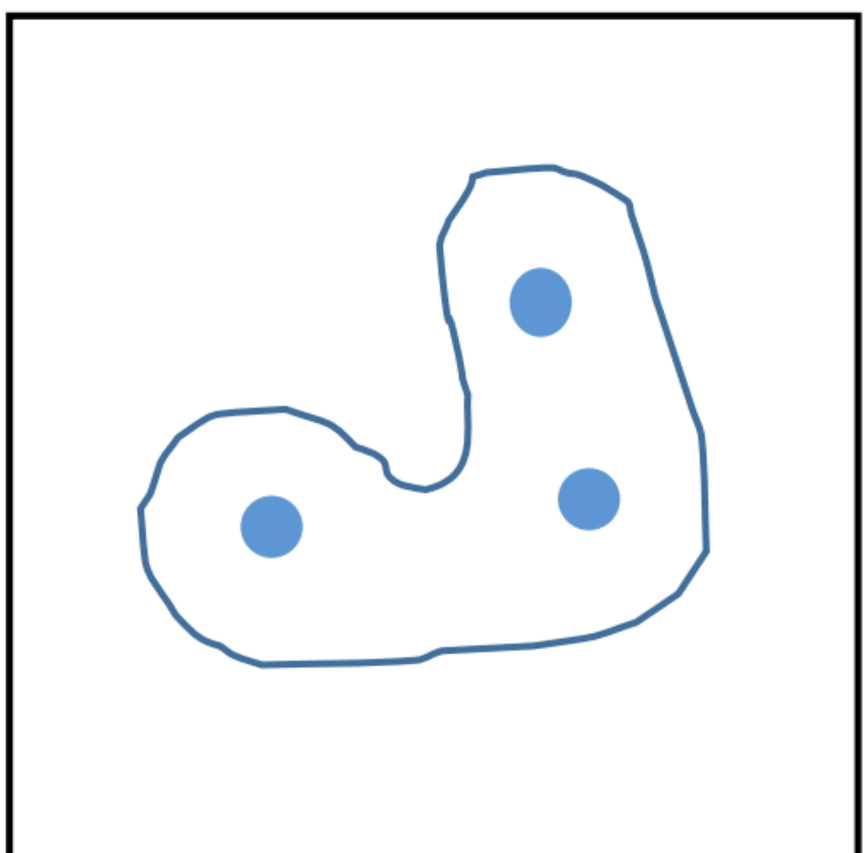
ByteDance 字节跳动

[John, Mou, Bahuleyan, Vechtomova, ACL2019]

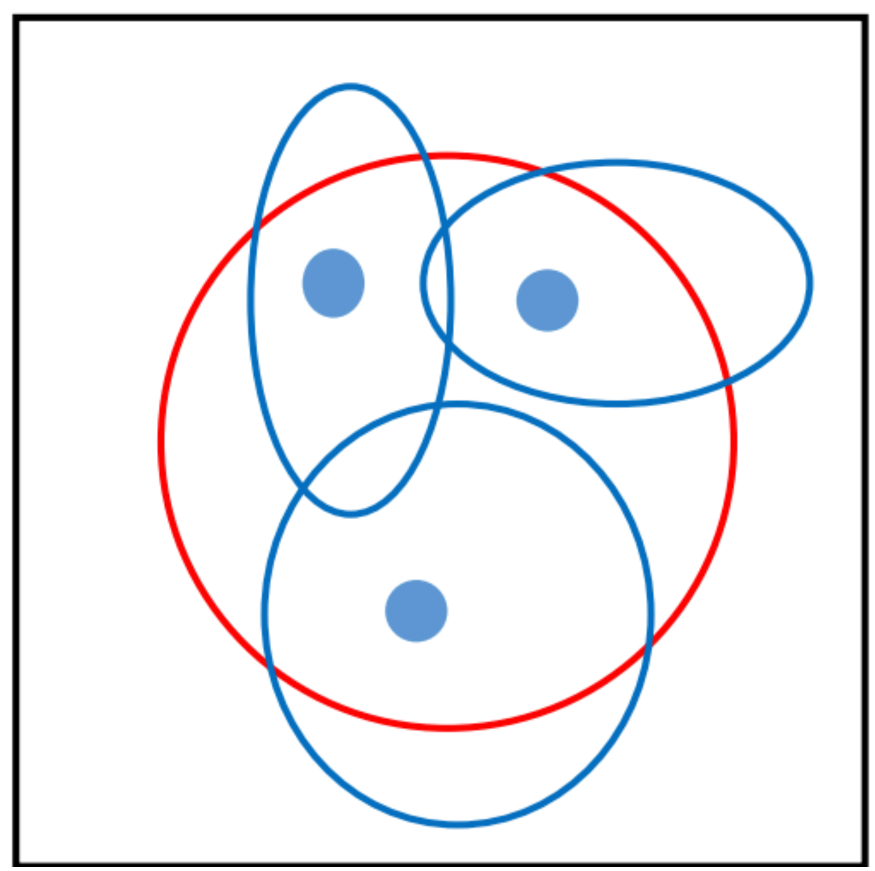
# Latent Space




AE



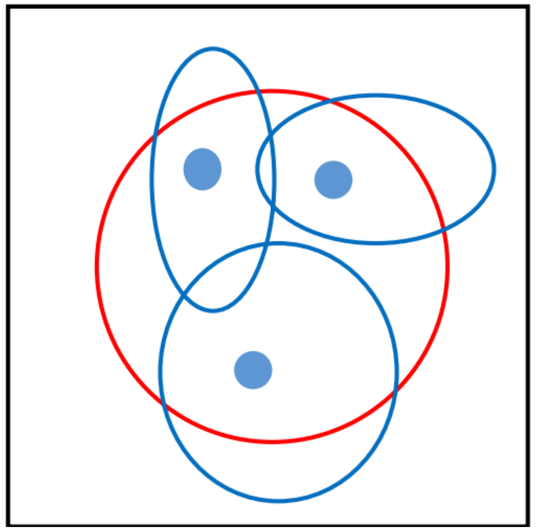
VAE



UNIVERSITY OF WATERLOO

ByteDance 字节跳动

# Disadvantages of VAE



- Two objective terms are conflicting
  - Perfect reconstruction => High KL
  - Perfect KL => no information captured in z

$$J = \mathop{\mathbb{E}}_{z \sim q(z|x)} [-\log p(x \,|\, z)] + \mathrm{KL}(q(z \,|\, x) \| p(z))$$

- Consequence: KL collapse
  - KL —> 0
  - Decoder -> Language model

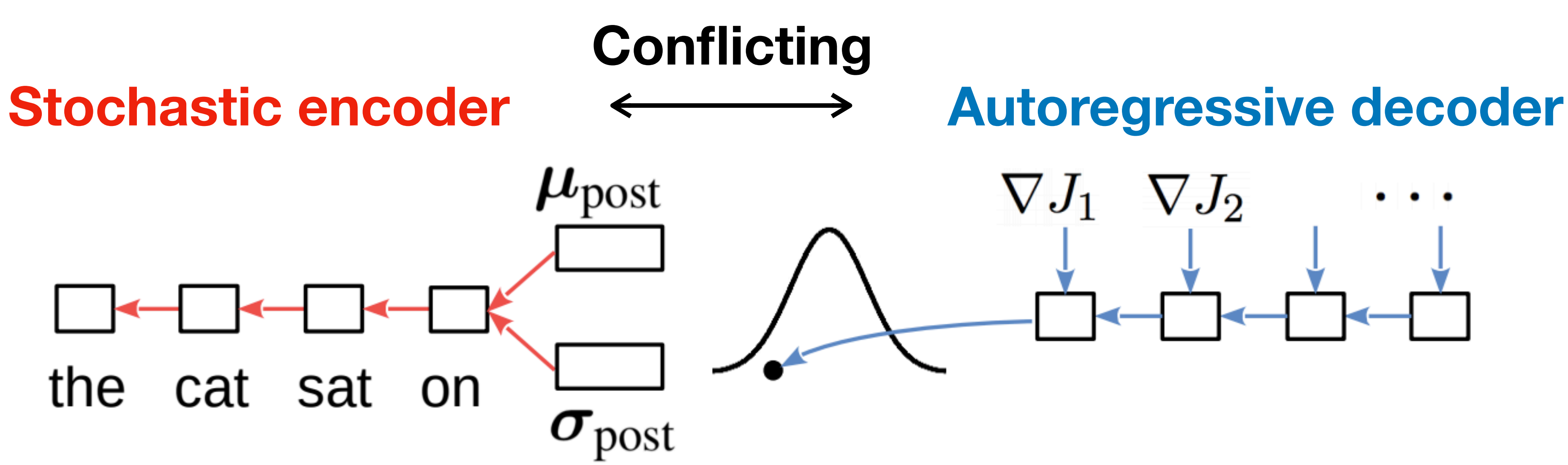
UNIVERSITY OF WATERLOO

ByteDance 字节跳动

# Engineering Fixes

[Bowman+, CoNLL, 2016]

- KL annealing
  - Reducing encoder's stochasticity
  - No KL penalty => $\sigma \to \mathbf{0}$ [Thm 1]
- Word dropout (in decoder)
  - Reducing decoder's auto-regressiveness



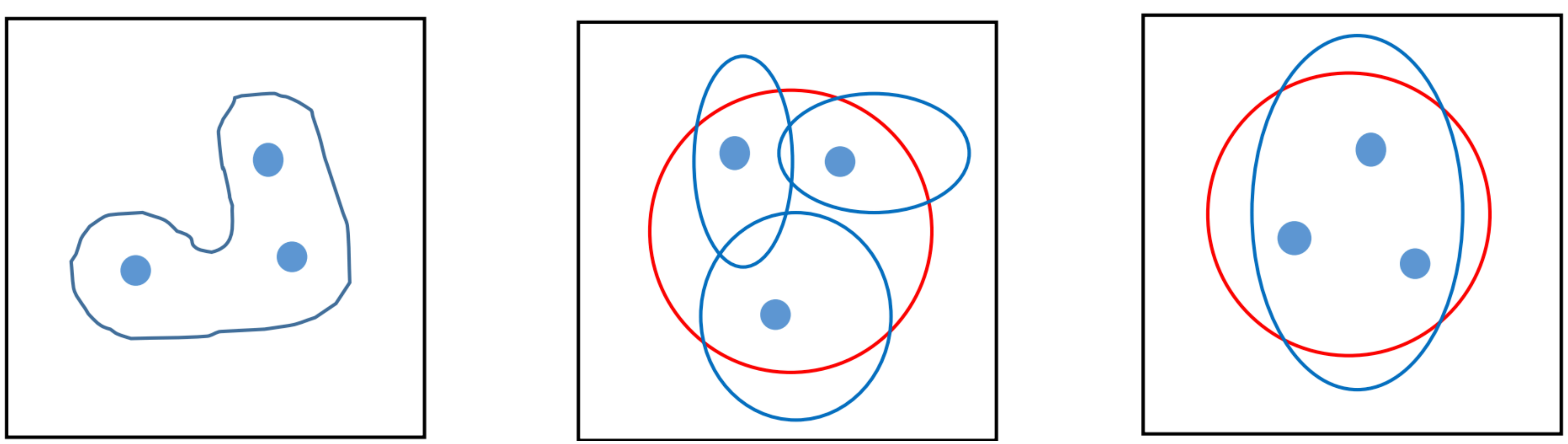
# Wasserstein Autoencoder

- VAE penalty

For any $x \in \mathscr{D}$, $\quad q(z\,|\,x) \xrightarrow{\text{close}} p(z)$

- WAE penalty

$$q(z) := \int_{x \in \mathscr{D}} q(z\,|\,x) p_{\mathscr{D}}(x)\ \mathrm{d}x \quad \xrightarrow{\text{set}} \quad p(z)$$



# Wasserstein Distance

- Constraint *q(z)* = *p(z)* relaxed by some "distance" *W( p(z), q(z) )*

$$q(z) := \int_{x\in\mathscr{D}} q(z\,|\,x)p_{\mathscr{D}}(x)\ \mathrm{d}x \quad \xrightarrow{\text{set}} \quad p(z)$$

- GAN-loss
- MMD-loss

$$\mathrm{MMD} = \left\| \int k(\boldsymbol{z},\cdot)\,\mathrm{d}P(\boldsymbol{z}) - \int k(\boldsymbol{z},\cdot)\,\mathrm{d}Q(\boldsymbol{z}) \right\|_{\mathcal{H}_k}$$

$$\widehat{\mathrm{MMD}} = \frac{1}{N(N-1)}\sum_{n\neq m} k(\boldsymbol{z}^{(n)},\boldsymbol{z}^{(m)}) + \frac{1}{N(N-1)}\sum_{n\neq m} k(\widetilde{\boldsymbol{z}}^{(n)},\widetilde{\boldsymbol{z}}^{(m)}) - \frac{1}{N^2}\sum_{n,m} k(\boldsymbol{z}^{(n)},\widetilde{\boldsymbol{z}}^{(m)})$$

Both based on samples of *p(z)* and *q(z)*

- Training objective

$$J = \mathop{\mathbb{E}}_{z\sim q(z|x)}[-\log p(x\,|\,z)] + W(q(z)\|p(z))$$

The two terms are not conflicting

# Stochastic Encoder Collapses

- Stochastic encoder is desired
  - Learning uncertainty of data
  - Posterior sampling
    - Unsupervised paraphrase generation [Bao+ACL19]

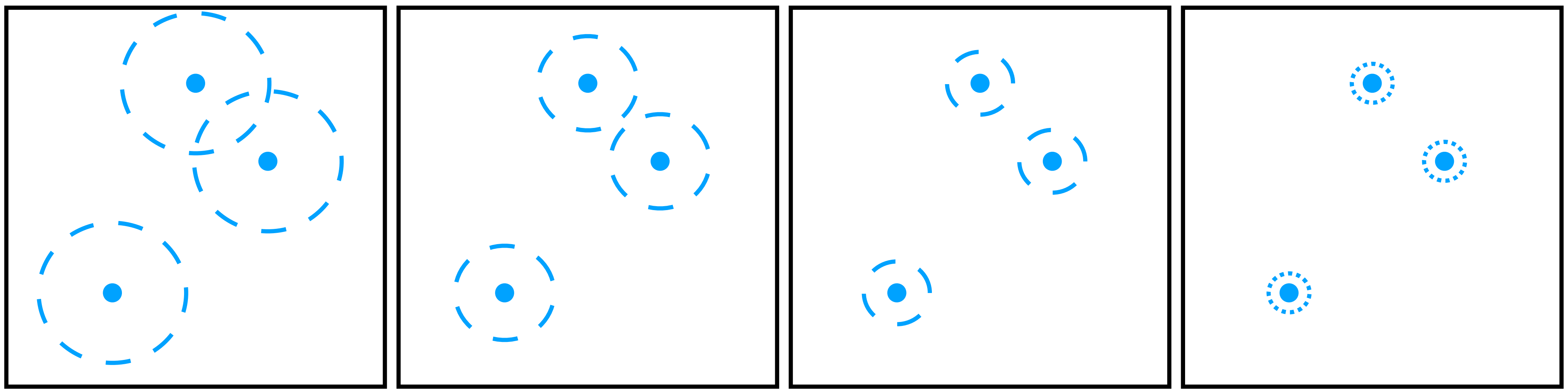
- Stochasticity collapse $q(z\,|\,x) \rightarrow \delta_\mu$

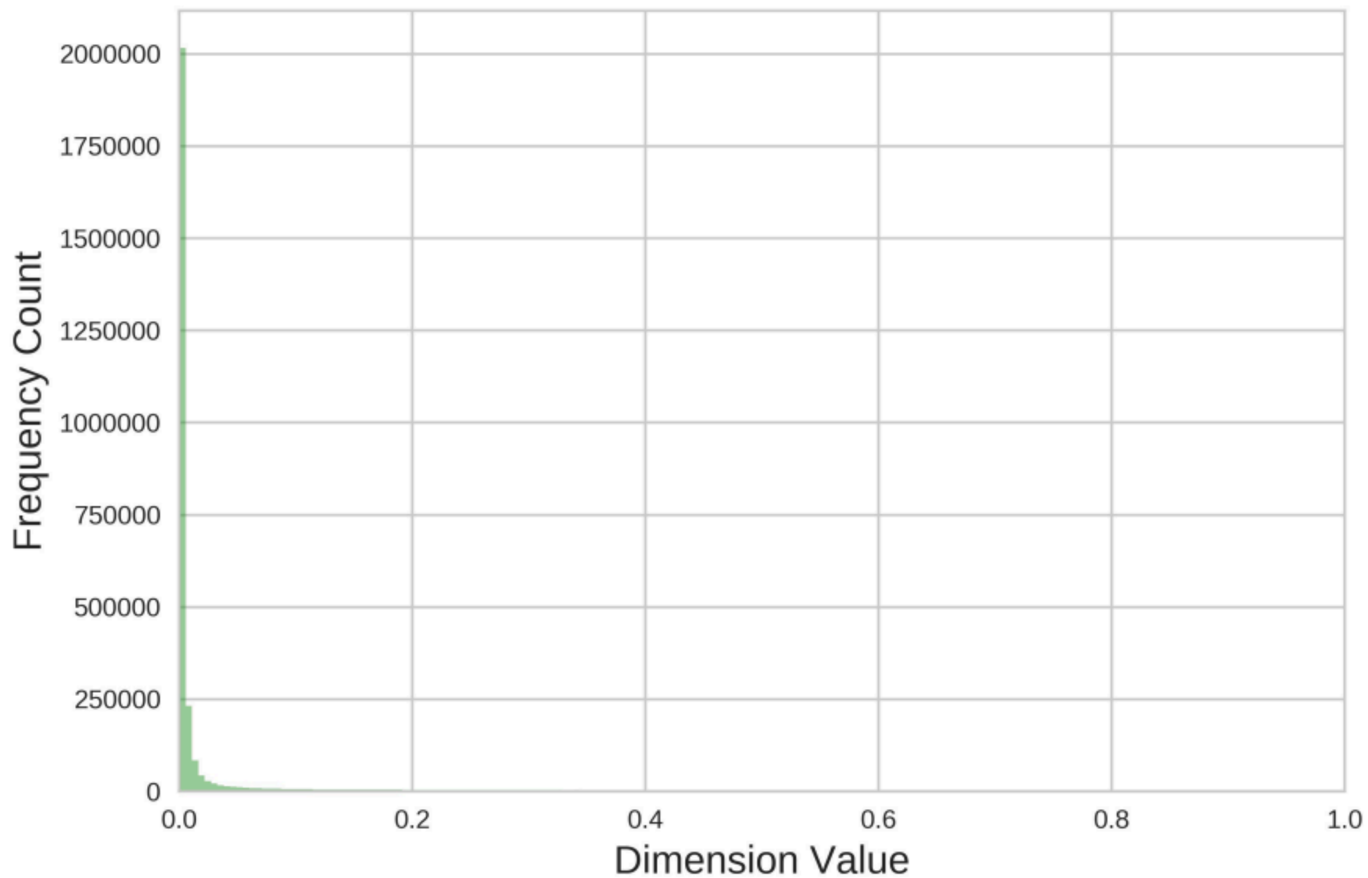
$$J = \mathop{\mathbb{E}}_{z \sim q(z|x)} [-\log p(x\,|\,z)] + W(q(z) \| p(z))$$

UNIVERSITY OF WATERLOO

ByteDance 字节跳动

# Illustration & Empirical evidence





# Why Stochasticity collapses?

- Direct optimization from a family of encoders
  - Stochasticity is bad for reconstruction
- Numerical optimization

**Theorem 1.** *Suppose we have a Gaussian family $\mathcal{N}(\boldsymbol{\mu}, \operatorname{diag} \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are parameters. The covariance is diagonal, meaning that the variables are independent. If the gradient of $\boldsymbol{\sigma}$ completely comes from sample gradient and $\boldsymbol{\sigma}$ is small at the beginning of training, then the Gaussian converges to a Dirac delta function with stochastic gradient descent, i.e., $\boldsymbol{\sigma} \to \mathbf{0}$.*

# Our Fix

- Penalizing a per-sample KL term against a Gaussian centered at the predicted mean

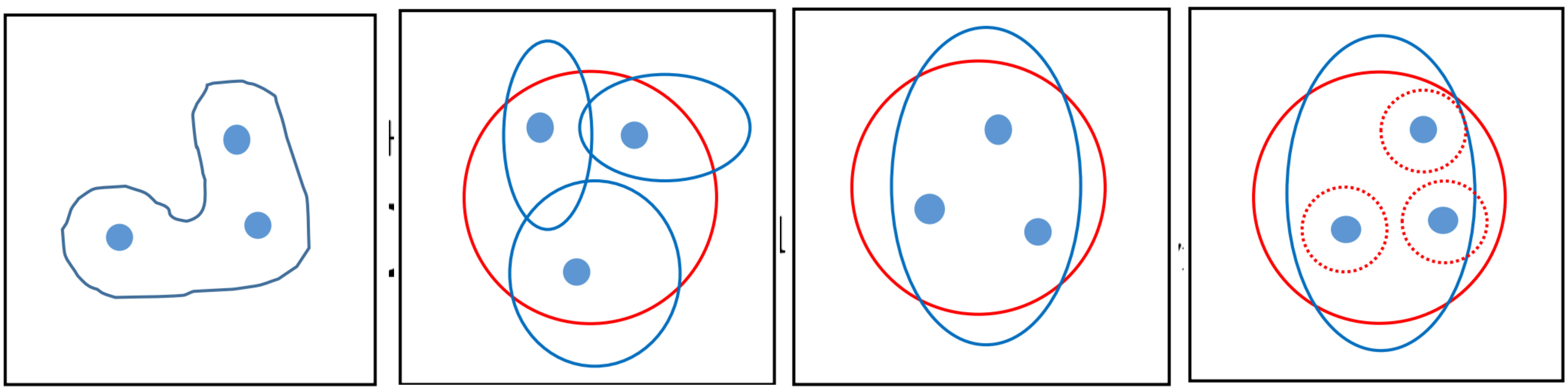
$$J = J_{\text{rec}} + \lambda_{\text{WAE}} \cdot \widehat{\text{MMD}}$$
$$+ \lambda_{\text{KL}} \sum\nolimits_{n} \text{KL}\left(\mathcal{N}(\boldsymbol{\mu}_{\text{post}}^{(n)}, \text{diag}(\boldsymbol{\sigma}_{\text{post}}^{(n)})^2) \| \mathcal{N}(\boldsymbol{\mu}_{\text{post}}^{(n)}, \mathbf{I})\right) \tag{5}$$

**Theorem 2.** *Objective (5) is a relaxed optimization of the WAE loss (4) with a constraint on* $\boldsymbol{\sigma}_{\text{post}}$.

$$\sum_n \sum_i \left[-\log \sigma_i^{(n)} + \frac{1}{2}(\sigma_i^{(n)})^2\right] < C$$

# Our Fix

- Penalizing a per-sample KL term against a Gaussian centered at



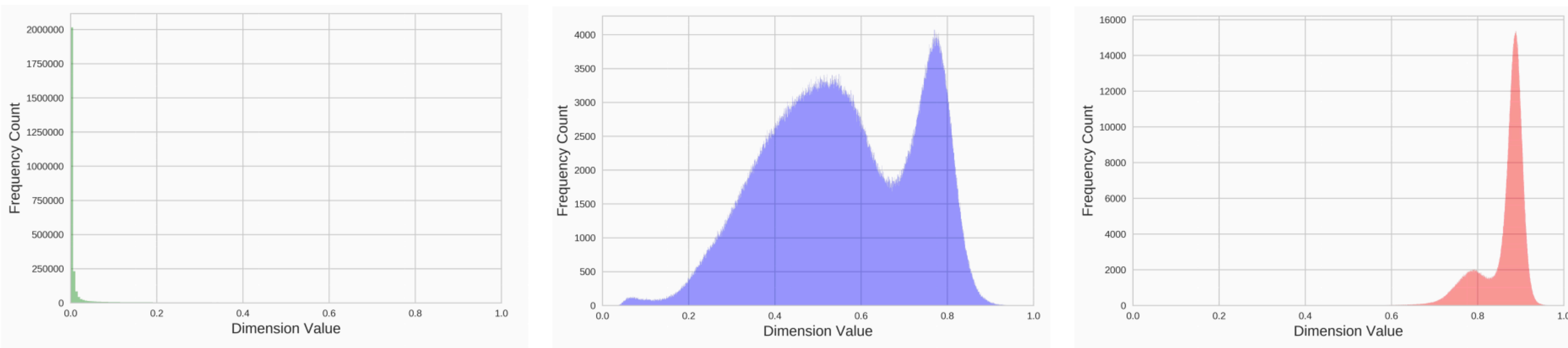
**Theorem 2.** *Objective (5) is a relaxed optimization of the WAE loss (4) with a constraint on* $\boldsymbol{\sigma}_{\text{post}}$.

$$\sum_n \sum_i \left[ -\log \sigma_i^{(n)} + \frac{1}{2} (\sigma_i^{(n)})^2 \right] < C$$

UNIVERSITY OF WATERLOO

ByteDance 字节跳动

# Distribution of $\sigma$'s




(a) $\lambda_{\mathrm{KL}} = 0$ (b) $\lambda_{\mathrm{KL}} = 0.01$ (c) $\lambda_{\mathrm{KL}} = 0.1$

- **Digression (hypothesis):**

  Two modes indicate two catchment basins

  - Language model (KL->0)
  - Reconstruction (Gaussian -> Dirac delta)

UNIVERSITY OF WATERLOO

ByteDance 字节跳动

# Experiment I: SNLI Generation

- Dataset: SNLI generation
  - Domain-specific sentence generation (similar to MNIST)
- Main results
  - WAE achieves close reconstruction performance to AE
    - Important for feature learning, conditional generation
  - WAE enjoys probabilistic properties as VAE
    - More fluent generated sentences, closer to corpus in distribution

| | BLEU↑ | PPL↓ | UniKL↓ | Entropy | AvgLen |
|---|---|---|---|---|---|
| **Corpus** | - | - | - | → 5.65 | → 9.6 |
| **DAE** | **86.35** | 146.2 | 0.178 | 6.23 | 11.0 |
| **VAE** (KL-annealed) | 43.18 | 79.4 | 0.081 | 5.04 | 8.8 |
| **WAE-D** $\lambda_{\text{WAE}}=3$ | 86.03 | 113.8 | 0.071 | 5.59 | 10.0 |
| **WAE-D** $\lambda_{\text{WAE}}=10$ | 84.29 | 104.9 | 0.073 | 5.57 | 9.9 |
| **WAE-S** $\lambda_{\text{KL}}=0.0$ | 75.66 | 115.2 | 0.069 | **5.61** | 9.9 |
| **WAE-S** $\lambda_{\text{KL}}=0.01$ | 82.01 | 84.9 | **0.058** | 5.26 | **9.4** |
| **WAE-S** $\lambda_{\text{KL}}=0.1$ | 47.63 | **62.5** | 0.150 | 4.65 | 8.7 |

UNIVERSITY OF WATERLOO

ByteDance 字节跳动

# Experiment II: Dialog Generation

- Dataset: DailyDialog [Li+, IJCNLP, 2017]
  - We deduplicate overlapping samples in the test set
- Main results
  - VAE inadmissible in this experiment

| | BLEU-2 | BLEU-4 | Entropy | Dist-1 | Dist-2 |
|---|---|---|---|---|---|
| **Test Set** | - | - | 6.15 | 0.077 | 0.414 |
| **DED** | 3.96 | 0.85 | 5.55 | 0.044 | 0.275 |
| **VED** | 3.26 | 0.59 | 5.45 | 0.053 | 0.204 |
| **WED-D** | **4.05** | **0.98** | 5.53 | 0.042 | 0.272 |
| **WED-S** | 3.72 | 0.69 | **5.59** | **0.066** | **0.309** |

UNIVERSITY OF WATERLOO

ByteDance 字节跳动

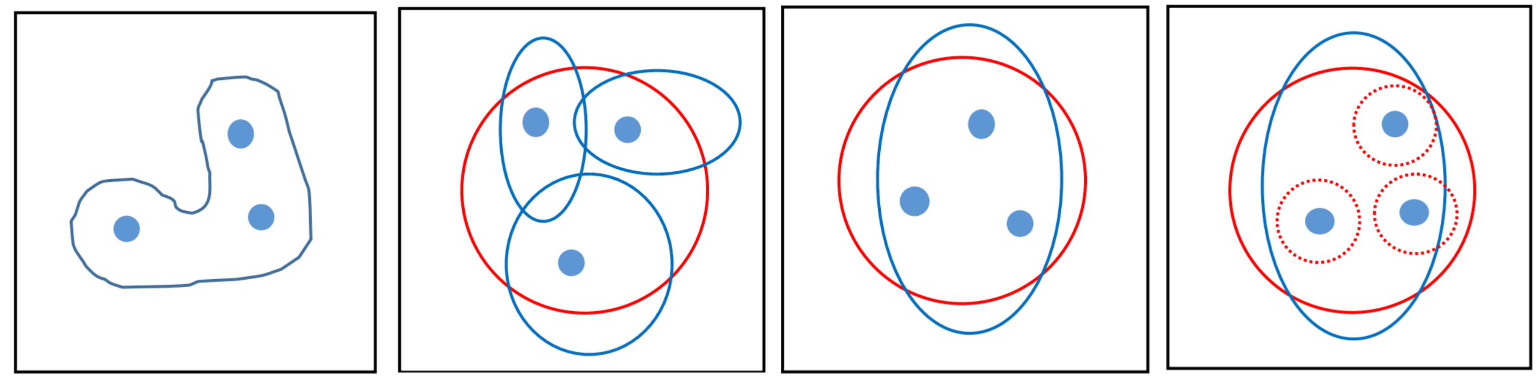
# Ease of Training

- No annealing needed
- Hyperparameters tuned on Exp. I
- Directly adopted to Exp. II

Our KL doesn't make WAE a language model

- Per-sample KL term doesn't force the posterior to be the same for different input sentences

# Conclusion

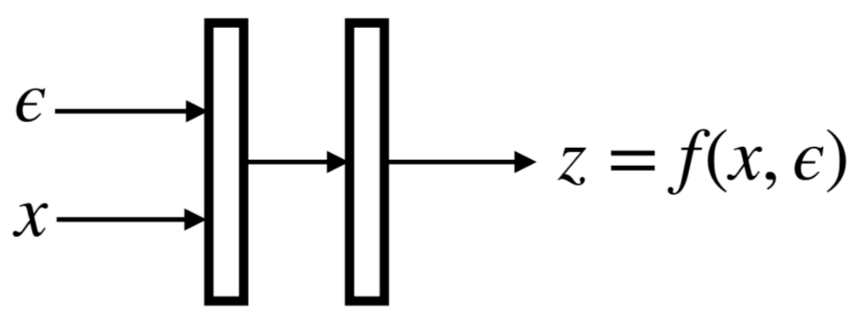


## Open questions

- A better understanding of KL collapse in VAE models
  - Two catchment basins? Flatter optimum?

**Stochastic encoder** $\xleftrightarrow{\textbf{Conflicting}}$ **Autoregressive decoder** But not exact!

- A thorough revisit of DGMs for stochasticity collapse
  - Non-Gaussian encoder? Non-reconstruction loss?




UNIVERSITY OF WATERLOO

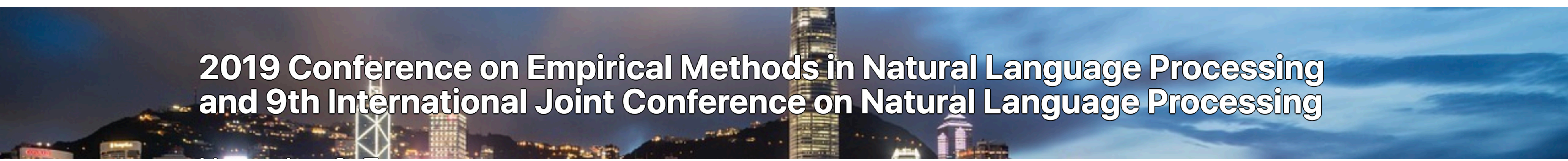
ByteDance 字节跳动

# Ads

UNIVERSITY OF ALBERTA

Lili Mou will be an assistant professor at U of Alberta

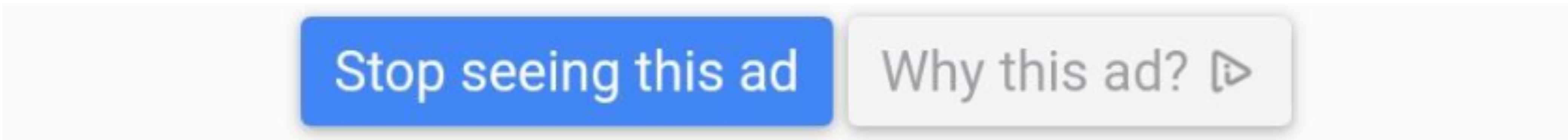
Admitting all-level students, postdocs, and visiting scholars




Lili Mou, Hao Zhou, and Lei Li

Discreteness in Neural Natural Language Processing

Tutorial @EMNLP-IJCNLP 2019




# Thank you!

# Q&A