Adversarial Feature Matching for Text Generation
Yizhe Zhang\textsuperscript{1,2}, Zhe Gan\textsuperscript{1}, Kai Fan\textsuperscript{2}, Zhi Chen\textsuperscript{1}, Ricardo Henao\textsuperscript{1}, Lawrence Carin\textsuperscript{1}

Department of Electronic and Computer Engineering\textsuperscript{1}, Duke University, Durham, NC, 27708
Department of Statistical Science\textsuperscript{2}, Duke University, Durham, NC, 27708

Motivation & Contribution
1) Estimating a distribution over sentences from a corpus, then use it to sample realistic-looking text.
2) Ameliorating mode-collapsing issue associated with standard GAN training.
3) Discretization approximations for text modeling.

Introductions
Generative adversarial network (GAN) aims to obtain the equilibrium of the following optimization objective:

\[ \mathcal{L}_{GAN} = \mathbb{E}_{x \sim p_r} \log D(x) + \mathbb{E}_{z \sim p_z} \log[1 - D(G(z))] \]

Minimizing the Jenson-Shannon Divergence (JSD) between the real data distribution and the synthetic data distribution.

TextGAN objective
We adopt a feature matching approach instead of vanilla GAN objective. Specifically, we consider the objective

\[ \mathcal{L}_D = \mathcal{L}_{GAN} - \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{MMD} \]

\[ \mathcal{L}_G = \mathcal{L}_{MMD} \]

\[ \mathcal{L}_{recon} = \mathbb{E}_{z \sim p_z} \log[1 - D(G(z))] \]

Discretization approximation
1) Instead of using original GAN loss, we consider a moment matching loss over CNN feature layer using maximum mean discrepancy (MMD).
2) The MMD measures the mean squared difference of two sets of samples over RKHS:

\[ \mathcal{L}_{MMD} = \| \mathbb{E}_{x \sim p_r} \phi(x) - \mathbb{E}_{y \sim p_g} \phi(y) \|^2_2 \]

\[ = \mathbb{E}_{x \sim p_r} \mathbb{E}_{y \sim p_g} [k(x, y)] + \mathbb{E}_{x \sim p_r} \mathbb{E}_{y \sim p_g} [k(y, x)] - 2 \mathbb{E}_{x \sim p_r} \mathbb{E}_{y \sim p_g} [k(x, y)] \]

With a Gaussian kernel \( k(x, y) = \exp(-\frac{|x-y|^2}{2\sigma^2}) \), the minimizing the MMD objective is can be visualized as matching all order of moments of two empirical distributions.

Results: empirical evaluation
Table: Sentences generated by textGAN.

![ figure: LSTM generator (left) and CNN discriminator (right) ]

MMD objective
- Instead of using original GAN loss, we consider a moment matching loss over \textit{CNN feature layer} using maximum mean discrepancy (MMD).
- The MMD measures the mean squared difference of two sets of samples over RKHS:

\[ \mathcal{L}_{MMD} = \| \mathbb{E}_{x \sim p_r} \phi(x) - \mathbb{E}_{y \sim p_g} \phi(y) \|^2_2 \]

\[ = \mathbb{E}_{x \sim p_r} \mathbb{E}_{y \sim p_g} [k(x, y)] + \mathbb{E}_{x \sim p_r} \mathbb{E}_{y \sim p_g} [k(y, x)] - 2 \mathbb{E}_{x \sim p_r} \mathbb{E}_{y \sim p_g} [k(x, y)] \]

With a Gaussian kernel \( k(x, y) = \exp(-\frac{|x-y|^2}{2\sigma^2}) \), the minimizing the MMD objective is can be visualized as matching all order of moments of two empirical distributions.

Discretization approximation
1) Score-function-based approaches, such as the REINFORCE algorithm, has very large variance of the gradient estimation.
2) We consider a Gumbel-softmax approach to approximate argmax operation.

\[ y_{t-1} = \text{softmax}(Vh_{t-1} \odot L) \]

where \( \odot \) represents the element-wise product. Note that when \( L \to \infty \), this approximation approaches argmax operation.

Alternative objective
- Problems: a minibatch (256) of data point is not densely sampled in feature space with high dimension (900).
- Alternative models:
  - Mapping feature space to lower dimension
  - Use accumulated batches, however kernel-based method is not available anymore. Instead we use Jensen-Shannon divergence:

\[ L_G = \mathbb{E}_{(\Sigma, \mu) \sim \pi} [\mu, (\mu_0 - \mu)] (\Sigma^{-1} (\mu - \mu_0)) \]

\( \Sigma \) and \( \mu \) are covariance and mean for the discriminative feature vector.

Quantitative comparison
Table: Quantitative results using BLEU-2,3,4 and KIDE.

![ figure: Moment matching comparison. Left: expectations of latent features from real vs. synthetic data. Right: elements of \( \Sigma \), \( \mu \) for real and synthetic data, respectively. ]

Conclusion
1) We introduced a novel approach for text generation using adversarial training.
2) We discussed several techniques to specify and train such a model.
3) We demonstrated that the proposed model delivers superior performance compared to related approaches.