Training language GANs from scratch

Mihaela Rosca
mihaelacr@google.com

Thanks to:
Cyprien de Masson d'Autume
Jack Rae
Shakir Mohamed

https://arxiv.org/abs/1905.09922

dialogue
question answering
translation
transcription
agent interaction

State of the art in language models

Maximum Likelihood training

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

Confined to autoregressive structures

$$p_\theta(\mathbf{x}) = \prod_{t=1}^{T} p_\theta(x_t \mid x_1, \ldots, x_{t-1})$$

Main improvements come from architecture changes (eg.Transformer)

MLE trained language models

Issues:

- MLE: too much mass around the data distribution
- Exposure bias: teacher forcing affecting sample quality

Alternative - GANs!

Intuition: generate data which is indistinguishable from real data according to a trained discriminator.

Great success with natural images:

https://arxiv.org/pdf/1809.11096.pdf

Why are text gans hard?

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

- gradient estimation
  - high variance (REINFORCE)
  - biased gradients (Gumbel Softmax/Concrete)
- discriminator/reward structure
- easy task for the discriminator early in training

State of current language GANs

Maximum likelihood pretraining.

But...

- “[current language GANs] barely achieve non-random results without supervised pre-training” [1]
- “the best-performing GANs tend to stay close to the solution given by maximum-likelihood training” [2]
- no performance GAN over MLE models [1, 2]

[1] https://arxiv.org/abs/1806.04936
[2] https://arxiv.org/abs/1811.02549

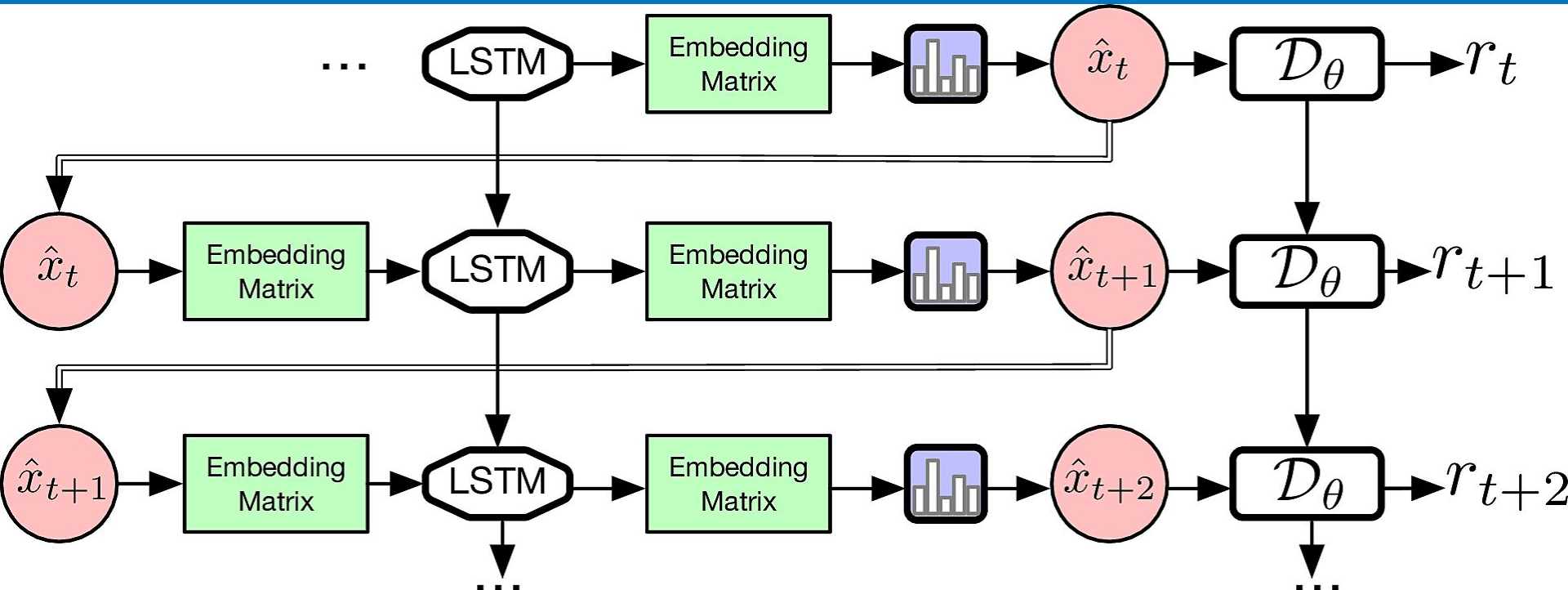
Solving text GANs - ScratchGAN

- Gradient variance
  - large batch sizes for REINFORCE
  - moving average baselines
- Dense rewards
- Discriminator regularization
  - layer normalization
  - dropout
  - l2 norm
- Model architecture

https://arxiv.org/abs/1905.09922

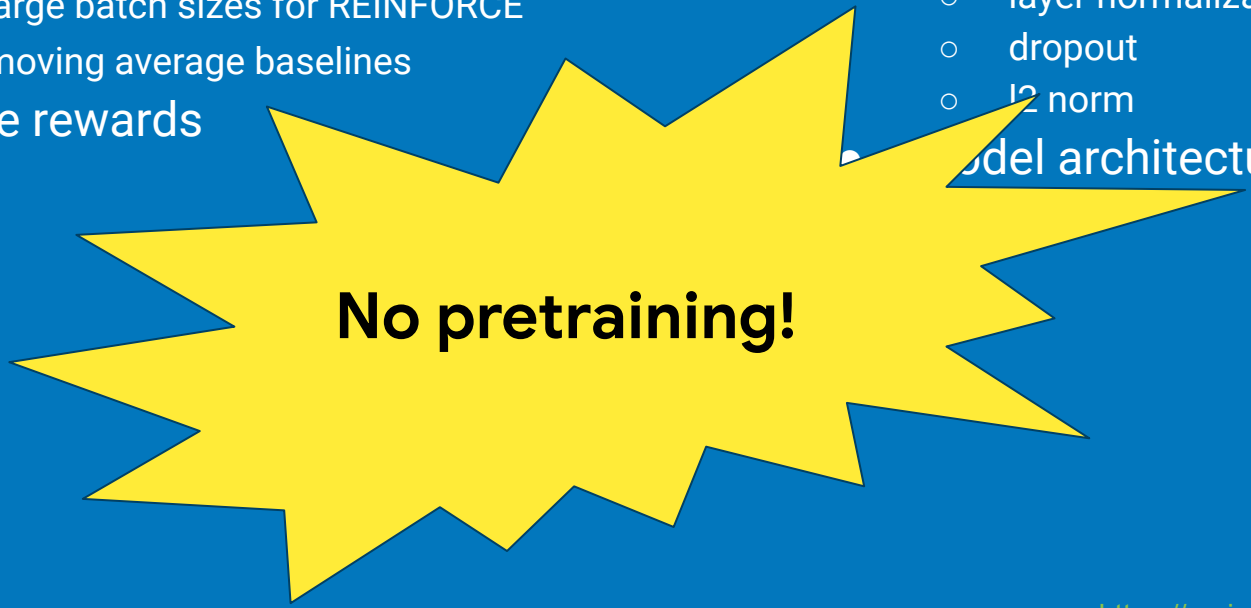
ScratchGAN

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



Solving text GANs - ScratchGAN

- Gradient variance
  - large batch sizes for REINFORCE
  - moving average baselines
- Dense rewards
- Discriminator regularization
  - layer normalization
  - dropout
  - l2 norm
- odel architecture



https://arxiv.org/abs/1905.09922

What we tried but did not work

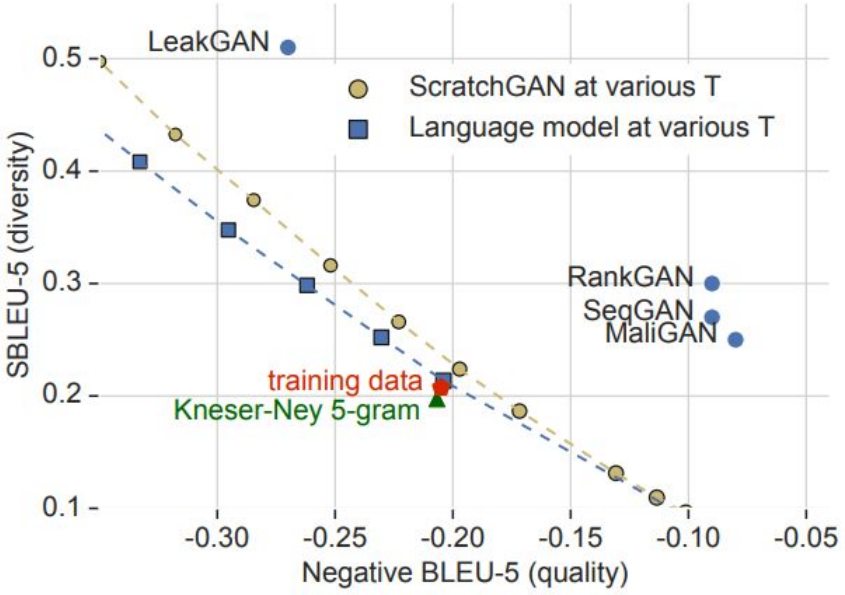
- Using a Wasserstein Loss on generator logits, with a straight-through gradient.
- Using ensembles of discriminators and generators.
- Training against past versions of generators/discriminators.
- Using a hand-designed curriculum.
- Other losses (Hinge loss) for the discriminator.
- Small datasets like Penn Tree Bank - overfitting.
- More.... (see paper)

Evaluating ScratchGAN

Datasets: EMNLP News 2017, WikiText 1.03

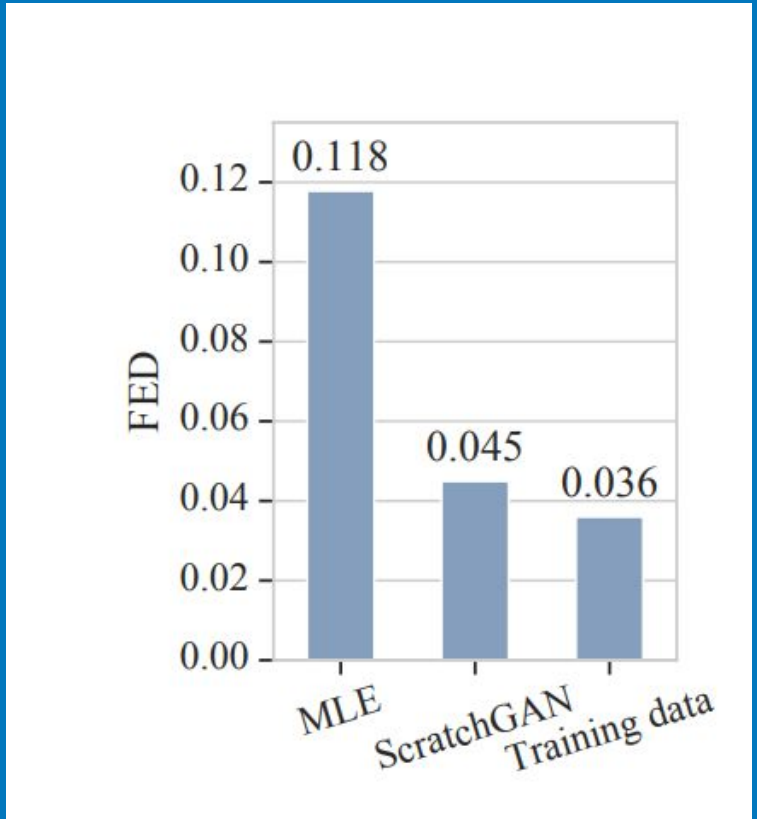
- N-gram metrics
  - local consistency
- Frechet Distance
- Longest matching n-grams
- Note: Can use perplexity
- More ... (see paper)

Results - local consistency (BLEU scores)



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

Results - global consistency (FED)



Lower is better.

Results - overfitting

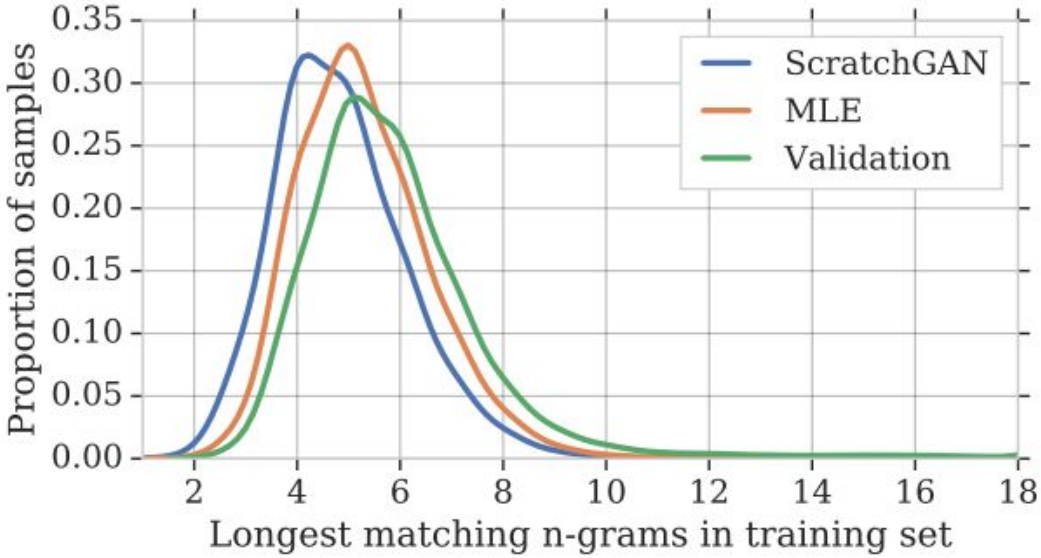


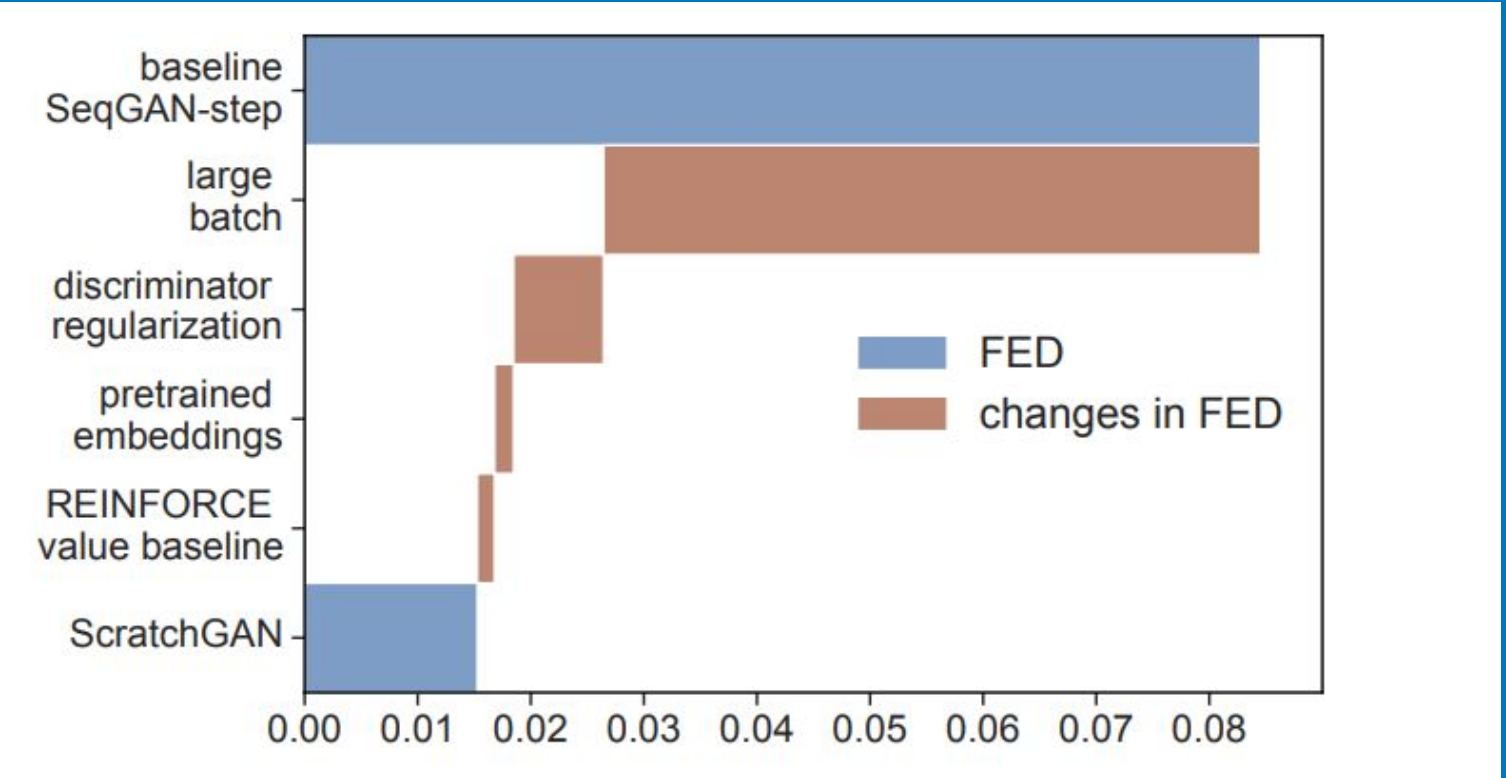
Figure 4: Matching n-grams in EMNLP2017.

Results - perplexity

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

Lower is better.

Ablation experiments



Next steps

- Further reduce variance by reducing vocabulary size
- Better architectures - transformer XL
- More data
- Removing autoregressivity

Thanks!

We are pleased for the trust and it was incredible , our job quickly learn the shape and get on that way.

There is task now that the UK will make for the society to seek secure enough government budget fund reduce the economy.

Keith is also held in 2005 and Ted ' s a successful campaign spokeswoman for students and a young brothers has took an advantage of operator .