MASKGAN: BETTER TEXT GENERATION VIA FILLING IN THE _____

William Fedus, Ian Goodfellow, Andrew M. Dai

Presented by: Joey Bose

February 16, 2018
Motivation

- Maximum Likelihood RNN’s are the most common generative model for sequences
- Teacher Forcing leads to unstable dynamics in the hidden states
- Professor Forcing does solve the above but does not encourage high sample quality
- GAN’s have shown incredible quality samples for images but discrete nature of text makes training a generator harder.
- Reinforcement Learning framework can be leveraged to train the generator by policy gradients
Teacher Forcing

Teacher forcing is a method for quickly and efficiently training recurrent neural network models that use the \textbf{true} output rather than the model output from a prior time step as input. (Figure taken from http://www.cedar.buffalo.edu/~srihari/CSE676)

\begin{center}
\textbf{Train time:}
We feed the correct output $y^{(t)}$ (from teacher) drawn from the training set as input to $h^{(t+1)}$
\end{center}

\begin{center}
\textbf{Test time:}
True output is not known. We approximate the correct output $y^{(t)}$ with the model's output $o^{(t)}$ and feed the output back to the model
\end{center}
Professor Forcing

- Training RNN's such that the network is indistinguishable during training and when sequences are self-generated.
- Train a second model to match the distribution of training and self-generated sequences
GAN training involves jointly training both a generator and discriminator network in a min-max setting where the value function is $\min_G \max_D V(D, G)$:

$$\mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$, \hspace{1cm} (1)

Similarly one can add side channel information and condition the generator:

$$\mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z|y)))]$$, \hspace{1cm} (2)

Common failure modes for GAN's include unstable training dynamics and mode collapse, which are exacerbated in a discrete setting.
Conditional GAN
SeqGAN

True data
Real World
Generate

Train

Next action
MC search

State

Reward

Policy Gradient
Start with a ground truth discrete sequence $x = (x_1, ..., x_T)$ and a binary mask of the same length, $m = (m_1, ..., m_T)$. Applying the mask on the input sequence creates, $m(x)$, a sequence with blanks:

For example:

$$x = a, b, c, d, e$$
$$m = 0, 1, 1, 0, 0$$
$$m(x) = a, _, _, d, e$$

The goal of the generator is then to autoregressively fill in the missing tokens conditioned on the previous tokens and the mask.
The Generator for MaskGAN is Seq2Seq network and models the following distribution:

\[
P(\hat{x}_1, \ldots, \hat{x}_T | m(x)) = \prod_{t=1}^{T} P(\hat{x}_t | \hat{x}_1, \ldots, \hat{x}_{t-1}, m(x)), \tag{3}
\]

\[
G(x_t) = P(\hat{x}_t | \hat{x}_1, \ldots, \hat{x}_{t-1}, m(x)), \tag{4}
\]
The discriminator is also a Seq2Seq network and receives the filled in sequence and \( m(x) \).

Failure Mode, Example

G: The director director guided the series
Possible Label 1: The *associate* director guided the series Possible Label 2: The director *expertly* guided the series

This gives an ambiguous signal to Generator when \( m(x) \) is not given as the generator doesn’t know which of the directors tokens is wrong. The discriminator computes \( P(\tilde{x}_t = x_t^{real} | \tilde{x}_0:T, m(x)) \)
**Reward** is \( \log \) of discriminator output \( r_t = \log P(\tilde{x}_t = x_t^{\text{real}}|x_0:T, m(x)) \)

- Additional Critic Head of the discriminator and estimates the value function \( R_t = \sum_{s=t}^{T} \gamma^s r_s \)
- The generator maximizes expected reward \( \mathbb{E}_{G(\theta)}[R] \) using the REINFORCE gradient estimator with the Critic as a baseline

\[
\nabla_{\theta} \mathbb{E}[R] = (R_t - b_t) \nabla_{\theta} \log G_{\theta}(\tilde{x}_t) \quad (5)
\]

The gradient to the generator for token \( \tilde{x}_t \) depends on all discounted future rewards assigned by the discriminator. The discriminator update is the same as vanilla GAN.
There are two stages of pretraining:

- Pretrain a language model with Maximum Likelihood and then use its weights to initialize Seq2Seq
- Pretrain Seq2Seq model on infilling task using Maximum Likelihood
- Select model with lowest validation perplexity on infilling task over 500 runs
- Authors report that adding a critic decreases variance by an order of magnitude
- Authors claim pretraining is not necessary for best results but significantly reduces computation time
Bleu Score for MaskGAN

Compare the number of unique n-grams produced by the Generator with the n-grams of the validation corpus and count the number of matches. Then compute the geometric average.

- Reducing validation perplexity does not equate to high sample quality.
- Generator can explore off-manifold and find alternative valid samples.
- This is justified as it is known that perplexity is not the best metric for language models.
Two modes of operation
1. Conditional Mode
2. Free-Running/Unconditional

Conditional Mode is the Regular MaskGAN with deterministic / stochastic mask applied to the input sequence. Free-Running mode is when the entire input sequence is masked reducing this task to traditional language modeling.
### Table 1: Conditional samples from PTB for both MaskGAN and MaskMLE models.

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>MaskGAN</th>
<th>MaskMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>the next day ’s show &lt;eos&gt; interactive telephone technology has taken a new leap in &lt;unk&gt; and television programmers are</td>
<td>the next day ’s show &lt;eos&gt; interactive telephone technology has taken a new leap in its retail business &lt;eos&gt; a</td>
<td>the next day ’s show &lt;eos&gt; interactive telephone technology has taken a new leap in the complicate case of the</td>
</tr>
<tr>
<td>MaskGAN</td>
<td>oct. N as the end of the year the resignations were approved &lt;eos&gt; the march N N &lt;unk&gt; was down</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Language model (unconditional) sample from PTB for MaskGAN.
<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>Pitch Black was a complete shock to me when I first saw it back in 2000 In the previous years I</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskGAN</td>
<td>Pitch Black was a complete shock to me when I first saw it back in 1979 I was really looking forward</td>
</tr>
<tr>
<td>MaskMLE</td>
<td>Black was a complete shock to me when I first saw it back in 1969 I live in New Zealand</td>
</tr>
</tbody>
</table>

Table 3: Conditional samples from IMDB for both MaskGAN and MaskMLE models.
| MaskGAN | **Positive**: Follow the Good Earth movie linked Vacation is a comedy that credited against the modern day era yarns which has helpful something to the modern day s best It is an interesting drama based on a story of the famed |

Table 4: Language model (unconditional) sample from IMDB for MaskGAN.
Perplexity of Generated Samples and Mode Collapse

Test the perplexity of MaskGAN and MaskMLE under the pretrained language model used to initialize it.

<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity of IMDB samples under a pretrained LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskMLE</td>
<td>273.1 ± 3.5</td>
</tr>
<tr>
<td>MaskGAN</td>
<td>108.3 ± 3.5</td>
</tr>
</tbody>
</table>

Evidence of Mode Dropping

<table>
<thead>
<tr>
<th>Model</th>
<th>% Unique bigrams</th>
<th>% Unique trigrams</th>
<th>% Unique quadgrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>40.6</td>
<td>75.2</td>
<td>91.9</td>
</tr>
<tr>
<td>MaskMLE</td>
<td>43.6</td>
<td>77.4</td>
<td>92.6</td>
</tr>
<tr>
<td>MaskGAN</td>
<td>38.2</td>
<td>70.7</td>
<td>88.2</td>
</tr>
</tbody>
</table>
### Human Evaluation on Mechanical Turk

<table>
<thead>
<tr>
<th>Preferred Model</th>
<th>Grammaticality %</th>
<th>Topicality %</th>
<th>Overall %</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>15.3</td>
<td>19.7</td>
<td>15.7</td>
</tr>
<tr>
<td>MaskGAN</td>
<td>59.7</td>
<td>58.3</td>
<td>58.0</td>
</tr>
<tr>
<td>LM</td>
<td>20.0</td>
<td>28.3</td>
<td>21.7</td>
</tr>
<tr>
<td>MaskMLE</td>
<td>42.7</td>
<td>43.7</td>
<td>40.3</td>
</tr>
<tr>
<td>MaskGAN</td>
<td>49.7</td>
<td>43.7</td>
<td>44.3</td>
</tr>
<tr>
<td>MaskMLE</td>
<td>18.7</td>
<td>20.3</td>
<td>18.3</td>
</tr>
<tr>
<td>Real samples</td>
<td>78.3</td>
<td>72.0</td>
<td>73.3</td>
</tr>
<tr>
<td>LM</td>
<td>6.7</td>
<td>7.0</td>
<td>6.3</td>
</tr>
<tr>
<td>Real samples</td>
<td>65.7</td>
<td>59.3</td>
<td>62.3</td>
</tr>
<tr>
<td>MaskGAN</td>
<td>18.0</td>
<td>20.0</td>
<td>16.7</td>
</tr>
</tbody>
</table>

Table 7: A Mechanical Turk blind heads-up evaluation between pairs of models trained on IMDB reviews. 100 reviews (each 40 words long) from each model are unconditionally sampled and randomized. Raters are asked which sample is preferred between each pair. 300 ratings were obtained for each model pair comparison.
1. Mode Dropping is less extreme than SeqGAN but still noticeable.

   It is a very funny film that is very funny It s a very funny movie and it s charming It

2. Matching Syntax at Boundaries

   Cartoon is one of those films me when I first saw it back in 200

3. Loss of Global Context

   This movie is terrible The plot is ludicrous The title is not more interesting and original This is a great movie Lord of the Rings was a great movie John Travolta is brilliant
The End