Adversarial Feature Matching for Text Generation

ICML @ Sydney
Thirty-fourth International Conference on Machine Learning

Yizhe Zhang[1,2], Zhe Gan[1], Kai Fan[2], Zhi Chen[1], Ricardo Henao[1], Lawrence Carin[1]

Department of Electronic and Computer Engineering[1], Duke University, Durham, NC, 27708
Department of Statistical Science[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:

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

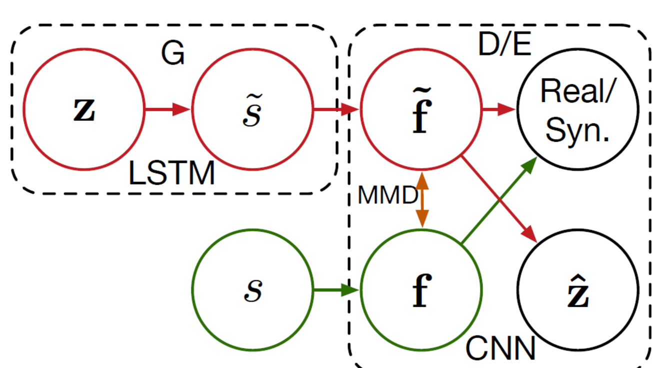
Minimizing the Jenson-Shannon Divergence (JSD) between the real data distribution and the synthetic data distribution.

TextGAN objective

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

$$\begin{aligned}
\mathcal{L}_D &= \mathcal{L}_{GAN} - \lambda_r \mathcal{L}_{recon} + \lambda_m \mathcal{L}_{MMD^2} \\
\mathcal{L}_G &= \mathcal{L}_{MMD^2} \\
\mathcal{L}_{GAN} &= \mathbb{E}_{s\sim \mathcal{S}} \log D(s) + \mathbb{E}_{z\sim p_z} \log[1 - D(G(z))] \\
\mathcal{L}_{recon} &= ||\hat{\mathbf{z}} - \mathbf{z}||^2 ,
\end{aligned}$$

- *Easier to train:* G try to match the sentence feature "fingerprint" rather than directly cheat D, which is more achievable.
- Enforce the generator to generate *different* sentence rather than a single one.
- The blue reconstruction loss in (3) enforce the most *representative* features for G is selected.
- The red mmd loss in (3) enforce the most *challenging* features for G is selected.



- Using CNN discriminator and LSTM generator.

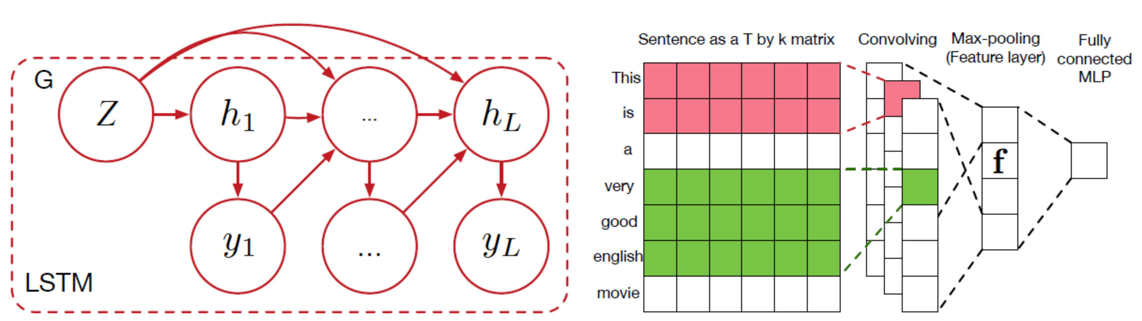


Figure: LSTM generator (left) and CNN discriminator (right)

MMD objective

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

$$\begin{aligned}
\mathcal{L}_{MMD^2} &= ||\mathbb{E}_{x\sim\mathcal{X}}\phi(x) - \mathbb{E}_{y\sim\mathcal{Y}}\phi(y)||^2_{\mathcal{H}} \\
&= \mathbb{E}_{x\sim\mathcal{X}}\mathbb{E}_{x'\sim\mathcal{X}}[k(x,x')] \\
&+ \mathbb{E}_{y\sim\mathcal{Y}}\mathbb{E}_{y'\sim\mathcal{Y}}[k(y,y')] - 2\mathbb{E}_{x\sim\mathcal{X}}\mathbb{E}_{y\sim\mathcal{Y}}[k(x,y)].
\end{aligned}$$

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

Discretization approximation

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

$$\boldsymbol{y}_{t-1} = \mathbf{W}_e \mathrm{softmax}(\mathbf{V}\boldsymbol{h}_{t-1} \odot \mathbf{L}) .$$

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:

  $$L_G = \mathrm{tr}(\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_r^{-1}\boldsymbol{\Sigma}_s) + (\boldsymbol{\mu}_s - \boldsymbol{\mu}_r)^T(\boldsymbol{\Sigma}_s^{-1} + \boldsymbol{\Sigma}_r^{-1})(\boldsymbol{\mu}_s - \boldsymbol{\mu}_r)$$

  - $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are covariance and mean for the discriminative feature vector.

Results: empirical evaluation

Table: Sentences generated by textGAN.

a	we show the joint likelihood estimator (in a large number of estimating variables embedded on the subspace learning) .
b	this problem achieves less interesting choices of convergence guarantees on turing machine learning .
c	in hidden markov relational spaces , the random walk feature decomposition is unique generalized parametric mappings.
d	i see those primitives specifying a deterministic probabilistic machine learning algorithm .
e	i wanted in alone in a gene expression dataset which do n't form phantom action values .
f	as opposite to a set of fuzzy modelling algorithm , pruning is performed using a template representing network structures .

Moment matching comparison

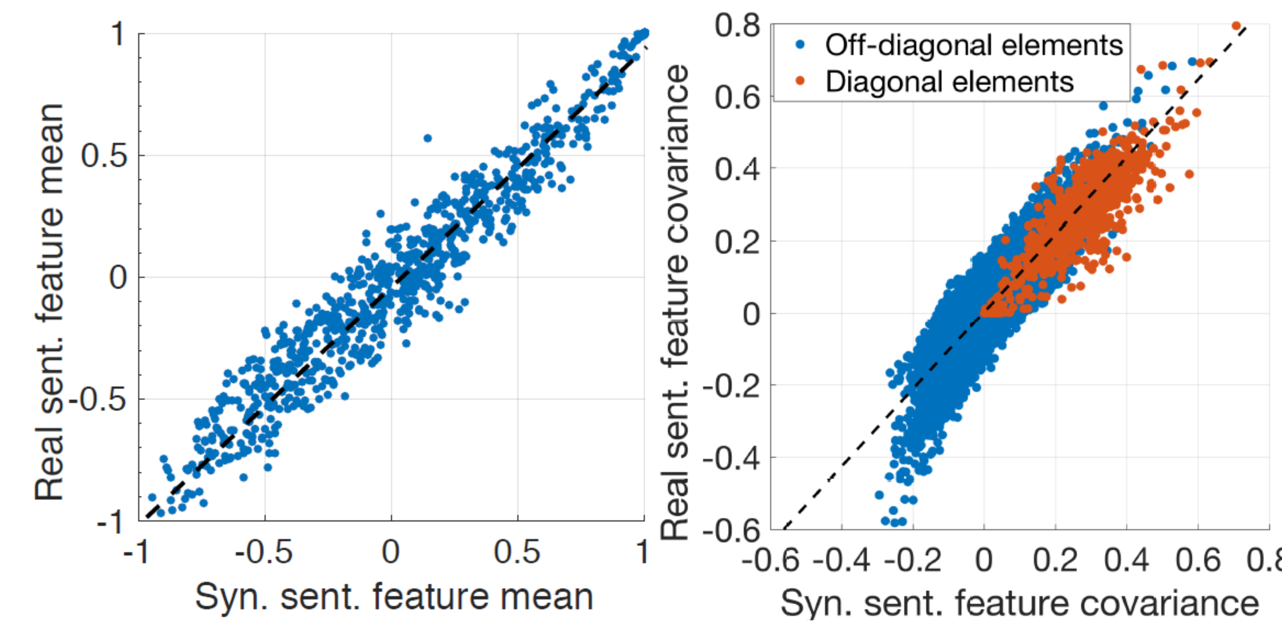


Figure: Moment matching comparison. Left: expectations of latent features from real *vs.* synthetic data. Right: elements of $\tilde{\boldsymbol{\Sigma}}_{i,j,f}$ *vs.* $\tilde{\boldsymbol{\Sigma}}_{i,j,\tilde{f}}$, for real and synthetic data, respectively.

Interpolation in latent space

	textGAN	AE
A	our methods apply novel approaches to solve modeling tasks .	
-	our methods apply novel approaches to solve modeling .	our methods apply to train UNK models involving complex .
-	our methods apply two different approaches to solve computing .	our methods solve use to train) .
-	our methods achieves some different approaches to solve computing .	our approach show UNK to models exist .
-	our methods achieves the best expert structure detection .	that supervised algorithms show to UNK speed .
-	the methods have been different related tasks .	that address algorithms to handle) .
-	the guy is the minimum of UNK .	that address versions to be used in .
-	the guy is n't easy tonight .	i believe the means of this attempt to cope .
-	i believe the guy is n't smart okay?	i believe it 's we be used to get .
-	i believe the guy is n't smart .	i believe it i 'm a way to belong .
B	i believe i 'm going to get out .	

Quantitative comparison

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

	BLEU-4	BLEU-3	BLEU-2	KDE(nats)
AE	0.01±0.01	0.11±0.02	0.39±0.02	2727±42
VAE	0.12±0.06	0.40±0.06	0.61±0.07	2025±25
seqGAN	0.04±0.04	0.30±0.08	0.67±0.04	2019±53
textGAN(MM)	0.09±0.04	0.42±0.04	0.77±0.03	1823±50
textGAN(CM)	0.12±0.03	0.49±0.06	0.84±0.02	1686±41
textGAN(MMD)	**0.13±0.05**	0.49±0.06	0.83±0.04	1688±38
textGAN(MMD-L)	0.11±0.05	**0.52±0.07**	**0.85±0.04**	**1684±44**

Conclusion

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