MASKGAN: TEXTUAL GENERATIVE ADVERSARIAL NETWORKS FROM FILLING-IN-THE-BLANK

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural text generation models are often autoregressive language models or seq2seq models. Neural autoregressive and seq2seq models that generate text by sampling words sequentially, with each word conditioned on the previous model, are state-of-the-art for several machine translation and summarization benchmarks. These benchmarks are often defined by validation perplexity even though this is not a direct measure of sample quality. Language models are typically trained via maximum likelihood and most often with teacher forcing. Teacher forcing is well-suited to optimizing perplexity but can result in poor sample quality because generating text requires conditioning on sequences of words that were never observed at training time. We propose to improve sample quality using Generative Adversarial Networks (GANs), which explicitly train the generator to produce high quality samples and have shown a lot of success in image generation. GANs were originally designed to output differentiable values, so discrete language generation is challenging for them. We introduce an actor-critic conditional GAN that fills in missing text conditioned on the surrounding context. We show qualitatively and quantitatively, evidence that this produces more realistic text samples compared to a maximum likelihood trained model.

1 INTRODUCTION

Recurrent Neural Networks (RNNs) \cite{Graves2012} are the most common generative model for sequences as well as for sequence labeling tasks. They have shown impressive results in language modeling \cite{Mikolov2010}, speech recognition \cite{Graves2013}, machine translation \cite{Wu2016} and text classification \cite{Miyato2016}. Text is typically generated from these models by sampling from a distribution that is conditioned on the previous word and a hidden state that consists of a representation of the words generated so far. These are typically trained with maximum likelihood in an approach known as teacher forcing, where ground-truth words are fed back into the model to be conditioned on for generating the following parts of the sentence. This causes problems when, during sample generation, the model is often forced to condition on sequences that were never conditioned on at training time. This leads to unpredictable dynamics in the hidden state of the RNN. Methods such as Professor Forcing \cite{Lamb2016} and Scheduled Sampling \cite{Bengio2015} have been proposed to solve this issue. These approaches work indirectly by either causing the hidden state dynamics to become predictable (Professor Forcing) or by randomly conditioning on sampled words at training time, however, they do not directly specify a cost function on the output of the RNN encouraging high sample quality. Our proposed method does so.

Generative Adversarial Networks (GANs) \cite{Goodfellow2014} are a framework for training generative models in an adversarial setup, with a generator generating images that is trying to fool a discriminator that is trained to discriminate between real and synthetic images. GANs have had a lot of success in producing more realistic images than other approaches but they have only been seen limited use for text sequences. This is due to the discrete nature of text making it infeasible to propagate the gradient from the discriminator back to the generator as in standard GAN training. We overcome this by using Reinforcement Learning (RL) to train the generator while the discriminator is still trained via maximum likelihood and stochastic gradient descent. GANs also commonly suffer from issues such as training instability and mode dropping, both of which are exacerbated in a textual setting. Mode dropping occurs when certain modalities in the training set are rarely generated by the
generator, for example, leading all generated images of a volcano to be the multiple variants of the same volcano. This becomes a significant problem in text generation since there are many complex modes in the data, ranging from bigrams to short phrases to longer idioms. Training stability is also an issue since unlike image generation, text is typically generated autoregressively and thus the loss from the discriminator is only observed after a complete sentence has been generated. This problem compounds when generating longer and longer sentences.

We reduce the impact of these problems by training our model on a text fill-in-the-blank or in-filling task. This is similar to the task proposed in Bowman et al. (2015) but we set up this task more robustly. In this task, portions of a body of text are deleted or redacted. The goal of the model is to then infill the missing portions of text so that it is indistinguishable from the original data. While in-filling text, the model operates autoregressively over the tokens it has thus far filled in, as in standard language modeling, while conditioning on the true known context. If the entire body of text is redacted, then this reduces to language modeling.

Designing error attribution per time step has been noted to be important in prior natural language GAN research (Yu et al., 2017; Li et al., 2017). The text infilling task naturally achieves this consideration since our discriminator will evaluate each token and thus provide a fine-grained supervision signal to the generator. Consider, for instance, if the generator produces a sequence perfectly matching the data distribution over the first \( t-1 \) time-steps, but then produces an outlier token \( y_t, (x_1:t-1, y_t) \). Despite the entire sequence now being clearly synthetic as a result of the errant token, a discriminative model that produces a high loss signal to the outlier token, but not to the others, will likely yield a more informative error signal to the generator.

This research also opens further inquiry of conditional GAN models in the context of natural language. There have been numerous impressive results in the continuous domain using conditional GAN approaches, beginning with the work of Mirza & Osindero (2014). In the work of Odena et al. (2016), the model is able to produce all 1000 classes of ImageNet Russakovsky et al. (2015) while conditioned on the class label. Another example of success of conditional GAN models was the transfer of style, object transfiguration and photo enhancement recently demonstrated by CycleGAN Zhu et al. (2017).

In the following sections,

- We introduce a text generation model trained on in-filling (MaskGAN).
- Consider the actor-critic architecture in extremely large action spaces.
- Consider new evaluation metrics and the generation of synthetic training data.

2 RELATED WORK

Research into reliably extending GAN training to discrete spaces and discrete sequences has been a highly active area. GAN training in a continuous setting allows for fully differentiable computations, permitting gradients to be passed through the discriminator to the generator. Discrete elements break this differentiability, leading researchers to either avoid the issue and reformulate the problem, work in the continuous domain or to consider RL methods.

SeqGAN (Yu et al., 2017) trains a language model by using policy gradients to train the generator to fool a CNN-based discriminator that discriminates between real and synthetic text. Both the generator and discriminator are pretrained on real and fake data before the phase of training with policy gradients. During training they then do Monte Carlo rollouts in order to get a useful loss signal per word.

Professor Forcing (Lamb et al., 2016) is an alternative to training an RNN with teacher forcing by using a discriminator to discriminate the hidden states of a generator RNN that is conditioned on real and synthetic samples. Since the discriminator only operates on hidden states, gradients can be passed through to the generator so that the hidden state dynamics at inference time follow those at training time.

Discrete GANs have been applied for translation (Yang et al., 2017) where they produced an improvement in BLEU score for Chinese-English translation. GANs have also been applied to dialogue generation (Li et al., 2017) showing improvements in adversarial evaluation and good results with
Human evaluation compared to a maximum likelihood trained baseline. Their method applies REINFORCE with Monte Carlo sampling on the generator.

Replacing the non-differentiable sampling operations with efficient gradient approximators (Jang et al., 2016; Maddison et al., 2016) led to some work applying GANs in toy tasks, (Kusner & Hernández-Lobato, 2016), however, there has not been strong evidence of the efficacy of these tools yet. Recent unbiased and low variance gradient estimate techniques such as Tucker et al. (2017) may prove more effective.

WGAN-GP (Gulrajani et al., 2017) avoids the issue of dealing with backpropagating through discrete nodes by generating text in a one-shot manner using a 1D convolutional network. Hjelm et al. (2017) proposes an algorithmic solution and uses a boundary-seeking GAN objective along with importance sampling to generate text. In Rajeswar et al. (2017), the discriminator operates directly on the continuous probabilistic output of the generator. However, to accomplish this, they recast the traditional autoregressive sampling of the text since the inputs to the RNN are predetermined. Che et al. (2017) instead optimize a lower-variance objective using the discriminator’s output, rather than the standard GAN objective.

Reinforcement learning methods have been explored successfully in natural language. Using a REINFORCE and cross entropy hybrid, MIXER, Ranzato et al. (2015) directly optimized BLEU score Papineni et al. (2002) and demonstrated improvements over baselines. More recently, actor-Critic methods in natural language were explored in Bahdanau et al. (2016) where instead of having rewards supplied by a discriminator in an adversarial setting.

Conditional text generation via GAN training has been explored in Rajeswar et al. (2017); Li et al. (2017).

Our work is distinct in that we employ an actor-critic training procedure on a task designed to provide rewards at every time step Li et al. (2017). We believe the in-filling may mitigate the problem of severe mode-collapse. This task is also harder for the discriminator which reduces the risk of the generator contending with a near-perfect discriminator. The critic in our method also helps the generator converge more rapidly by reducing the high-variance of the gradient updates in an extremely high action-space environment when operating at word-level in natural language.

3 MaskGAN

3.1 Notation

Let \((x_t, y_t)\) denote pairs of input and target tokens. Let \(<m>\) denote a masked token (indicates that the original token is replaced with a hidden token) and let \(\hat{x}_t\) denote the filled-in token. Finally, \(\hat{x}_t\) is a filled-in token passed to the discriminator which may be either real or fake.

3.2 Architecture

The task of imputing missing tokens requires that our MaskGAN architecture condition on information from both the past and the future. We choose to use a seq2seq Sutskever et al. (2014) architecture. Our generator consists of an encoding module and decoding module. For a discrete sequence \(x = (x_1, \cdots, x_T)\), a binary mask is generated (deterministically or stochastically) of the same length \(m = (m_1, \cdots, m_T)\) where each \(m_t \in \{0, 1\}\) selects which tokens will remain. The token at time \(t\), \(x_t\) is then replaced with a special mask token \(<m>\) if the mask is 0 and remains unchanged if the mask is 1.

The encoder reads in the masked sequence, which we denote as \(m(x)\), where the mask is applied element-wise. The encoder provides access to future context for the MaskGAN during decoding.

As in standard language-modeling, the decoder fills in the missing tokens auto-regressively, however, it is now conditioned on both the masked text \(m(x)\) as well as what it has filled-in up to that point. The generator decomposes the distribution over the sequence into an ordered conditional sequence

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

\[
G(x_t) \equiv P(\hat{x}_t | \hat{x}_1, \cdots, \hat{x}_{t-1}, m(x))
\] (1)
The discriminator has an identical architecture to the generator except that the output is a scalar probability at each time point, rather than a distribution over the vocabulary size. The discriminator is given the filled-in sequence from the generator, but importantly, it is given the original real context \( m(x) \). We give the discriminator the true context, otherwise, this algorithm has a critical failure mode. For instance, without this context, if the discriminator is given the filled-in sequence the \textit{director} director guided the series, it will fail to reliably identify the \textit{director} director bigram as fake text, despite this bigram potentially never appearing in the training corpus (aside from an errant typo). The reason is that it is ambiguous which of the two occurrences of director is fake; the \textit{associate} director guided the series or the director \textit{expertly} guided the series are both potentially valid sequences. Without the context of which words are real, the discriminator was found to assign equal probability to both words. The result, of course, is an inaccurate learning signal to the generator which will not be correctly penalized for producing these bigrams. To prevent this, our discriminator \( D_\phi \) computes the probability of each token \( \tilde{x}_t \) being real given the true context of the masked sequence \( m(x) \).

\[
D_\phi(\tilde{x}_t|\tilde{x}_{0:T}, m(x)) = P(\tilde{x}_t = x_t^{\text{real}}|\tilde{x}_{0:T}, m(x)) \tag{2}
\]

In our formulation, the logarithm of the discriminator estimates are regarded as the reward

\[
r_t \equiv \log D_\phi(\tilde{x}_t|\tilde{x}_{0:T}, m(x)) \tag{3}
\]

Our third network is the Critic network, which is implemented as an additional head off the discriminator. The critic estimates the value function, which is the discounted total return of the filled-in sequence

\[
R_t = \sum_{s=t}^{T} \gamma^s r_s,
\]

where \( \gamma \) is the discount factor.

### 3.3 Training

Our model is not fully-differentiable due to the sampling operations on the generator’s probability distribution for the next token. Therefore, to train the generator, we estimate the gradient with respect to its parameters \( \theta \) via policy gradients \cite{Sutton2000}. Reinforcement learning was first employed to GANs for language modeling in Yu et al. \cite{Yu2017}. Analogously, here the generator seeks to maximize the cumulative total reward \( R = \sum_{t=1}^{T} R_t \). We optimize the parameters of the generator (\( \theta \)) by performing gradient ascent on \( E[R] \). Using one of the REINFORCE family of algorithms, we can find an unbiased estimator of this as \( \nabla_\theta E_G[R_t] = R_t \nabla_\theta \log G_\theta(\hat{x}_t) \). The variance of this estimator may be reduced by using the learned value function as a baseline \( b_t = V^G(x_{1:t}) \). This results in the update equation

\[
\nabla_\theta E_G[R_t] = (R_t - b_t) \nabla_\theta \log G_\theta(\hat{x}_t) \tag{4}
\]

In the nomenclature of RL, the quantity \( R_t - b_t \) may be interpreted as an estimate of the advantage

\[
A(a_t, s_t) = Q(a_t, s_t) - V(s_t).
\]

Here, the action \( a_t \) is the token chosen by the generator \( a_t \equiv \tilde{x}_t \) and
the state \( s_t \) are the current tokens produced up to that point \( s_t \equiv \hat{x}_1, \cdots, \hat{x}_{t-1} \). This approach is an actor-critic architecture where \( G \) determines the policy and the baseline \( b_t \) is the critic (Sutton & Barto 1998; Degris et al. 2012).

For this task, we design rewards at each time step for a single sequence in order to aid with credit assignment (Li et al. 2017). As a result, a token generated at time-step \( t \) will influence the rewards received at that time step and subsequent time steps. Our gradient update for the discounted total return \( R = \sum_{t=1}^{T} R_t \) then becomes,

\[
\nabla \theta E[R] = \mathbb{E}_{\hat{x}_t \sim G} \left[ \sum_{t=1}^{T} (R_t - b_t) \nabla \theta \log(G_{\theta}(\hat{x}_t)) \right] 
\]

(5)

\[
\nabla \theta E[R] = \mathbb{E}_{\hat{x}_t \sim G} \left[ \sum_{t=1}^{T} \left( \sum_{s=t}^{T} \gamma^{s-t} r_s - b_t \right) \nabla \theta \log(G_{\theta}(\hat{x}_t)) \right] 
\]

(6)

Intuitively, this shows that the gradient to the generator associated with producing \( \hat{x}_t \) will depend on all the discounted subsequent rewards assigned by the discriminator. For a non-zero \( \lambda \) discount factor, the generator will be penalized for greedily selecting a token that earns a high reward at that time-step alone.

Finally, as in conventional GAN training, our discriminator will be updated according to the gradient

\[
\nabla \phi \frac{1}{m} \sum_{i=1}^{m} \left[ \log D(x^{(i)}) + \log(1 - D(G(z^{(i)}))) \right] 
\]

(7)

### 3.4 Method Details

We have pretraining stages, first we train a language model using standard maximum likelihood training. We then use the pretrained language model weights for the seq2seq encoder and decoder modules. With these language models, we now pretrain the seq2seq model on the in-filling task using maximum likelihood, in particular, the attention parameters as described in Luong et al. (2015). We select the model producing the lowest validation perplexity on the masked task via a hyperparameter sweep over 500 runs.

### 4 Training Details

Our model was trained with the Adam method for stochastic optimization (Kingma & Ba 2014) with the default Tensorflow exponential decay rates of \( \beta_1 = 0.99 \) and \( \beta_2 = 0.999 \). Our model uses 2-layers of 650 unit LSTMs for both the generator and discriminator, 650 dimensional word embeddings, variational dropout. We used Bayesian hyperparameter tuning to tune the variational dropout rate and learning rates for the generator, discriminator and critic. We perform 3 gradient descent steps on the discriminator for every step on the generator and critic.

We share the embedding and softmax weights of the generator as proposed in Bengio et al. (2003); Press & Wolf (2016); Inan et al. (2016). Furthermore, to improve convergence speed, we share the embeddings of the generator and the discriminator. Additionally, as noted in our architectural section, our critic shares all of the discriminator parameters with the exception of the separate output head to estimate the value. Both our generator and discriminator use variational dropout (Kingma et al. 2015).

### 5 Evaluation

Evaluation of generative models continues to be an open-ended research question. We seek heuristic metrics that we believe will be correlated with human-evaluation. BLEU score (Papineni et al. 2002) is used extensively in machine translation where one can compare the quality of candidate translations
from the reference. Motivated by this metric, we compute the number of unique \( n \)-grams produced by the generator that occur in the validation corpus for small \( n \). Then we compute the geometric average over these metrics to get a unified view of the performance of the generator.

From our maximum-likelihood trained benchmark, we were able to find GAN hyperparameter configurations that led to small decreases in validation perplexity on \( O(1) \)---point. However, we found that these models did not yield considerable improvements to the sample quality so we abandoned trying to reduce validation perplexity. One of the biggest advantages of GAN-trained NLP models, is that the generator can produce alternative, yet realistic language samples, but not be unfairly penalized by not producing with high likelihood the single correct sequence. As the generator explores `off-manifold` in the free-running mode, it may find alternative options that are valid, but do not maximize the probability of the underlying sequence. We therefore choose not to focus on architectures or hyperparameter configurations that led to small reductions in validation perplexity, but rather, searched for those that improved our heuristic evaluation metrics.

6 Experiments

6.1 The Penn Treebank (PTB)

The Penn Treebank dataset [Marcus et al. (1993)] has a vocabulary of 10,000 unique words. The training set contains 930,000 words, the validation set contains 74,000 words and the test set contains 82,000 words. For our experiments, we train on the training partition.

We first pretrain a language model with parameter dimensions common to MaskGAN in the [Zaremba et al. (2014)] code base to a validation perplexity of 78. After then loading the weights of from the language model into the MaskGAN generator we further pretrain with masking rate of 0.5 (half the text blanked) to a validation perplexity of 55.3. Finally, we then pretrain the discriminator on the samples produced from the current generator and real training text.

We present both conditional and unconditional samples generated on this data set. In the following section, MaskGAN refers to our GAN-trained variant and MaskMLE refers to our maximum-likelihood trained variant.

6.1.1 Conditional Samples

We produce samples conditioned on surrounding text. Underlined sections of text are portions here and throughout the text have been filled in via either the MaskGAN or MaskMLE algorithm.

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>the next day ‘s show &lt;eos&gt; interactive telephone technology has taken a new leap in &lt;unk&gt; and television programmers are</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskGAN</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>
</tr>
<tr>
<td></td>
<td>the next day ‘s show &lt;eos&gt; interactive telephone technology has long dominated the &lt;unk&gt; of the nation ‘s largest economic</td>
</tr>
<tr>
<td></td>
<td>the next day ‘s show &lt;eos&gt; interactive telephone technology has exercised a N N stake in the u.s. and france</td>
</tr>
<tr>
<td>MaskMLE</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></td>
<td>the next day ‘s show &lt;eos&gt; interactive telephone technology has been &lt;unk&gt; in a number of clients ‘ estimates mountain-bike</td>
</tr>
<tr>
<td></td>
<td>the next day ‘s show &lt;eos&gt; interactive telephone technology has instituted a week of &lt;unk&gt; by &lt;unk&gt; &lt;unk&gt; wis. auto</td>
</tr>
</tbody>
</table>

Table 1: Conditional samples from PTB for both MaskGAN and MaskMLE models.
6.1.2 **Language Model (Unconditional) Samples**

We may also run MaskGAN in an unconditional mode, where the entire context is blanked out, thus making it equivalent to a language model. We present a set of language model samples below in Table 2.

| MaskGAN | a <unk> basis despite the huge after-tax interest income <unk> from $ N million <eos> in west germany N N oct. N as the end of the year the resignations were approved <eos> the march N N <unk> was down the world ’s most corrupt organizations act as a multibillion-dollar <unk> atmosphere or the metropolitan zone historic array with their |

Table 2: Language model (unconditional) samples from PTB for both MaskGAN.

6.2 **IMDB Movie Dataset**

The IMDB dataset [Maas et al. 2011] consists of 100,000 movie reviews taken from IMDB. Each review may contain several sentences. The dataset is divided into 25,000 labeled training instances, 25,000 labeled test instances and 50,000 unlabeled training instances. The label indicates the sentiment of the review and may be either positive or negative. We use the first 40 words of each review in the training set to train our models, which leads to a dataset of 3 million words.

Identical to the training process in PTB, we pretrain a language model to a validation perplexity of 105.6. After then loading the weights of from the language model into the MaskGAN generator we further pretrain with masking rate of 0.5 (half the text blanked) to a validation perplexity of 87.1. Finally, we then pretrain the discriminator on the samples produced from the current generator and real training text.

6.2.1 **Conditional Samples**

<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></td>
<td>Pitch Black was a complete shock to me when I first saw it back in 1976 The promos were very well</td>
</tr>
<tr>
<td></td>
<td>Pitch Black was a complete shock to me when I first saw it back in the days when I was a</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>
<tr>
<td></td>
<td>Pitch Black was a complete shock to me when I first saw it back in 1951 It was funny All Interiors</td>
</tr>
<tr>
<td></td>
<td>Pitch Black was a complete shock to me when I first saw it back in the day and I was in</td>
</tr>
</tbody>
</table>

Table 3: Conditional samples from IMDB for both MaskGAN and MaskMLE models.

6.2.2 **Language Model (Unconditional) Samples**

As in the case with PTB, we generate IMDB samples unconditionally, equivalent to a language model. We present length-40 samples of both positive and negative sentiment in Table 4.

6.3 **Perplexity of Generated Samples**

As of this date, GAN training has not achieved state-of-the-art word level validation perplexity on the Penn Treebank dataset. Rather, the top performing models are still maximum-likelihood trained
Table 4: Language model (unconditional) samples from IMDB for both MaskGAN.

<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity of samples under a pretrained LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskMLE</td>
<td>273.1 ± 3.5</td>
</tr>
<tr>
<td>MaskMLE overfitting at 500k steps</td>
<td>155.3 ± 3.3</td>
</tr>
<tr>
<td>MaskGAN</td>
<td>108.3 ± 3.5</td>
</tr>
</tbody>
</table>

Table 5: The samples generated by the MaskGAN model consistently lower perplexity than samples generated by the MaskMLE model. The perplexity is calculated using a pre-trained language model that is equivalent to the decoder (in terms of architecture and size) used in the MaskMLE and MaskGAN models.

models, such as the recently well-regularized multilayer LSTM by Melis et al. (2017) followed by the architectures found via neural architecture search in Zoph & Le (2016). An extensive hyperparameter search with MaskGAN further supported that GAN training does not improve the validation perplexity results set via state-of-the-art models. However, we instead seek to understand the quality of the sample generation. As highlighted earlier, a fundamental problem of generating in free-running mode potentially leads to ‘off-manifold’ sequences which can result in poor sample quality for teacher-forced models. We seek to quantitatively evaluate this dynamic present only during sampling. This is commonly done with BLEU but as shown by Wu et al. (2016), BLEU is not necessarily correlated with sample quality. We believe the correlation may be even less in the in-filling task since there are many potential valid in-fillings and BLEU would penalize valid ones.

Instead, we calculate the perplexity of the generated samples by MaskGAN and MaskMLE by using the language model that was used to initialize MaskGAN and MaskMLE. Both MaskGAN and MaskMLE produce samples autoregressively (free-running mode), building upon the previously sampled tokens to produce the distribution over the next.

The MaskGAN model produces samples which are more likely under the initial model than the MaskMLE model. The MaskMLE model generates improbable sentences, as assessed by the initial language model, during inference as compounding sampling errors result in a recurrent hidden states that are never seen during teacher forcing. Conversely, the MaskGAN model operates in a free-running mode while training and this supports that it is more robust to these sampling perturbations.

6.4 Mode collapse

In contrast to image generation, mode collapse can be measured by directly calculating certain n-gram statistics. In this instance, we measure mode collapse by the percentage of unique n-grams in a set of 100K generated IMDB samples. We generate each sample (consisting of 40 words) conditioned only on the label. This results in almost 4M bi/tri/quad-grams.

In addition, all complete samples (taken as a sequence) are unique. We also observed during RL training an initial small drop in perplexity on the ground-truth validation set but then a steady increase in perplexity as training progressed. Despite this, sample quality remained relatively consistent. The final samples were generated from a model that had a perplexity on the ground-truth of 400. We
Table 6: As often witnessed in GAN literature, our discrete GAN is subject to some mode collapse.

hypothesise that mode dropping is occurring near the tail end of sequences since generated samples are unlikely to generate all the previous words correctly in order to properly model the distribution over words at the tail. [Theis et al., 2015] also shows how validation perplexity is not a good measurement of sample quality.

6.5 HUMAN EVALUATION

We evaluate the quality of the generated samples of the MaskGAN and the MaskMLE in a blind heads-up comparison using Amazon Mechanical Turk. We pay raters to compare the quality of two extracts along 3 axes (grammaticality, topicality and overall quality). They are asked if the first extract, second extract or neither is higher quality.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unique bigrams</th>
<th>Unique trigrams</th>
<th>Unique quadgrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskMLE</td>
<td>24.6</td>
<td>57.6</td>
<td>81.2</td>
</tr>
<tr>
<td>MaskMLE overfitting at 500k steps</td>
<td>13.7</td>
<td>40.3</td>
<td>66.6</td>
</tr>
<tr>
<td>MaskGAN</td>
<td>16.7</td>
<td>41.5</td>
<td>64.6</td>
</tr>
</tbody>
</table>

Table 7: A Mechanical Turk heads-up evaluation between MaskGAN and MaskMLE IMDB samples when asked which model’s samples raters preferred out of 100 samples. 279 ratings were obtained.

<table>
<thead>
<tr>
<th>Model</th>
<th>Grammaticality</th>
<th>Topicality</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskGAN</td>
<td>54.1</td>
<td>44.8</td>
<td>50.9</td>
</tr>
<tr>
<td>MaskMLE</td>
<td>25.8</td>
<td>29.4</td>
<td>26.2</td>
</tr>
</tbody>
</table>

Table 8: A Mechanical Turk heads-up evaluation between MaskGAN and MaskMLE PTB samples when asked which model’s samples raters preferred out of 100 samples. 94 ratings were obtained.

<table>
<thead>
<tr>
<th>Model</th>
<th>Grammaticality</th>
<th>Topicality</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskGAN</td>
<td>44.7</td>
<td>27.7</td>
<td>45.7</td>
</tr>
<tr>
<td>MaskMLE</td>
<td>37.2</td>
<td>36.3</td>
<td>40.4</td>
</tr>
</tbody>
</table>

The Mechanical Turk results show that MaskGAN generates superior human-looking samples to MaskMLE on the IMDB dataset. However, on the smaller PTB dataset, the results are closer. We also show results with SeqGAN (trained with the same network size and vocabulary size) as MaskGAN, which show similar results to MaskGAN on PTB but with a smaller margin. In additional tests of generated samples, we found that 20.6% of 4-grams in SeqGAN samples appeared in the training set compared to 10.8% of 4-grams in MaskGAN samples, showing that our model is less prone to mode collapse well generating high-quality samples.

7 DISCUSSION

We generally found training with contiguous masks to produce better samples. We believe this is simply an issue of error attribution, especially if the discriminator is not conditioned on the mask. For instance, suppose a random mask is sampled that has masked-out words with many real words in-between, an easy solution for the generator is to simply repeat the preceding or following word. It would then be very difficult for the discriminator to determine which word in the repeat was the real word and which was generated. Indeed, we observed this behavior in early samples.

We also found the use of attention was important for the in-filled words to be sufficiently conditioned on the input context. Without attention, the in-filling would fill in reasonable subsequences that became implausible in the context of the surrounding words.
Table 9: A Mechanical Turk heads-up evaluation between SeqGAN and MaskMLE PTB samples when asked which model’s samples raters preferred out of 100 samples. 96 ratings were obtained. We were unable to obtain sufficient responses on Mechanical Turk and thus we are unable to rule out the null hypothesis $p = 0.5$ with high significance.

In general we think the proposed contiguous in-filling task is a good approach to reduce mode collapse and help with training stability for textual GANs. We show that MaskGAN samples on a larger dataset (IMDB reviews) is significantly better than the corresponding tuned MaskMLE model as shown by human evaluation. We also show we can produce high-quality samples despite the MaskGAN model having much higher perplexity on the ground-truth test set.
REFERENCES


A.1 Mode Collapse

As widely witnessed in GAN-training, we also find a common failure of mode collapse across various n-gram levels. The mode collapse may not be as extreme to collapse at a 1-gram level (ddddddd, ...) as described by Gulrajani et al. (2017), but it may manifest as grammatical, albeit, inanely repetitive phrases, for example,

<table>
<thead>
<tr>
<th>MaskGAN</th>
<th>It is a very funny film that is very funny It s a very funny movie and it s charming It</th>
</tr>
</thead>
</table>

Table 10: Mode collapse failure mode occurring at a higher order n-gram.

Of course the discriminator may discern this as an out-of-distribution sample, however, in certain failure modes, we observed the generator to move between common modes frequently present in the text.

A.2 Matching Syntax for Future Context

We notice that the MaskGAN architecture often struggles to produce syntactically correct sequences when it must condition on future context. This is also a relatively challenging task for humans, because the filled in text must not only be contextual but also match syntactically at the boundary between the blank and where the text is present.

<table>
<thead>
<tr>
<th>MaskGAN</th>
<th>Cartoon is one of those films me when I first saw it back in 2000</th>
</tr>
</thead>
</table>

Table 11: Demonstration of syntax not being correct at the boundary of generated and present text.

As noted in this failure mode, the intersection between the filled in text and the present text is non grammatical.
A.3 LOSS OF GLOBAL CONTEXT

Similar to failure modes present in GAN image generation, the produced samples often can lose global coherence, despite being sensible locally.

| MaskGAN | 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 |

Table 12: Unconditional and conditional MaskGAN samples failing to maintain global coherence in the IMDB dataset.

A.4 n-GRAM METRICS MAY BE MISLEADING PROXIES

In the absence of a global scalar objective to optimize while training, we monitor various $n$-gram language statistics to assess performance. However, these only are crude proxies of the quality of the produced samples.

![Figure 2: Particular failure mode succeeding in the optimization of a 4-gram metric at the extreme expense of validation perplexity. The resulting samples are shown below.](image)

For instance, MaskGAN models that led to improvements of a particular $n$-gram metric at the extreme expense of validation perplexity as seen in Figure 2 could devolve to a generator of very low sample diversity. In Table 13, we produce several samples from this particular model which, despite the dramatically improved 4-gram metric, has lost diversity.

| MaskGAN | It is a great movie It's just a tragic story of a man who has been working on a home It's a great film that has a great premise but it's not funny It's just a silly film It's not the best movie I have seen in the series The story is simple and very clever but it |

Table 13: Demonstration of low sample diversity when the MaskGAN model over-optimizes $n$-gram metrics.

Capturing the complexities of natural language with these metrics alone is clearly insufficient.

A.5 SEQGAN SAMPLES

We modified SeqGAN to train and generate PTB samples using the same size architecture for the generator as in the MaskGAN generator.
are removed <eos> another takeover target lin ‘s directors attempted through october <unk> and british airways is allowed three funds poor utilization of the company ‘s facilities and its headquarters for virtually most of its previous age but remains small cineplex odeon corp. shares made fresh out of the group purchase one part of a revised class of <unk> british include its $ N billion in bonds and by far short interest <eos> on friday the company said it offered there are <unk> <unk> and <unk> about the <unk> seed <eos> they use pcs are <unk> and their performance <eos>