

# OPTAGAN: ENTROPY-BASED FINETUNING ON TEXT VAE-GAN

Paolo Tirota<sup>1</sup> and Stefano Lodi<sup>2</sup>

<sup>1</sup>Department of Statistics, University of Bologna, Italy

<sup>2</sup>Department of Computer Science, University of Bologna, Italy

## ABSTRACT

*Transfer learning through large pre-trained models has changed the landscape of current applications in natural language processing (NLP). Recently Optimus, a variational autoencoder (VAE) which combines two pre-trained models, BERT and GPT-2, has been released, and its combination with generative adversarial networks (GANs) has been shown to produce novel, yet very human-looking text. The Optimus and GANs combination avoids the troublesome application of GANs to the discrete domain of text, and prevents the exposure bias of standard maximum likelihood methods. We combine the training of GANs in the latent space, with the finetuning of the decoder of Optimus for single word generation. This approach lets us model both the high-level features of the sentences, and the low-level word-by-word generation. We finetune using reinforcement learning (RL) by exploiting the structure of GPT-2 and by adding entropy-based intrinsically motivated rewards to balance between quality and diversity. We benchmark the results of the VAE-GAN model, and show the improvements brought by our RL finetuning on three widely used datasets for text generation, with results that greatly surpass the current state-of-the-art for the quality of the generated texts.*

## KEYWORDS

*Text Generation, Variational Autoencoders, Generative Adversarial Networks, Reinforcement Learning.*

## 1. INTRODUCTION

Unsupervised text generation finds its use on a plethora of real-world application, ranging from machine translation [1], to summarization [2] and dialogue generation [3]. A general approach to modelling text sequences is to auto regressively generate the next token given the previous ones, and the most successful and widespread technique is to train a model using maximum likelihood estimation (MLE). This approach, however, is not without fault. At training time the model learns to generate a token given the ground truth, while at inference time it takes as input its own generated sequence of words. This dissimilarity leads to the so-called exposure bias [4], where the accumulation of errors during inference can produce poor outputs. Furthermore, the loss function of MLE is very strict. For each sequence, only the token accounted by the training sample is considered as correct [5], and the model learns precisely to mimic the given samples, often leading to quite dull and homogeneous outputs.

An alternative to MLE methods are generative adversarial networks (GANs) [6], where a generator learns to create outputs that can fool a discriminator into believing they are real. Thus, GANs do not have the strict loss function of MLE, and do not suffer from exposure bias, as they learn to sample during training. Nonetheless, the application of GANs to the text realm has been rather complicated. Due to the discreteness of text, the sampling of each token results in a non-

differentiable function, which does not allow to back propagate the loss of the discriminator. Countermeasures include the use of reinforcement learning (RL) [7][8][9][10], the use of the Gumbel-Softmax relaxation [11][12], or to avoid the discrete space altogether and work with continuous embeddings using autoencoders [13][14][15]. However, methods which utilize RL often rely on MLE pre-training, and usually do not improve over them [16]. Instead, for both the approaches using the Gumbel-Softmax distribution, and even more so for autoencoders, the discriminator considers a continuous representation of text, so it is not able to judge effectively the single word-by-word generation.

In the past few years, natural language processing (NLP) applications have found huge improvements with the introduction of the attention mechanism and the transformer architecture, with notable examples of BERT, GPT-2 and GPT-3 among others [17][18][19][20]. These kind of language models are large deep neural networks that are able to understand the dependencies between words thanks to attention and are trained over huge amounts of unannotated data. As such, pre-trained language models provide better language understanding over recurrent neural networks, can be very easily finetuned on a downstream task, including text generation, and reached state-of-the-art results in many areas. Recently Optimus, a text variational autoencoder (VAE), that is an autoencoder which maps sentences to a meaningful latent space, has been proposed [21]. It combines both BERT and GPT-2, as encoder and decoder respectively, and can be employed both as a generative model, and as a tool for language understanding tasks.

In this work, we aim to benchmark the results obtained from combining Optimus and GANs, similarly as indicated in the original paper. In doing so, we also investigate the GAN structure and compare the adaptive update strategy presented in [22] with the standard update strategy of GANs. Furthermore, we combine the training in the continuous space, with the finetuning of the decoder of Optimus in the discrete text space, in a similar fashion as done in ConCriteGAN [23]. However, differently from most approaches which use RL, we do not use REINFORCE, but add an additional value head to GPT-2, which outputs the intermediate rewards [24]. Moreover, we modify the reward function by considering the entropy of the model when generating tokens, and favour diversity in the output by adding an intrinsic reward.

Thus, our model OptAGAN<sup>1</sup> is able to model both the higher level sentence structure, and has more control over single word generation, in a way that favours both quality and diversity for the generated sentences. We measure such criteria using standard automatic metrics: BLEU for quality, Backwards-BLEU for diversity, and Fréchet distance, on which we also present further analysis. We consider the image caption dataset COCO, the Stanford Natural Language Inference (SNLI) dataset, and the EMNLP News 2017 dataset for unconditional text generation, and also provide results for the conditional review dataset YELP.

Results show that the base VAE-GAN model already improves over other GAN methods, especially with regards to quality. OptAGAN, further improves over these results, and manages to handle the quality-diversity trade-off very well. Moreover, we show a further experiment that helps understanding the strengths and weaknesses of our finetuning approach.

## 2. BACKGROUND

In this section we introduce the mathematical notation and briefly describe the main theoretical tools which are used in OptAGAN. We also present an overview of the other methods of text generation.

---

<sup>1</sup>Opt(imus) A(ugmented) GAN – Implementation can be found at <https://github.com/Egojr/optagan>

## 2.1. Variational Autoencoders

VAEs are generative models formed by two independent models, an encoder  $q_\phi$  and a decoder  $p_\theta$ . The encoder is tasked with mapping the input  $x$  to a latent space  $z$  that allows for interpolation. The decoder maps from  $z \rightarrow \tilde{x}$ , providing an approximation of the original input. Thanks to the introduction of a local variation from sampling the encoder output, it is possible to induce a smooth latent representation of the inputs, which differs from the rigid space of autoencoders.

**Optimus** Optimus combines the autoregressive nature of the GPT-2 text generation with the latent produced by the encoder, such that text generation is done as:

$$p_\theta(x|z) = \prod_{t=1}^n p_\theta(x_t|x_1, \dots, x_{t-1}, z) = \prod_{t=1}^n p_\theta(x_t|x_{<t}, z), \quad (1)$$

where the probability of each token is estimated conditionally on the latent embedding and the previous tokens. The latent vector  $z$ , which comes from the output of the BERT encoder, controls the high-level characteristics of the sentence, such as length, tense, style and topic, and allows for the guided generation of text.

## 2.2. Generative Adversarial Networks

GANs are also generative models formed by two models: a generator and a discriminator. Differently from VAEs, the generator  $G$  samples from a random variable to produce output that can fool the discriminator  $D$  into believing they are real, while the discriminator is constantly learning to distinguish between the real and generated data. The objective of the two models can be formulated as:

$$\min_G \max_D (V(D, G)) = E_{x \sim p_D} [\log(D(x))] + E_{\epsilon \sim p_\epsilon} [\log(1 - D(G(\epsilon)))], \quad (2)$$

where  $p_D$  and  $p_\epsilon$  are the distribution of the data and of the input noise of the generator, respectively. When combining GANs with autoregressive text generation, the operation of sampling the next tokens is non-differentiable, so the application of GANs relies on either the use of policy gradient algorithms, or the use of continuous approximations, such as the combination of GANs with autoencoders or VAEs.

## 2.3. Reinforcement Learning

Approaches that use policy gradient algorithms to allow the training of GANs consider the generator as the policy  $\pi_\theta$  to train, the sampling of the token as an action  $A$  from a state  $S$ , and the output of the discriminator as the reward  $R$ . As the discriminator only calculates the reward over finished sentences, the intermediate rewards are obtained through the REINFORCE algorithm and Monte-Carlo rollout. Optimization of the parameters is performed through gradient ascent:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta_t), \quad (3)$$

$$\nabla_\theta J(\theta_t) \propto E_\pi [G_t \nabla_\theta \ln \pi(A_t | S_t, \theta)], \quad (4)$$

where the gradient of the REINFORCE objective,  $\nabla_\theta J(\theta_t)$  is proportional to the discounted returns  $G_t = \sum_{j=0}^{\infty} \gamma^j R_{t+j}$ , so that higher return actions are favoured, and is inversely proportional to the probability of being selected, so that higher probability actions are not at an advantage compared

to low probability ones. Equation 4 also provides a value that can be sampled at each time step and only depends on the policy  $\pi$ .

## 2.4. Related Work

Many works have dealt with the training of GANs in the discrete realm, starting with SeqGAN [7], LeakGAN [8] and RankGAN [9], where all of them share a similar structure, mostly differing in the form of the discriminator, and require MLE pre-training followed by adversarial training with REINFORCE. ScratchGAN [10] is the first model to show that MLE pre-training can be avoided by carefully combining existing techniques. Other works that train GANs using continuous relaxations include ARAE [13] and LATEX-TGAN [15], which use autoencoders to learn a continuous latent representation. Models based on the Gumbel-Softmax distribution are RelGAN [12] and GSGAN [11]. On the comparison of these methods and the evaluation metrics, [16][25] have shown the inadequacy of current GANs when compared to MLE and the need for metrics that can better measure the quality and diversity of the models.

On the topic of exploration of text GAN models and RL are ColdGANs [26], which delve deeper into the effects of temperature for the text generation. An approach similar to ours, which involves the use of large pre-trained models and RL is TextGAIL [27], where both GPT-2 as the generator, and RoBERTa as the discriminator are used for the task of text generation.

Regarding text VAEs, Optimus [21] is the first large pre-trained model of such kind, whereas previously researchers had tried developing VAEs using either recurrent neural networks [28] or semi-amortized inference [29].

## 3. OPTAGAN

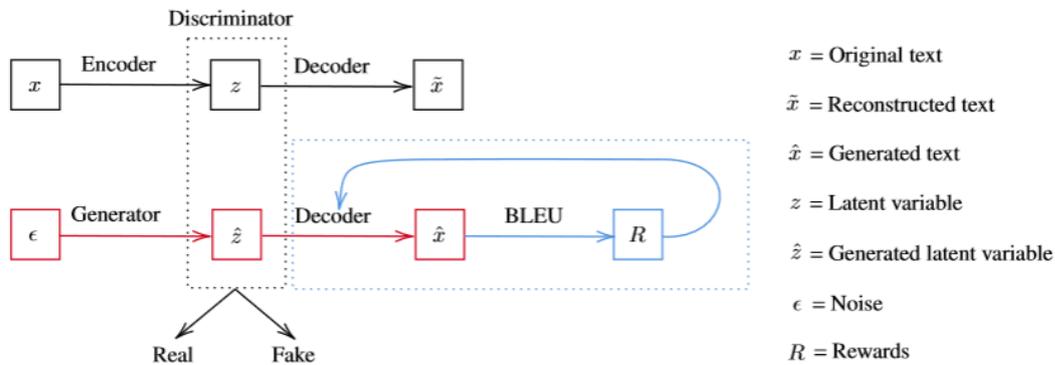


Figure 1. Structure of our proposed model OptAGAN. In red is the main process to generate new text, while in black we show the parts that are present in the training of the model. Finally, in blue is the last bit of the RL finetuning process.

The architecture that we present in this work is composed of three main processes, as can be seen from Figure 1. Each process is independent of the others, so each part is trained sequentially.

- In order to fully utilize the strengths of Optimus, we finetune both the encoder and the decoder on the target dataset. The end results are a more separated and distinct latent space for each sentence, and a decoder which better reconstructs the original sentences.
- Next, we train the GAN model composed of the generator and the discriminator. In the case of conditional generation, we also add a classifier, whose loss is then passed to the

generator. Both the generator and the discriminator only consider the continuous latent embeddings, so they are much lighter and faster to train compared to other text GANs.

- Finally, we finetune the decoder on discrete text using a value head, which estimates the reward of each single token in a sentence for the generated sequences. The estimated rewards are also augmented by considering the model entropy of each generated token. The gradient is then passed to the decoder through simple policy gradient.

The structure of both the generator and the discriminator is a simple feed-forward neural network. Current literature does not give clear answers about which loss function specification for GANs is best for continuous data such as text latent embeddings. Our experiments show that Wasserstein GANs with gradient penalty (WGAN-GP) perform the best over sliced Wasserstein distances, which produce very homogeneous outputs. Thus, the loss function to optimize for the generator and discriminator is:

$$\min_G \max_D E_{x \sim p_D} [D(x)] - E_{\epsilon \sim p(\epsilon)} [D(G(\epsilon))] + \lambda E_{\hat{x} \sim p(\hat{x})} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2], \quad (5)$$

### 3.1. Update Strategy

When training GANs, the common update strategy is to have  $k$ , usually in the range [5][10], update steps for the discriminator for each step of the generator, as it has been shown to give stable training for GANs. We indeed also use this update method, but we also experiment with an adaptive update strategy proposed in [22], where the choice to update the discriminator or the generator is given by a comparison of the loss change ratio of the two networks.

$$r_G = \frac{|(L_G^c - L_G^p)|}{L_G^p + c}, \quad (6) r_D = \frac{|(L_D^c - L_D^p)|}{L_D^p + c}, \quad (7)$$

where the relative change between the current loss  $L^c$  and previous loss  $L^p$  for both networks is used in determining which one gets updated. We also add an arbitrarily small constant  $c$  in case the losses are too close to 0.

A weight  $\lambda$  can be also introduced, so that if  $r_D > \lambda r_G$  the discriminator is updated, and vice versa. Contrarily to the original paper, which suggests a value  $\lambda \square \square \geq 1$ , we notice that at the beginning of training there is a stark imbalance in the number of updates between the networks resulting in slower convergence. To better balance the training, we end up using a value of  $\lambda < 1$ , which converges to 1 with each passing epoch of training.

### 3.2. Value Head

Due to the high computational costs of implementing a text discriminator with a vocabulary of size equal to GPT-2, we rely on an external value head, whose scalar output for each token corresponds to the intermediate reward. Let the hidden states of the decoder  $\rho$  and the value head  $v$ , the rewards calculated from an external metric  $R^{ex}$  and each state  $S^t$  are:

$$R_t = v(S_t | \rho), \quad \text{with } v(S_t | \rho) \text{ minimizing } \left| \sum_{t=0}^T (v(S_t | \rho)) - R^{ex} \right|, \quad (8)$$

The external head takes as input the frozen hidden states of GPT-2. Freezing the hidden parameters is necessary because we are not modelling the parameters of the VAE model during RL. The loss is then calculated according to a MAE objective and passed back to the value head.

This process estimates how much each token contributes to the reward, and comparatively to a text discriminator is faster to train and showed better results.

### 3.3. Entropy-Based Rewards

Our external rewards are calculated based on the quality metric BLEU, which we briefly describe in Section 4.1. Under many considered reward specifications, which included the addition of diversity metrics, maximum entropy RL, changes of temperature, or a combination of these, the increase in quality is counterbalanced with a drop in the diversity of the generated sentences. One approach that managed to balance the quality-diversity trade-off was the addition of an intrinsically motivated penalty based on the confidence of the model when generating the token, calculated by the entropy. If again we consider the hidden states  $\rho$ , the last layer calculating the logits as our policy  $\pi$  with parameters  $\Theta$  and the entropy as  $H(\pi(\cdot|\rho, S_t))$ , we calculate the intrinsic rewards and the performance objective as:

$$R_t^{in} = \log(\text{clamp}(H(\pi(\cdot|\rho, S_t)), 0.2, 1)), \quad (9)$$

$$\nabla_{\theta} J(\theta_t) = E_{\pi} \left[ \left( \sum_{k=0}^t (\gamma^k R_k) + R_t^{in} \right) \nabla_{\theta} \ln \pi(A_t | \rho, S_t, \theta) \right], \quad (10)$$

This specification favours high-reward actions with high entropy, while low-entropy actions have to have a high enough reward to be able to keep their high probability, resulting in a more diverse generation. As a rule of thumb, we found out that penalties should be lower than the maximum overall reward.

## 4. EXPERIMENTAL SETTINGS

In this section, we introduce the automatic metrics and the datasets used for evaluation. For comparison we consider a MLE model, SeqGAN, RankGAN, as implemented by the benchmarking platform Texus[30], and ScratchGAN. We also provide further details on our RL finetuning and present issues with the current evaluation metrics.

### 4.1. Evaluation Metrics

BLEU is a metric that measures the overlapping n-grams between a hypothesis text and all the reference texts. The final score is calculated as the average of the scores over all hypothesis sentences. Studies have shown [16][25] that BLEU can only detect small syntax problems, resulting in poor correlation with human evaluations, however it still remains the standard when evaluating the quality of generated texts. To measure diversity we utilize Backwards-BLEU (BBLEU) where the generated texts are the reference and the test set becomes the hypothesis, giving a measure of how much the test set is represented.

Additionally, we consider the Fréchet Distance with the InferSent embedding model (FID). It has been shown that FID responds better than BLEU at identifying mode collapse and changes in words usage. However, we show that it can be biased due to its distributional assumptions, mainly for differences in sentence length distribution. Nonetheless, it can be useful in identifying very homogeneous outputs, especially in conjunction with BLEU and BBLEU scores.

## 4.2. Datasets

We consider three of most widely used datasets for unconditional text generation: image COCO [30], Stanford Natural Language Inference (SNLI) [31] and the EMNLP News 2017 dataset<sup>2</sup>. Moreover, we consider the YELP review dataset for conditional text generation [33].

Table 1. Average length of the train set and number of sentences for each of the datasets used for evaluation.

	<b>COCO</b>	<b>SNLI</b>	<b>EMNLP</b>	<b>YELP</b>
<b>Conditional</b>	X	X	X	✓
<b>Average sentence length</b>	11.3	9.7	28.8	96.4
<b>Size of train set</b>	10k	100k	270k	100k
<b>Size of dev set</b>	10k	10k	10k	10k
<b>Size of test set</b>	10k	10k	10k	10k

Each of the datasets presents different challenges when training: COCO and SNLI are a small and medium-sized dataset with short sentences, respectively. EMNLP is a large dataset with longer sentences. Lastly, the YELP dataset is a conditional, medium-sized dataset with very long sentences. The preprocessing on the datasets is minimal and only on the YELP dataset.

## 4.3. Experimental Results

Tables 2, 3 and 4 show the results for the quality and diversity of the models. The changes between standard and adaptive updates mostly favour the adaptive one, with larger gains in diversity, with the exception of the SNLI dataset. Additional considerations about the two updates can be found in Appendix B. Therefore, we only apply the RL finetuning on the adaptively trained model to obtain the OptAGAN results. Our approach shows improvements under all metrics, albeit small for the COCO and SNLI datasets. In comparison with the other GAN models, OptAGAN boasts the highest quality, and average, or higher diversity. In comparison with the MLE model, the quality-diversity trade-off favours our model for quality, and the MLE approach for diversity. Notably, for the EMNLP dataset, the curiosity-driven finetuning allows OptAGAN to surpass all models for both BLEU and BBLEU.

We believe that the difference in the magnitude of change between the EMNLP task and the COCO and SNLI ones is due the starting quality of the model. In fact, as the RL finetuning slightly prioritizes quality, so increases in BLEU score, over diversity, the actual changes on the word-by-word generation are very few for those two datasets.

Compared to the other methods, ours is also better able to reproduce longer sequences of words, as the growing differences between 2,3 and 4 n-grams metrics show. Regarding the FID scores, they mostly show the same behaviour as BBLEU. However, we show in section 4.5 that the FID is biased for the sentence length distribution, that, in contrast with other methods such as ScratchGAN, we do not model.

---

<sup>2</sup><http://www.statmt.org/wmt17/>

Table 2. EMNLP results of the automatic metrics for OptAGAN, the two base VAE-GAN models with standard and adaptive updates and the other models implemented for comparison.

Metrics	MLE	SeqGAN	RankGAN	ScratchGAN	Standard	Adaptive	OptAGAN
<b>BLUE-2</b>	0.829	0.796	0.764	0.835	0.825	0.816	<b>0.860</b>
<b>BLUE-3</b>	0.548	0.471	0.399	0.556	0.554	0.544	<b>0.605</b>
<b>BLEU-4</b>	0.304	0.228	0.159	0.313	0.285	0.284	<b>0.356</b>
<b>BBLEU-2</b>	0.840	0.762	0.728	0.824	0.765	0.805	<b>0.841</b>
<b>BBLEU-3</b>	0.563	0.563	0.383	0.545	0.488	0.526	<b>0.586</b>
<b>BBLEU-4</b>	0.317	0.317	0.221	0.303	0.261	0.278	<b>0.350</b>
<b>FID</b>	0.926	1.934	3.509	<b>0.466</b>	1.153	0.784	0.674

Table 3. SNLI results of the automatic metrics for OptAGAN, the two base VAE-GAN models with standard and adaptive updates and the other models implemented for comparison.

Metrics	MLE	SeqGAN	RankGAN	ScratchGAN	Standard	Adaptive	OptAGAN
<b>BLUE-2</b>	0.841	0.838	0.784	0.795	0.867	0.888	<b>0.889</b>
<b>BLUE-3</b>	0.635	0.599	0.514	0.564	0.693	0.726	<b>0.727</b>
<b>BLEU-4</b>	0.428	0.380	0.309	0.363	0.484	0.524	<b>0.525</b>
<b>BBLEU-2</b>	<b>0.843</b>	0.768	0.771	0.800	0.786	0.764	0.764
<b>BBLEU-3</b>	<b>0.639</b>	0.546	0.523	0.564	0.570	0.547	0.548
<b>BBLEU-4</b>	<b>0.433</b>	0.347	0.347	0.362	0.374	0.361	0.362
<b>FID</b>	<b>0.376</b>	0.919	1.486	0.539	1.424	1.764	1.765

Table 4. COCO results of the automatic metrics for OptAGAN, the two base VAE-GAN models with standard and adaptive updates and the other models implemented for comparison.

Metrics	MLE	SeqGAN	RankGAN	ScratchGAN	Standard	Adaptive	OptAGAN
<b>BLUE-2</b>	0.854	0.866	0.834	0.862	0.917	0.919	<b>0.920</b>
<b>BLUE-3</b>	0.664	0.667	0.641	0.678	0.792	0.792	<b>0.794</b>
<b>BLEU-4</b>	0.459	0.452	0.423	0.478	<b>0.627</b>	0.617	0.620
<b>BBLEU-2</b>	<b>0.844</b>	0.813	0.775	0.791	0.755	0.775	0.778
<b>BBLEU-3</b>	<b>0.656</b>	0.616	0.553	0.573	0.550	0.579	0.584
<b>BBLEU-4</b>	<b>0.455</b>	0.415	0.344	0.377	0.368	0.396	0.399
<b>FID</b>	<b>0.588</b>	0.756	2.347	1.001	2.297	1.908	1.903

From the computational cost point of view, the full training of our model took at most 24 hours using a single Tesla V100 GPU. Additional details about the training can be found in Appendix A.

#### 4.4. Conditional Generation

For the conditional generation task, we follow the same procedure as the unconditional one, with the only exception of the addition of a classifier network to better model the GAN generation depending on the label. The results of table 5 are very similar to the ones of the COCO and SNLI, where the entropy regularized finetuning performs slightly better than the base adaptive VAE-GAN model, with small gains in diversity and quality. We present some examples of the generated sentences in Appendix C.

Table 5. Yelp results of the automatic metrics models and the entropy regularized OptAGAN.

Metrics	Standard	Adaptive	OptAGAN
<b>BLUE-2</b>	0.880	0.886	<b>0.887</b>
<b>BLUE-3</b>	0.675	0.683	<b>0.685</b>
<b>BLEU-4</b>	0.448	0.456	<b>0.458</b>
<b>BBLEU-2</b>	<b>0.860</b>	0.854	0.854
<b>BBLEU-3</b>	<b>0.666</b>	0.660	0.661
<b>BBLEU-4</b>	<b>0.453</b>	0.449	0.451
<b>FID</b>	2.598	<b>2.539</b>	2.562

#### 4.5. Analysis on FID Space

We discuss problems with the FID metric by showing an heatmap of the two-dimensional principal component analysis (PCA) representation and the length distribution of the sentences for the SNLI dataset. Previous works [10] already investigated the dependency of the FID scores on the length distribution, which can overshadow other problems with the generated samples.

We further reinforce those analysis, as a prime example of this issue can be seen in Figure 2, where the FID scores get progressively higher the more the length distribution of the generated sentences is close to the one of the test data, while ignoring the fact that the actual distribution may not match the test one.

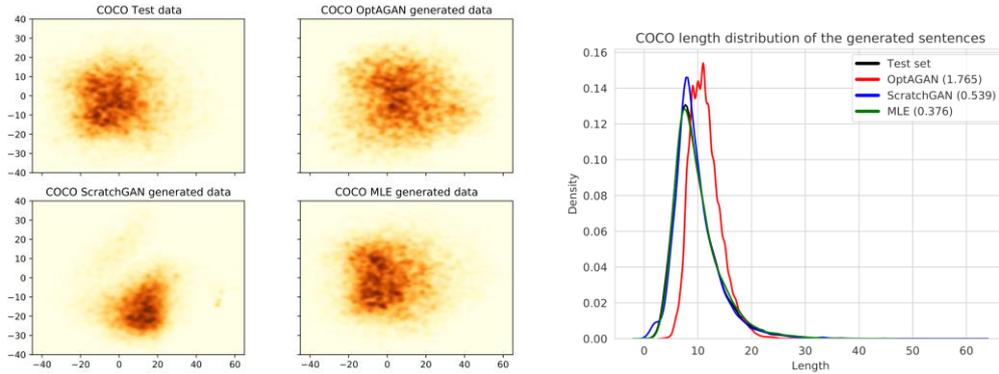


Figure 2. On the left, the PCA representation of the FID space of the SNLI test set (top left), OptAGAN (top right), ScratchGAN (bottom left) and MLE (bottom right) dataset. On the right, the length distribution of the four datasets and the FID scores for the generated data.

In fact, the score for the ScratchGAN model is much lower than the one for OptAGAN, although the distribution of the sentences of ScratchGAN in the space misses most of the distribution of the test sentences, as can be seen from the PCA representation. Although only 10-15% of the overall variability, over the InferSent embedding dimensions, is explained by PCA, there is a huge mismatch that is not addressed by the use of the Fréchet distance, which favours homogeneous length distributions over correct representation of the space.

#### 4.6. Experiment on RL Finetuning

In order to fully gauge the strengths and weaknesses of our entropy penalty approach we set up an experiment where we train a GAN model using the pre-trained Optimus encoder and decoder that are not finetuned on the dataset, so we can finetune a lower quality model. We compare this model with two curiosity-driven regularized model for 1000 and 5000 epochs, respectively. We also use a higher learning rate as we are not interested in preserving the structure of the original sentences and present the results in Table 6.

The starting model achieves much worse results than the fully optimized OptAGAN, especially with regards to the BLEU score. After 1000 epochs of RL finetuning, the model improves over both quality and diversity. However, when finetuning for longer, the model cannot balance anymore between exploration and exploitation.

Table 6. Yelp results of the automatic metrics models and the entropy regularized OptAGAN.

Metrics	Base VAE-GAN	RL 1000 Epochs	RL 5000 Epochs
<b>BLUE-2</b>	0.641	0.707	<b>0.829</b>
<b>BLUE-3</b>	0.370	0.444	<b>0.624</b>
<b>BLUE-4</b>	0.198	0.243	<b>0.402</b>
<b>BBLEU-2</b>	<b>0.752</b>	<b>0.752</b>	0.717
<b>BBLEU-3</b>	0.478	<b>0.497</b>	0.479
<b>BBLEU-4</b>	0.266	<b>0.286</b>	0.283
<b>FID</b>	2.441	<b>2.369</b>	2.892

We believe this behaviour is because of the bias that the model has in finding high reward tokens. The RL finetuning does not evaluate all the tokens in the vocabulary, so there might be multiple good scoring tokens which are never considered during the finetuning. This means that our approach is limited by the quality of the original model.

A countermeasure that might prevent this kind of behaviour from happening could be to increase the penalty for models with higher average entropy, and tuning its value depending on the use case, to further encourage heterogeneous generation.

## 5. CONCLUSIONS AND DISCUSSION

In this work, we benchmark the combination of Optimus and GANs for a text VAE-GAN model, with results that already surpass current methods for the quality of generated texts. We further improve this baseline using entropy-based curiosity-driven rewards to improve both the quality and the diversity of the model. This novel approach could benefit many models utilizing RL for text generation, and supplementary research could be done into exploring advantage policy gradient, or proximal policy optimization with intrinsic rewards. This specification of the reward also allows researchers to prioritize quality, diversity, or to balance between both.

Due to our limited computational resources, we utilize smaller batch sizes than we would otherwise have preferred, as larger batch sizes could help in reducing the high variance gradients of these approaches. Moreover, further research on the automatic metrics could be beneficial not only for evaluation, but also for better reward signals to improve the speed and quality of the finetuning.

## REFERENCES

- [1] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, NishantPatil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. *Google's neural machine translation system: Bridging the gap between human and machine translation*.
- [2] Mehdi Allahyari, Seyed Amin Pouriye, Mehdi Assefi, SaeidSafaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys J. Kochut. *Text summarization techniques: A brief survey*.
- [3] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. *Deep reinforcement learning for dialogue generation*.
- [4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. *Scheduled sampling for sequence prediction with recurrent neural networks*.
- [5] Ofir Press, Amir Bar, Ben Bogin, Jonathan Berant, and Lior Wolf. *Language generation with recurrent generative adversarial networks without pre-training*.
- [6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, SherjilOzair, Aaron Courville, and Yoshua Bengio. *Generative adversarial networks*, 2014.
- [7] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. *Seqgan: Sequence generative adversarial nets with policy gradient*.
- [8] Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. *Adversarial ranking for language generation*.
- [9] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. *Long text generation via adversarial training with leaked information*, 2017.
- [10] Cyprien de Masson d'Autume, Mihaela Rosca, Jack Rae, and Shakir Mohamed. *Training language gans from scratch*, 2020.
- [11] Matt J. Kusner and José Miguel Hernández-Lobato. *Gans for sequences of discrete elements with the gumbel-softmax distribution*, 2016.
- [12] WeiliNie, Nina Narodytska, and Ankit Patel. *RelGAN: Relational generative adversarial networks for text generation*. In International Conference on Learning Representations, 2019.
- [13] Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. *Adversarially regularized autoencoders for generating discrete structures*.
- [14] David Donahue and Anna Rumshisky. *Adversarial text generation without reinforcement learning*.
- [15] Md. Akmal Haidar, Mehdi Rezagholizadeh, Alan Do-Omri, and Ahmad Rashid. *Latent code and text-based generative adversarial networks for soft-text generation*.
- [16] Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. *Language gans falling short*, 2020.
- [17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention is all you need*. CoRR, abs/1706.03762, 2017.
- [18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: pre-training of deep bidirectional transformers for language understanding*.
- [19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. *Language models are unsupervised multitask learners*. 2019.
- [20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. *Language models are few-shot learners*, 2020.
- [21] Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. *Optimus: Organizing sentences via pre-trained modeling of a latent space*, 2020.
- [22] Xu Ouyang and GadyAgam. *Accelerated wgan update strategy with loss change rate balancing*, 2020.
- [23] Yanghoon Kim, Seungpil Won, Seunghyun Yoon, and Kyomin Jung. *Collaborative training of gans in continuous and discrete spaces for text generation*, 2020.
- [24] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. *Fine-tuning language models from human preferences*.

- [25] Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. *On accurate evaluation of gans for language generation*, 2019.
- [26] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. *Coldgans: Taming language gans with cautious sampling strategies*, 2020.
- [27] Qingyang Wu, Lei Li, and Zhou Yu. *Textgail: Generative adversarial imitation learning for text generation*, 2021.
- [28] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, RafalJózefowicz, and SamyBengio. *Generating sentences from a continuous space*.
- [29] Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, and Alexander M. Rush. *Semi-amortized variational autoencoders*, 2018.
- [30] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. *Texygen: A benchmarking platform for text generation models*.
- [31] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. *Microsoft COCO captions: Data collection and evaluation server*.
- [32] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. *A large annotated corpus for learning natural language inference*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
- [33] Nabihha Asghar. *Yelp dataset challenge: Review rating prediction*.

## APPENDIX

### A. TRAINING DETAILS

For the finetuning of Optimus we follow the original work and train for one epoch with the hyperparameters that give the best reconstruction quality, namely:

- Pre-trained model epoch = 508523
- Training epochs<sup>3</sup> = 1
- Learning rate =  $5 \cdot 10^{-5}$
- Batch size = 5
- Latent size = 768
- $\beta = 0$
- Annealing ratio = 0.5
- Ratio increase = 0.25

We follow a similar approach for the training of the GAN part of the model, where we use standard hyperparameters for the training of WGAN-GP. The best performing epoch of the GAN according to the sum of the BLEU and BBLEU partial scores with 500 texts for both reference and hypothesis is saved.

- Training epochs = 50
- Learning rate =  $10^{-4}$
- Batch size = 256
- Latent size = 768
- Maximum sequence length = 100
- Number of blocks of generator and discriminator = 10
- Gradient penalty  $\lambda = 10$

---

<sup>3</sup>Due to the size of the COCO dataset, it is the only one where we finetune for 5 epochs

Finally, these are the details for the entropy regularized finetuning. Empirical results showed next to no difference for the BLEU n-gram choice, so we choose 1-gram due to slightly faster computational times, and it also translates into a clearer understanding of the intermediate values. Moreover, we use a small learning rate, in order to keep the same structure as the original sentences.

- BLEU reward n-grams = 1
- Finetuning epochs = 1000
- Learning rate =  $10^{-6}$
- Batch size = 32
- Epochs value head pre-training = 200
- Learning rate value head pre-training =  $10^{-4}$

## B. ADAPTIVE STRATEGY DETAILS

Figure 3 shows the validation BLEU and BBLEU for the COCO and YELP datasets. We show the results over these two datasets due to the stark difference between them.

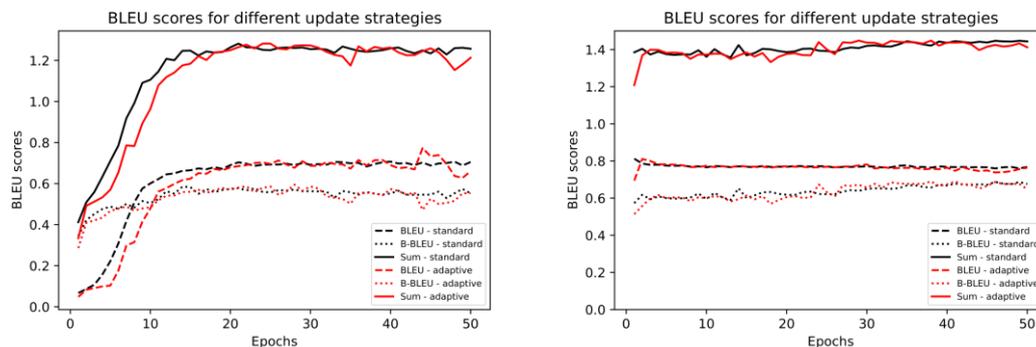


Figure 3. Differences in BLEU scores during training for the standard and adaptive update in GANs, evaluated on the COCO dev set on the left and on the YELP dev set on the right.

It is evident that the adaptive strategy is slightly slower to converge. Since the YELP dataset is 10 times larger than COCO, in both cases after about 200,000 samples the two models reach the same quality. The remainder of the training appears to be very stable for the standard update, with little to no changes in both the scores. Meanwhile, the adaptive updates look slightly more volatile, with changes that impact the two scores both positively and negatively, usually following the quality-diversity trade-off principle. However, it also means that the adaptive updates is more likely to find higher scores than the standard one, as happens for all the datasets analysed in this work.

## C. GENERATED SAMPLES

Table 7. Examples of the generated unconditional sentences from OptAGAN trained on COCO, SNLI and EMNLP dataset.

Dataset	Sentences
COCO	a man posing with a bike inside of a forest . a man sitting on a swinging chair pulled by some purple ducks . a woman holding a bird over some flowers on the beach .

SNLI	<p>the man is sweating because they are blue .</p> <p>the man is getting more thank with the dog .</p> <p>a man in white clothes stands next to a marketplace where he can store plastic .</p>
EMNLP	<p>Republican presidential nominee , Donald Trump , has said that 20 to 30 years might be the way to try and narrow it out .</p> <p>Whether or not agenda minutes can deliver , it would , therefore , encourage the majority of Scottish MPs to think about that .</p> <p>Earlier in the day , Bell travelled to Sydney ' s Supreme Court and was effectively blocking a vote of no - one that would produce the album .</p>

Table 8. Examples of the generated sentences from OptAGAN trained on the YELP dataset, conditioned on the label.

Stars	Generated review
1	<p>i was disappointed with this company . for the 2 visits to this location on the \_UNK highway i paid a ridiculous amount to visit . got stuck there , and ignored and messed up on my money . . the wait staff must know the difference between sink of vinegar and fresh vinegar !haha . i have seen orange juice \_UNK when they used to serve it but now the juice i am getting from juice bar in sanfrancisco had no consistency at all .i asked the waitress if they could rotate me out of soda in the microwave in exchange for a new glass ... hmmm , if that's your way of saying bad customer service you are talking a rip off . wait staff can do anything .</p>
2	<p>i have been to some great buffets , but this was mediocre at best . my husband ordered a turkey sandwich and it was just like anyone else's sandwich . for the service , there wasn't much seating . the food and the \_UNK were stale , awful bread . something to do if you go to vegas for dinner and want to have some classic awesomeness ... maybe try a dim sum instead .</p>
3	<p>this place is pretty good .i like the burgers , the selection is pretty good and they have a ginormous amount of steak . being a vegetarian ,i took one bite of everything i ordered and went back again . for the friday afternoon rush - they brought me the french fries instead of the turkey , a mousse and mgr .</p>
4	<p>loved this place .i had the corned beef burrito , which was very good , though a bit greasy and lacking . they have a good selection of veggie options and happy hour specials and are very attentive to your meal . it's clean , spacious , and the ambiance is great . i have definitely come here when my friends visit vegas to see if they have any other option . the best part about going here is the sitting area outside where you can hang out while eating all you could eat .</p>
5	<p>love this place .i have visited all of the great restaurants that offer all sorts of flavors , and to top it all off , all the facilities are extremely clean . \_UNK is the guy . he came for a quick check up , took my dad to the bar and came out free of charge !! seriously !</p>