Training language GANs from Scratch

Cyprien de Masson d'Autume, Mihaela Rosca, Jack Rae, Shakir Mohamed

2019

Plan for today

- Recap (generative text models)
- Recap (GANs)
- Why do we need GANs for text?
- ScratchGAN architecture & Losses
- Evaluation & findings

Generative text modelling recap

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(x_t | x_1, ..., x_{t-1}) \qquad \hat{x}_t \sim p_{\boldsymbol{\theta}}(x_t | \hat{x}_1, ..., \hat{x}_{t-1})$$

Generative text modelling recap

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(x_t | x_1, ..., x_{t-1}) \qquad \hat{x}_t \sim p_{\boldsymbol{\theta}}(x_t | \hat{x}_1, ..., \hat{x}_{t-1})$$

$$\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p^*(\mathbf{x})} \log p_{\theta}(\mathbf{x})$$

GANs recap

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \mathbb{E}_{p^*(\mathbf{x})} \left[\log \mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x})\right] + \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{x})} \left[\log(1 - \mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x}))\right]$$

GANs recap

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \mathbb{E}_{p^*(\mathbf{x})}\left[\log \mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x})\right] + \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{x})}\left[\log(1 - \mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x}))\right]$$

REINFORCE gradient estimator

$$\nabla_\theta \mathbb{E}_{p_\theta(\mathbf{x})}[R(\mathbf{x})] = \mathbb{E}_{p_\theta(\mathbf{x})}\left[R(\mathbf{x}) \nabla_\theta \log p_\theta(\mathbf{x})\right]$$

From D: $R(\mathbf{x}) = \frac{p^*(\mathbf{x})}{p_\theta(\mathbf{x})}$

GANs for text (problems)

Train from scratch:

- gradient estimation
- mode collapse
- overfitting to the training set

GANs for text (problems)

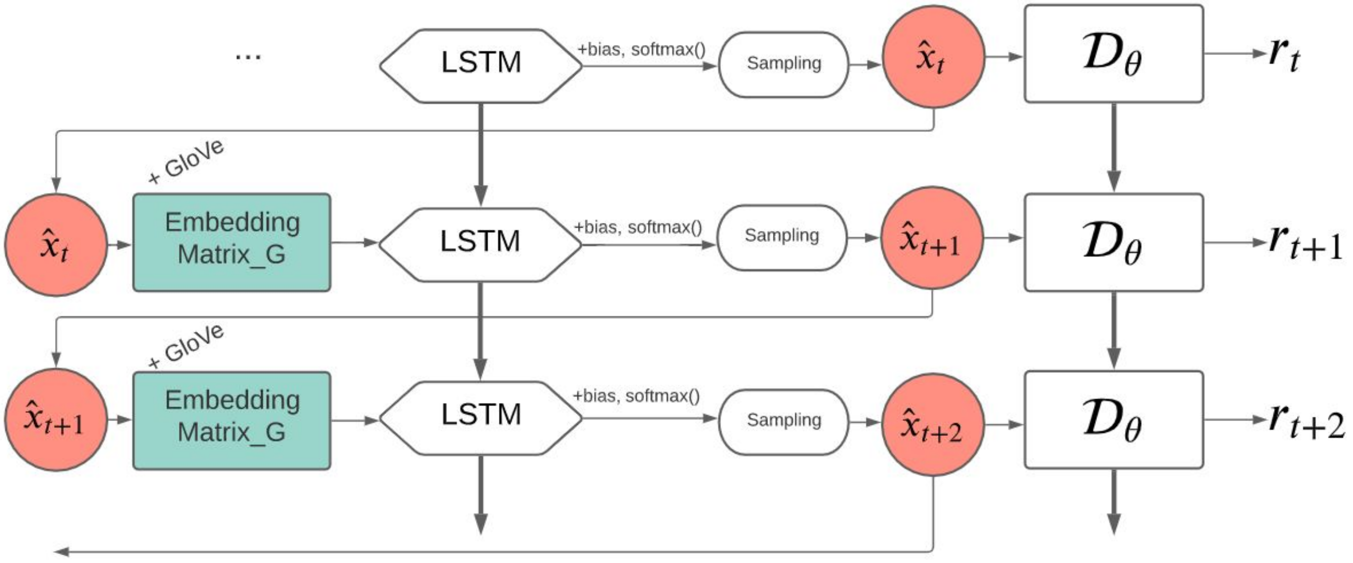
Train from scratch:

- gradient estimation
- mode collapse
- overfitting to the training set

Pretrain with MLE (adversarial fine-tuning):

- limited, close to the original, nothing new
- do not improve over maximum likelihood-trained models

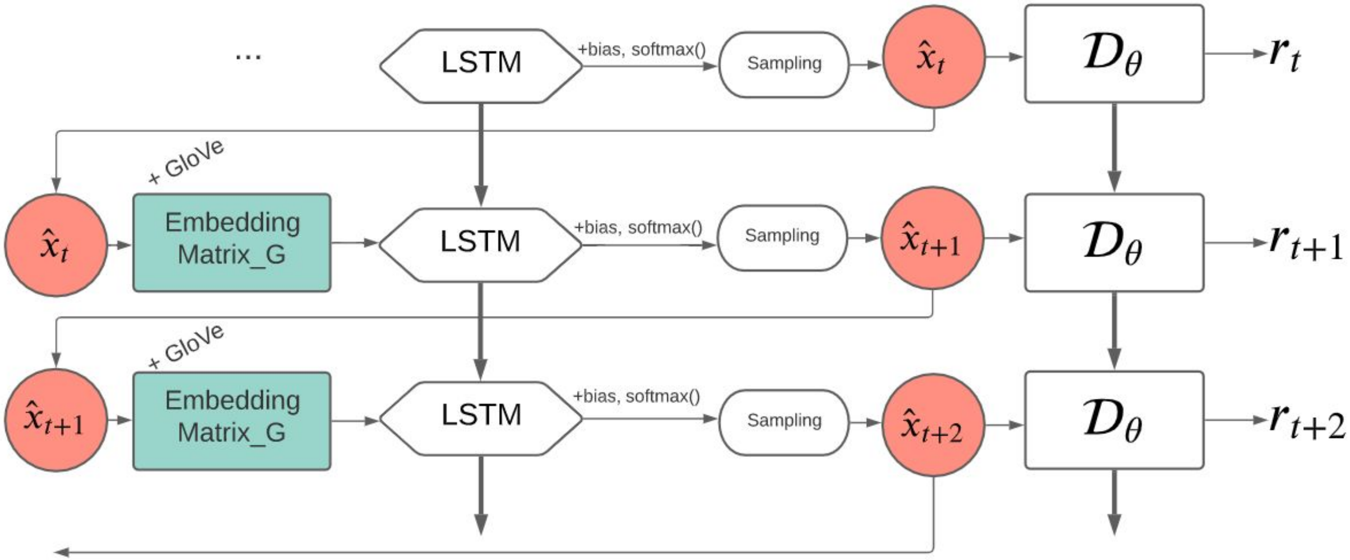
ScratchGAN



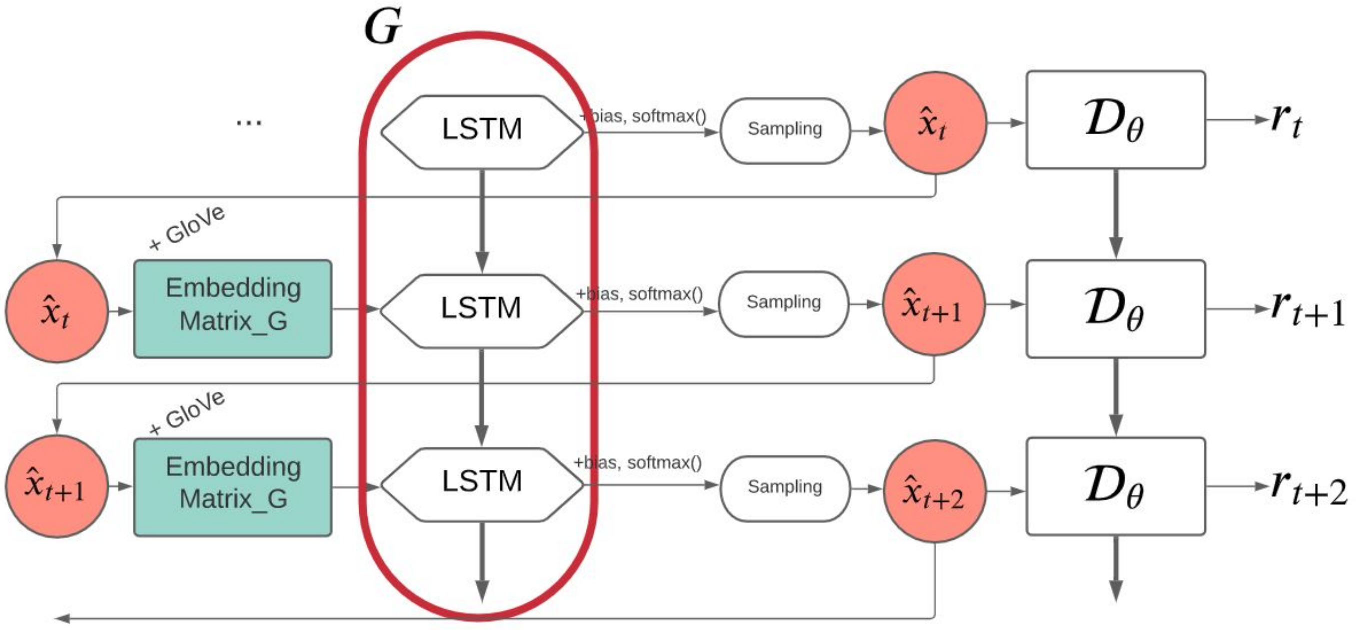
...
LSTM
+bias, softmax()
Sampling
$\hat{x}_t$
$\mathcal{D}_\theta$
r_t
+ GloVe
$\hat{x}_t$
Embedding Matrix_G
LSTM
+bias, softmax()
Sampling
$\hat{x}_{t+1}$
$\mathcal{D}_\theta$
r_{t+1}
+ GloVe
$\hat{x}_{t+1}$
Embedding Matrix_G
LSTM
+bias, softmax()
Sampling
$\hat{x}_{t+2}$
$\mathcal{D}_\theta$
r_{t+2}

ScratchGAN

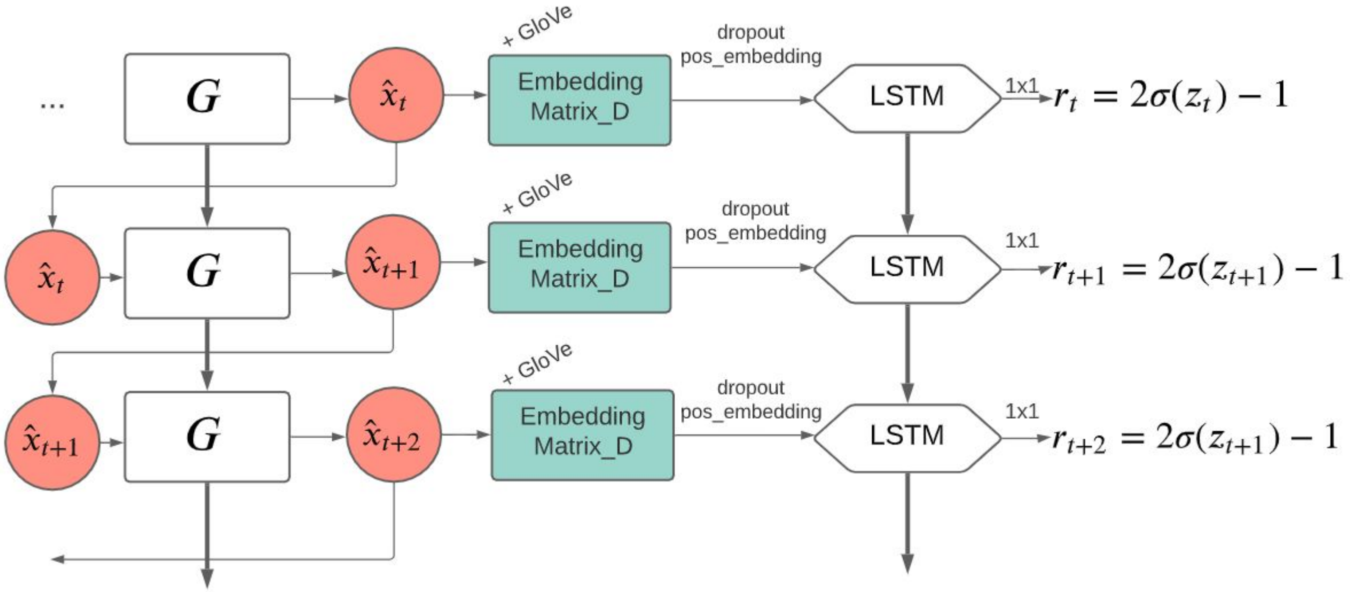
$$\boxed{r_t = 2\mathcal{D}_\phi(\hat{x}_t | x_{t-1} \ldots x_1) - 1}$$



Generator in detail

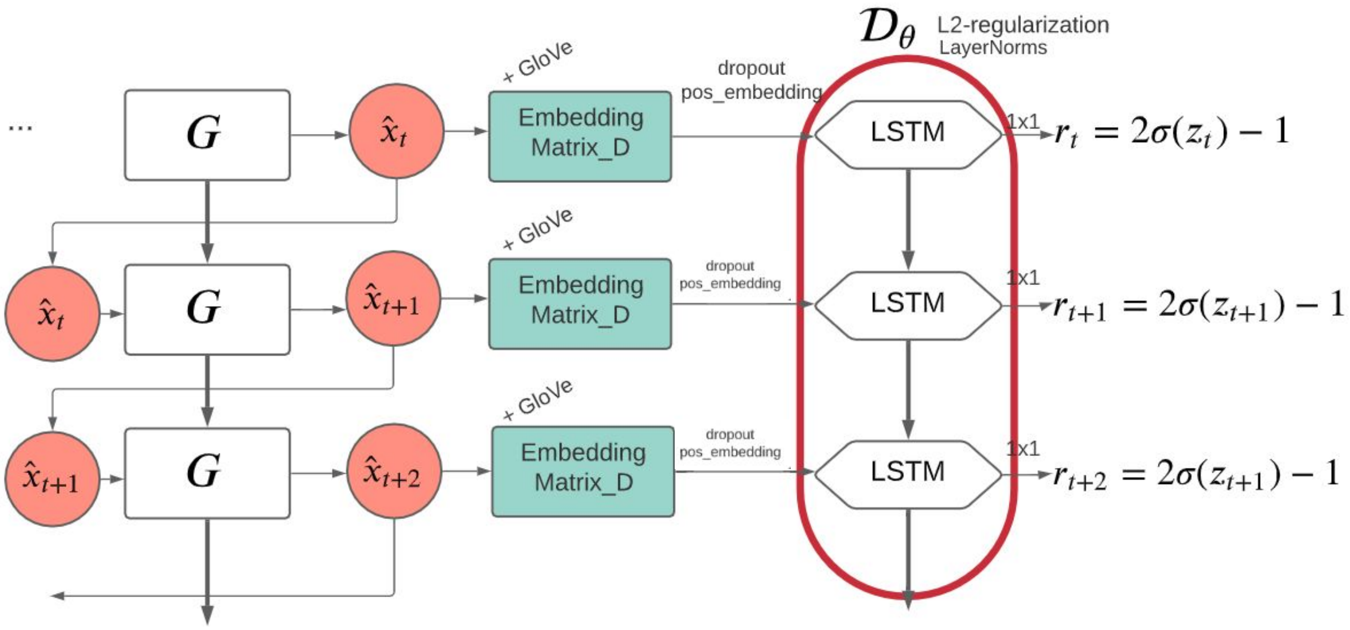


ScratchGAN



...
G
$\hat{x}_t$
+ GloVe
Embedding Matrix_D
dropout
pos_embedding
LSTM
1x1
$r_t = 2\sigma(z_t) - 1$
$\hat{x}_t$
G
$\hat{x}_{t+1}$
+ GloVe
Embedding Matrix_D
dropout
pos_embedding
LSTM
1x1
$r_{t+1} = 2\sigma(z_{t+1}) - 1$
$\hat{x}_{t+1}$
G
$\hat{x}_{t+2}$
+ GloVe
Embedding Matrix_D
dropout
pos_embedding
LSTM
1x1
$r_{t+2} = 2\sigma(z_{t+1}) - 1$

Discriminator in detail



G loss at time step i

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \mathbb{E}_{p^*(\mathbf{x})}\left[\log \mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x})\right] + \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{x})}\left[\log(1 - \mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x}))\right]$$

G loss at time step i

$$\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{x})}\left[\log(1 - \mathcal{D}_{\boldsymbol{\phi}}(\mathbf{x}))\right]$$

G loss at time step i

$$r_t = 2\mathcal{D}_{\boldsymbol{\phi}}(\hat{x}_t | x_{t-1}...x_1) - 1$$

G loss at time step i

$$r_t = 2\mathcal{D}_{\boldsymbol{\phi}}(\hat{x}_t | x_{t-1} ... x_1) - 1$$

$$R_t = \sum_{s=t}^{T} \gamma^{s-t} r_s$$

G loss at time step i

$$r_t = 2\mathcal{D}_\phi(\hat{x}_t | x_{t-1} \ldots x_1) - 1$$

$$R_t = \sum_{s=t}^{T} \gamma^{s-t} r_s$$

$$L_{ti}^{G} = \boxed{-(R_t - b_i) \ln p_\theta(x_t)}$$

$$\nabla_\theta \mathbb{E}_{p_\theta(\mathbf{x})}[R(\mathbf{x})] = \mathbb{E}_{p_\theta(\mathbf{x})}\left[\boxed{R(\mathbf{x}) \nabla_\theta \log p_\theta(\mathbf{x})}\right]$$

From D: $R(\mathbf{x}) = \frac{p^*(\mathbf{x})}{p_\theta(\mathbf{x})}$

G loss at time step i

$$r_t = 2\mathcal{D}_\phi(\hat{x}_t | x_{t-1} ... x_1) - 1$$

$$R_t = \sum_{s=t}^{T} \gamma^{s-t} r_s$$

$$L^G_{ti} = -(R_t - b_i) \ln p_\theta(x_t)$$

at training step i is $\sum_t L^G_{ti}$

G loss at time step i

$$r_t = 2\mathcal{D}_\phi(\hat{x}_t | x_{t-1} ... x_1) - 1$$

$$b_i = \lambda b_{i-1} + (1 - \lambda)\bar{R}_i$$

$$R_t = \sum_{s=t}^{T} \gamma^{s-t} r_s$$

$$L_{ti}^G = -(R_t - b_i) \ln p_\theta(x_t)$$

at training step i is $\sum_t L_{ti}^G$

G loss at time step i

$$r_t = 2\mathcal{D}_\phi(\hat{x}_t | x_{t-1} \ldots x_1) - 1$$

$$b_i = \lambda b_{i-1} + (1 - \lambda)\bar{R}_i$$

$$R_t = \sum_{s=t}^{T} \gamma^{s-t} r_s$$

$$L^G_{ti} = -(R_t - b_i) \ln p_\theta(x_t)$$

at training step i is $\sum_t L^G_{ti}$

$\bar{R}_i$ is the mean cumulative reward over all sequence timesteps and over the current batch

D loss is ...

D loss is just Cross Entropy!

D loss is just Cross Entropy!

$$\max_{\boldsymbol{\phi}} \sum_{t=1}^{T} \mathbb{E}_{p^*(x_t|x_1,\ldots,x_{t-1})}\left[\log \mathcal{D}_{\boldsymbol{\phi}}(x_t|x_1,...x_{t-1})\right] + \sum_{t=1}^{T} \mathbb{E}_{p_{\boldsymbol{\theta}}(x_t|x_1,\ldots,x_{t-1})}\left[\log(1-\mathcal{D}_{\boldsymbol{\phi}}(x_t|x_1,...x_{t-1}))\right]$$

D loss is just Cross Entropy!

$$\max_{\boldsymbol{\phi}} \sum_{t=1}^{T} \mathbb{E}_{p^*(x_t|x_1,\dots,x_{t-1})}\left[\log \mathcal{D}_{\boldsymbol{\phi}}(x_t|x_1,\dots x_{t-1})\right] + \sum_{t=1}^{T} \mathbb{E}_{p_{\boldsymbol{\theta}}(x_t|x_1,\dots,x_{t-1})}\left[\log(1-\mathcal{D}_{\boldsymbol{\phi}}(x_t|x_1,\dots x_{t-1}))\right]$$

1 step G; 1 step D

Key ScratchGAN components

Key ScratchGAN components

- Recurrent D & REINFORCE

Key ScratchGAN components

- Recurrent D & REINFORCE
- D regularization (L2 & LNs)

Key ScratchGAN components

- Recurrent D & REINFORCE
- D regularization (L2 & LNs)
- Large batch size ?

Key ScratchGAN component is large batch size

$$\nabla_\theta = \sum_{n=1}^{N} \sum_{t=1}^{T} (R_t^n - b_t) \nabla_\theta \log p_{\boldsymbol{\theta}}(\hat{x}_t^n | \hat{x}_{t-1}^n ... \hat{x}_1^n), \qquad \hat{x}_t^n \sim p_{\boldsymbol{\theta}}(x_t^n | \hat{x}_{t-1}^n ... \hat{x}_1^n)$$

G optimization

Key ScratchGAN component is large batch size

$$\nabla_\theta = \sum_{n=1}^{N} \sum_{t=1}^{T} (R_t^n - b_t) \nabla_\theta \log p_{\boldsymbol{\theta}}(\hat{x}_t^n | \hat{x}_{t-1}^n \ldots \hat{x}_1^n), \qquad \hat{x}_t^n \sim p_{\boldsymbol{\theta}}(x_t^n | \hat{x}_{t-1}^n \ldots \hat{x}_1^n)$$

G optimization

Evaluation difficulties

Perplexity

Model	**World level perplexity**
Random	5725
ScratchGAN	154
MLE	**42**

Table 2: EMNLP2017 News perplexity.

LSTM

EMNLP2017 News

Evaluation difficulties

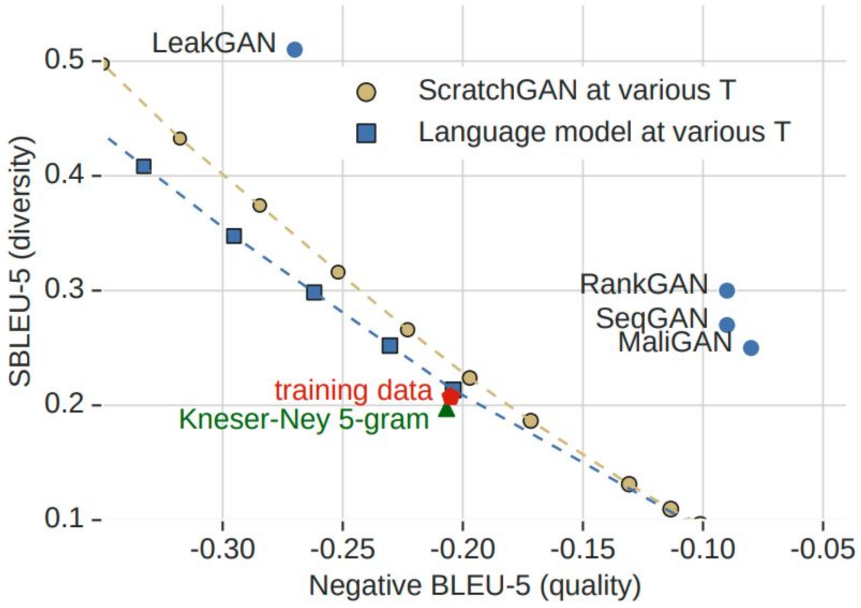
- Diversity and quality

BLEU

- local consistency and detect relatively simple problems with syntax
- DOES NOT capture semantic variation

BLEU

- local consistency and detect relatively simple problems with syntax
- DOES NOT capture semantic variation



(a) Negative BLEU-5 versus Self-BLEU-5.

EMNLP2017 News

Evaluation difficulties

- Diversity and quality ✓
- Local and global consistency

Fréchet Embedding Distance (FED)

- Universal Sentence Encoder

- Distance between two Gaussian distributions fitted to data embeddings

https://tfhub.dev/google/universal-sentence-encoder/2

Fréchet Embedding Distance (FED)

- Universal Sentence Encoder
- Distance between two Gaussian distributions fitted to data embeddings

- Global consistency
- Both quality and diversity
- Correlates with human evaluation
- Less sensitive to word order than BLEU metrics
- FID proven useful for images.

https://tfhub.dev/google/universal-sentence-encoder/2

Evaluation difficulties

- Diversity and quality ✓
- Local and global consistency ✓
- Generalization beyond the training set

USE	Nearest Neighbours	3-gram	Nearest Neighbours
Sample: *A nice large part of Trump has to plan exactly what Pence would worth , for Trump to choose him strongly in Florida, where he can be 100 percent away.*			
0.77	His name , of course , is Donald Trump , the billionaire businessman who leads most national polls for the Republican nomination .	0.13	It ' s like the situation in Florida , where he didn ' t pay taxes on his golf course .
0.75	But to get there , Rubio believes he needs to cut significantly into Cruz ' s support in Iowa , a state dominated by social conservatives .	0.12	Donald Trump is spending his third straight day in Florida , where he ' s already made six campaign stops since Sunday .
0.72	On the Republican side , the Iowa poll shows Ted Cruz leading Donald Trump by four points , but Trump has a 16 - point lead in New Hampshire .	0.10	He has long been mentioned as a possible candidate for governor in Florida , where he has a home in Miami with his wife and four school - age children .

EMNLP2017 News

Evaluation difficulties

- Diversity and quality
- Local and global consistency
- Generalization beyond the training set

Training stability

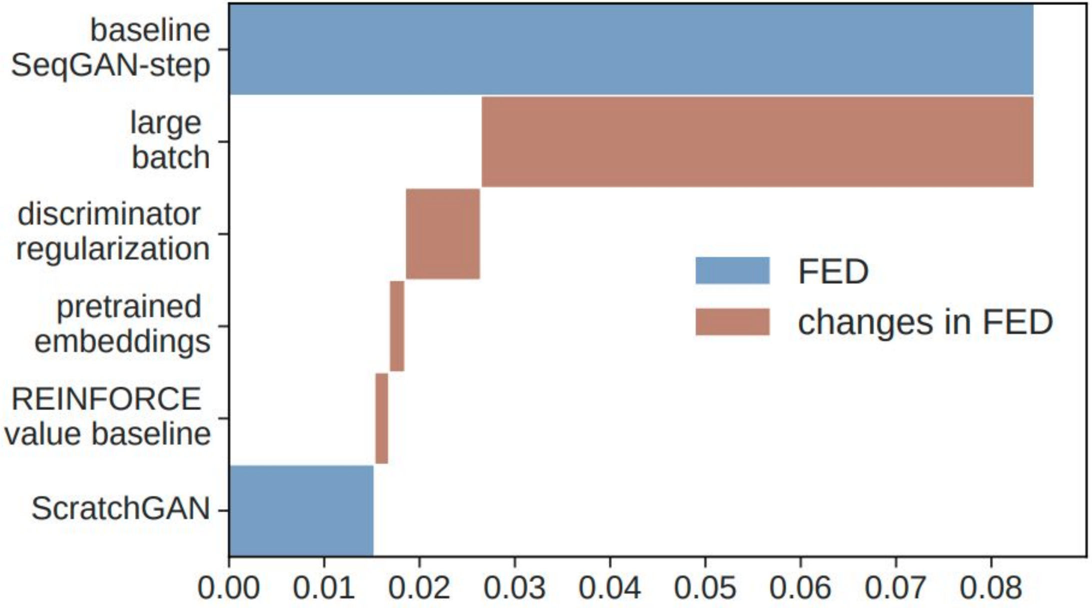
-

Table 4: FED sensitivity on EMNLP2017 News.

Variation	FED
Hyperparameters	0.021 ± 0.0056
Seeds (best hypers)	0.018 ± 0.0008

Ablation study

-



(c) ScratchGAN ablation study.

EMNLP2017 News

C Negative results

Here we list some approaches that we tried but which proved unsuccessful or unnecessary:

- Using a Wasserstein Loss on generator logits, with a straight-through gradient. This was unsuccessful.
- Using ensembles of discriminators and generators. The results are on par with those obtained by a single discriminator-generator pair.
- Training against past versions of generators/discriminators. Same as above.
- Using bi-directional discriminators. They can work but tend to over-fit and provide less useful feedback to the generator.
- Using several discriminators with different architectures, hoping to have the simple discriminators capture simple failure modes of the generators such as repeated words. It did not improve over single discriminator-generator pair.
- Training on small datasets such as Penn Tree Bank. The discriminator quickly over-fit to the training data. This issue could probably be solved with stronger regularization but we favoured larger datasets.
- Using a Hinge loss [44] on the discriminator. This did not improve over the cross-entropy loss.
- Using a hand-designed curriculum, where the generator is first trained against a simple discriminator, and later in training a more complex discriminator is substituted. This was unsuccessful. We suspect that adversarial training requires a difficult balance between discriminator quality and generator quality, which is difficult to reach when either component has been trained independently from the other.

Thanks for your attention!