Sequence Generation with Optimal-Transport-Enhanced Reinforcement Learning

Liqun Chen, Ke Bai, Chenyang Tao, Yizhe Zhang, Guoyin Wang, Wenlin Wang, Ricardo Henao, Lawrence Carin
Duke University

Abstract

Reinforcement learning (RL) has been widely used to aid training in language generation. This is achieved by enhancing standard maximum likelihood objectives with user-specified reward functions that encourage global semantic consistency. We propose a principled approach to address the difficulties associated with RL-based solutions, namely, high-variance gradients, uninformative rewards and brittle training. By leveraging the optimal transport distance, we introduce a regularizer that significantly alleviates the above issues. Our formulation emphasizes the preservation of semantic features, enabling end-to-end training instead of ad-hoc fine-tuning, and when combined with RL, it controls the exploration space for more efficient model updates. To validate the effectiveness of the proposed solution, we perform a comprehensive evaluation covering a wide variety of NLP tasks: machine translation, abstractive text summarization and image caption, with consistent improvements over competing solutions.

Introduction

Sequence generation is one of the central research topics in natural language processing (NLP) studies, covering a broad spectrum of applications including machine translation (Sutskever, Vinyals, and Le, 2014; Cho et al., 2014), abstractive summarization (Rush, Chopra, and Weston, 2015; Chopra, Auli, and Rush, 2016), image captioning (Vinyals et al., 2015; Xu et al., 2015), and style transfer (Shen et al., 2017; Prabhumoye et al., 2018).

Maximum likelihood estimation (MLE) is employed widely for training a sequence generation model. Specifically, given the sequential nature of language, MLE training follows an autoregressive design that maximizes the expected log-likelihood of a word conditioned on its previous context, i.e., the information encoded by the already observed input sequence. This practice leads to the well-known exposure bias issue, where the lack of exposure of the model to generated sequences during the training phase leads to the fast degeneration of semantic consistency for long sequences at test time (Bengio et al., 2015). This has motivated development of regularization schemes that involve the model-generated sequence at training time, resulting in marked performance gains (Rennie et al., 2016; Norouzi et al., 2016; Rennie et al., 2017).

Among various empirical attempts to regularize MLE, the use of reinforcement learning (RL) techniques has been investigated extensively (Rennie et al., 2016; Norouzi et al., 2016; Rennie et al., 2017; Paulus, Xiong, and Socher, 2018; Wu et al., 2016). RL leverages a set of principled learning algorithms, such as REINFORCE (Williams, 1992) and TRPO (Schulman et al., 2015), to optimize the model over desirable but often non-differentiable performance metrics. In RL-based approaches these metrics are termed reward functions. Relative to the likelihood, reward functions allow the model to be informed by specialized performance metrics that can be understood as regularizers to the standard MLE objective (Tan et al., 2018).

Crucial to the success of RL-regularized training is the choice of the reward function. Current practice largely falls into two categories: (i) static rewards, such as phrase-matching-based metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and CIDEr (Vedantam, Lawrence Zitnick, and Parikh, 2015), routinely used in the evaluation of language models; and more recently, (ii) dynamic rewards that track changes of the model in feature space during training, and are typically implemented via adversarial learning (Yu et al., 2017; Chen et al., 2018).

Despite encouraging results (Rennie et al., 2017; Paulus, Xiong, and Socher, 2018), limitations of both types of rewards have been widely recognized. Static rewards usually serve as imperfect proxies for human evaluation, resulting in undesirable biases (Wang et al., 2018). While being more objective, adversarial supervision in dynamic rewards relies on the delicate balance of a mini-max objective, which can be easily undermined in practice (Arjovsky and Bottou, 2017; Zhang et al., 2017; Chen et al., 2018). Both types of rewards can incur destabilization when training, manifested as mode-dropping and gradient-vanishing. Consequently, careful fine tuning is often required in the training of RL-regularized NLP models.

*Optimal transport* (OT) has emerged recently as a powerful learning framework, demonstrating success on some of the frontier challenges in machine learning, such as natural image generation (Salimans et al., 2018) and graph learning (Xu et al., 2019). OT-based learning generally reformu-
Reinforcement learning (RL) can be used as a means for evaluating and optimizing model parameters over flexible performance metrics in NLP tasks. Examples of these metrics are BLEU, ROUGE and CIDEr. Any dynamic (autoregressive) generative model for sequence generation can be viewed as an agent that interacts with an environment, i.e., word (token) sequences. Well-known examples of these generative models are the long-short term memory (LSTM) cell [Hochreiter and Schmidhuber (1997)] and the Transformer [Vaswani et al. (2017)].

Predictions from the agent that result in the (unobserved) next word of the sequence constitute the action of an executed policy $p_{\theta}$, where $\theta$ denotes the parameters of the agent, i.e., the generative model, $\mathcal{M}(\cdot; \theta)$. During training, once the agent has reached the end of the sequence, a reward is issued. The reward function $r(\cdot; \cdot)$, e.g., an evaluation metric such as BLEU [Papineni et al. (2002)], is calculated by comparing the sequence generated by the model with a reference, i.e., the ground-truth sequence. The loss function in RL-based training can be written as

$$\mathcal{L}_{\text{RL}}(\theta) = -E_{y \sim p_\theta} [r(y, y^*)],$$

(1)

where $y$ is the sequence sampled from $p_\theta$, $y^*$ is the reference, and the expectation is with respect to the policy $p_\theta$.

In scenarios where either sequence sampling or the reward function are not differentiable, we can use the REINFORCE algorithm [Williams (1992)] to approximate the gradient of (1). Specifically, we can approximate the gradient of $\mathcal{L}_{\text{RL}}(\theta)$ as:

$$\nabla_{\theta} \mathcal{L}_{\text{RL}}(\theta) = -E_{y \sim p_\theta} [r(y, y^*)] \nabla_{\theta} \log p_\theta(y),$$

(2)

where the expectation is approximated by Monte Carlo sampling from $p_\theta$, i.e., the probability of each generated word, resulting in actions manifested as each word of sequence $y$.

To decrease the variance of the gradient of (1), a baseline (function) to the reward can be added to (2), while preserving the unbiasedness of the gradient estimation [Paisley, Blei, and Jordan (2012)]. This baseline may be an arbitrary function depending on the history of $y$ not its future, i.e., a parameterized value-function whose purpose is to estimate the unbiased future reward [Kanazato et al. (2016)]. The self-critic sequence training algorithm [Rennie et al. (2017)] considers the baseline function $b(\cdot; \cdot)$ as the reward collected at test time, which implies $b \triangleq r(y, y^*)$, where $y$ is the greedy decoded sequence whose elements are obtained from the generator at step $t$ via $y_t = \arg \max_{y_t} \, p_\theta(y_t| y_{<t}).$ Note the baseline is obtained at the end of the generating process and is not dependent of $y^*$ (the ground-truth sequence). Aided by the baseline, we can re-formulate (2) as

$$\nabla_{\theta} \mathcal{L}(\theta) = -E_{y \sim p_\theta} [(r(y, y^*) - b(y))] \nabla_{\theta} \log p_\theta(y),$$

(3)

where $r(\cdot; \cdot)$ and $b(\cdot; \cdot)$ compare the generated sequence with the reference and greedily decoded sequence, respectively. As a result, we seek to maximize the reward and at the same time minimize the difference between the most likely, greedily decoded sequence, and samples from $p_\theta$. Intuitively, if the evaluation metric value of the sampled sentence $y$ is larger than that of the greedily decoded sentence $y^*$, minimizing the loss is equivalent to maximizing the probability of the greedily chosen words [Rennie et al. (2017)].
OT for Semantic Matching

Optimal transport (OT) can provide a dynamic (independent of position) 1-gram matching scheme for semantic matching [Chen et al., 2019]. In a nutshell, it computes the moving distance between the generated sequence, and similarly for the ground-truth one-hot-encoded input sequence but using the ground-truth one-hot-encoded input sequence.

In an optimal transport approach for sequence generation, we need to evaluate distances between words, that here we perform in word embedding space, to learn both the parameters of the model \( \theta \) and the embedding matrix \( E \). In particular, we use the word embeddings predicted by the model, given by \( \tilde{z}_t = E^T \hat{w}_t \), where \( E \in \mathbb{R}^{V \times d} \) is the word embedding matrix, \( V \) is the vocabulary size, \( d \) is the dimensionality of the embedding vector, and \( \hat{w}_t \) is the word generated by the model by sampling from a Gumbel-softmax design [Jang, Gu, and Poole, 2016] [Maddison, Mnih, and Teh 2017]. Under Gumbel-softmax, outputs from the sequence generation model \( v_t \), produce \( \hat{w}_t = \text{softmax}(\frac{E v_t + \tau}{\tau}) \), where \( \tau \) is the temperature parameter, and elements of \( g \) are sampled i.i.d. from Gumbel(0,1) [Maddison, Mnih, and Teh, 2017].

This Gumbel-softmax formulation is used for convenience since directly applying the non-differentiable \( \arg \max \) operator to obtain (discrete) one-hot represented tokens, as in the RL case, results in the inability to backpropagate gradients through the model to obtain parameter updates [Zhang et al., 2017; Yu et al., 2017]. By applying the Gumbel-softmax trick, we can collect embeddings for the entire predicted sequence \( Z_g = \{ \tilde{z}_i \}_{i=1}^n \). Similarly we denote the reference sequence embeddings as \( Z_r = \{ z_j \}_{j=1}^m \).

Using \( Z_r \) and \( Z_g \), we can compute the sequence-level optimal transport matrix \( T_{ij} \), using the inexact proximal point method for optimal transport (IPOT) algorithm [Xie et al., 2018].

We use the cosine similarity as the cost function: \( C_{ij} = C(\tilde{z}_i, z_j) = 1 - \frac{\tilde{z}_i^T z_j}{\| \tilde{z}_i \| \| z_j \|} \). The optimal transport distance is then formulated as:

\[
\mathcal{L}_{OT} = \min_{T_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} T_{ij} C_{ij},
\]

subject to \( \sum_{i} T_{ij} = d_j, \forall j \in [1, m] \) \( \sum_{j} T_{ij} = d_i, \forall i \in [1, n] \),

where \( T_{ij} \) is the transport matrix, \( d_i = \frac{a_i}{\sum_j a_j} \), \( a_i \) is the number of times the \( i \)-th word appears in the generated sentence, and similarly for \( d_j \) and ground-truth sequence. The loss in (4) can be optimized with the IPOT algorithm [Xie et al., 2018], outlined in Algorithm 1. Note that an iterative algorithm can be problematic for building the computational graph in modern deep learning frameworks; however, it can be proved by the envelope algorithm that the gradient flow through the transport matrix \( T_{ij} \) can be ignored [Carter 2001]. Therefore, this algorithm can be executed efficiently as gradients only depend on the cost function \( C(\cdot, \cdot) \) (cosine similarity).

**Optimal-Transport-Enhanced RL**

OT-only sequence modeling solution considered in [Chen et al., 2019] basically describes a dynamic 1-gram matching scheme, whose training is generally stable. However, the diversity of natural language largely comes from the compositionality of \( k \)-grams, which simple 1-gram cannot capture. It is challenging to extended OT to dynamic \( k \)-gram matching. Specifically, we will have \( \binom{n}{k} \times \binom{m}{k} \) distinct combinations of \( k \)-gram pairs to match between two sentences of length \( n \) and \( m \). On the other hand, RL-only solutions suffer weak learning signals due to high gradient variance; though it is capable of matching long phrases, i.e., \( k \)-grams for arbitrary \( k \).

So motivated, we combine OT and RL to improve models for sequence generation. OT helps stabilizing training while encouraging semantic consistency, while RL helps to capture consistency in long phrases. The proposed OT-enhanced RL (OTRL) approach is constructed as a hybrid algorithm. The complete loss function for the model is written as:

\[
\mathcal{L} = \lambda_1 \mathcal{L}_{MLE} + \lambda_2 \mathcal{L}_{OT} + \lambda_3 \mathcal{L}_{RL},
\]

where \( \mathcal{L}_{MLE} \) denotes the standard maximum likelihood objective of the sequence generation model of choice.

In practice, the sequence generation model will be first trained with the most restrictive loss, i.e., \( \mathcal{L}_{MLE} \), to get a good initial starting point. Then we gradually expand the exploration space by introducing the optimal transport loss \( \mathcal{L}_{OT} \). This allows the model to search words during the generation process that account for both dynamic position matching and key word preservation. Finally we add the RL-based self-critic loss \( \mathcal{L}_{RL} \), to fine-tune the model. This procedure can be implemented by parameter annealing. For instance, we can decrease \( \lambda_1 \) and increase \( \lambda_2 \) and \( \lambda_3 \) gradually when training. A similar annealing strategy is also employed in MIXER [Ranzato et al., 2016]. Detailed discussion about the choice of the hyper-parameters is provided in Section 2.

### Algorithm 1 IPOT algorithm.

1: **Input**: Embeddings \( Z_g = \{ \tilde{z}_i \}_{i=1}^n \), \( Z_r = \{ z_j \}_{j=1}^m \) and generalized step size \( \beta \).
2: \( \sigma = \frac{1}{m} 1_m \); \( T^{(1)} = 1_n 1_n^T \).
3: \( C_{ij} = C(\tilde{z}_i, z_j) \);
4: **for** \( t = 1, 2, 3 \ldots **do**
5: \( Q = A \odot T^{(t)} // \odot \) is the Hadamard product
6: \( \delta = \frac{1}{m} Q \sigma \); \( \sigma = m_{Q} \delta \).
7: **end for**
8: \( T^{(t+1)} = \text{diag}(\delta) Q \text{diag}(\sigma) \).
9: **end for**
10: Return \( T, C = \{ C_{ij} \} \).
To improve the performance of RL-based methods, recent regularized policy optimization (ERPO) (Tan et al., 2018) rowing ideas from MLE and policy gradient. All these algorithms (RAML) (Norouzi et al., 2016) is a hybrid RL method borrowing ideas from MLE and policy gradient. Here, the reward augmented maximum likelihood algorithm (RAML) (Norouzi et al., 2016) is a hybrid RL method borrowing ideas from MLE and policy gradient. All these algorithms can be understood as some special cases of entropy-regularized policy optimization (ERPO) (Tan et al., 2018). To improve the performance of RL-based methods, recent work revisits traditional variance-reduction techniques, i.e., offsetting the reward with a greedy roll-out baseline [2]. Our work considers a new framework that seeks to execute systematic policy exploration space via OT, to decrease the variance of the policy optimization.

### Related Work

**RL in sequence generation** A variety of RL approaches have been proposed to further improve sequence generation over MLE-based learning. In Ranzato et al. (2016) used the REINFORCE algorithm (Williams, 1992) and Bahdanau et al. (2017) applied the actor-critic algorithm. These two approaches only use a RL-inspired loss to fine-tune the model pre-trained via MLE. In Ding and Soricut (2017), they proposed the softmax policy gradient (SPG) algorithm to allow for end-to-end training without a pre-training phase. Further, the reward augmented maximum likelihood algorithm (RAML) (Norouzi et al., 2016) is a hybrid RL method borrowing ideas from MLE and policy gradient. All these algorithms can be understood as some special cases of entropy-regularized policy optimization (ERPO) (Tan et al., 2018). To improve the performance of RL-based methods, recent work revisits traditional variance-reduction techniques, i.e., offsetting the reward with a greedy roll-out baseline [2]. Our work considers a new framework that seeks to execute systematic policy exploration space via OT, to decrease the variance of the policy optimization.

**GAN in sequence generation** The generative adversarial network (GAN) (Goodfellow et al., 2014) has achieved success in many computer vision tasks. Several works in NLP have also tried to incorporate GAN-inspired losses for text generation (Yu et al., 2017; Zhang et al., 2017; Chen et al., 2018). Specifically, a discriminator is designed to compare the generated sequences with references to tell which ones are real or synthesized. This implies that the feedback from the discriminator into the generator is a sequence-based (global) loss. However, the discriminator of GAN is often hard to train for discrete observations. This happens because it is very easy for the discriminator to differentiate between real and synthetic sequences, and consequently the gradient from the discriminator can easily vanish. Further, there is still no principled way to avoid the so-called mode-collapsing problem in GAN training (Arjovsky, Chintala, and Bottou, 2017a). Alternatively, our framework, unlike GAN, is robust to misspecification and does not require an additional design for the discriminator.

### Optimal transport in NLP

Optimal transport has been widely applied in computer vision applications (Rubner, Tomasi, and Guibas, 2000), and recently it has been applied to NLP tasks. In Kusner et al. (2015), the authors propose word mover’s distance (WMD) for document classification and word embedding refinement, while in Alvarez-Melis and Jaakkola (2018) there is a goal of aligning the word embedding space with Gromov-Wasserstein distance (Memoli, 2011) in an unsupervised way. Most recently, Chen et al. (2019) used OT as a sequence-level loss for sequence generation, as a regularization term in MLE-based learning. Here, optimal transport is used initially to refine word embeddings and capture key phrases, and later to regularize the RL objective.

### Datasets and setup

We consider three tasks to evaluate our model: i) **Machine translation**: the commonly used English-German translation dataset, IWSLT 2014 (Cettolo et al., 2014), which was also applied in Ranzato et al. (2016); Norouzi et al. (2016); Chen et al. (2019). We employ the same pre-processing in Ranzato et al. (2016); Norouzi et al. (2016). The EN-DE dataset has 146K/7K/7K paired sentences for training/validation/testing, respectively. The vocabulary size for English is 23K, and 32K for German. For EN-VI dataset, we train the model on IWSLT15 with 133K paired sentences, and test on IWSLT 2012 news dataset (NT12) and 2013 news dataset (NT13) following the same setup as Chen et al. (2019). ii) **Abstractive summarization**: two different datasets are employed for the summarization task, English gigawords (Graff et al., 2003) and CNN/Daily-mail (Hermann et al., 2015; Nallapati et al., 2016), which are standard summarization benchmarks. The English gigawords corpus consists of 200K/8K/2K source and target sentence pairs in training/validation/testing, respectively. The CNN/Daily-mail dataset has 287,113 training pairs, 13,368 validation pairs and 11,490 testing pairs. iii) **Image captioning**: we also consider an image captioning task with the COCO dataset (Lin et al., 2014), which contains 123,287 images in total and each image is annotated with at least 5 captions. Following the split method proposed in (Karpayvich and Fei-Fei, 2015), 113,287 images are used for training and 5,000 images are used for validation and testing.

For fair comparison, we use Texar (Hu et al., 2018) as our codebase to implement all proposed methods and baselines, and to compared models for IWSLT 2014 and gigawords.
Table 1: Results of abstractive summarization on CNN/Daily-mail. First section is the MLE-training baseline with Seq2Seq model. Models in the second section are results trained using OpenNMT codebase \cite{Gehrmann2018}. The last section shows the results of our baselines and the proposed model.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>SummaRuNNer \cite{Nallapati2017}</td>
<td>39.6</td>
<td>16.2</td>
<td>35.3</td>
</tr>
<tr>
<td>Pointer-Generator \cite{See2017}</td>
<td>36.44</td>
<td>15.66</td>
<td>33.42</td>
</tr>
<tr>
<td>Pointer-Generator + Coverage \cite{See2017}</td>
<td>39.53</td>
<td>17.28</td>
<td>36.38</td>
</tr>
<tr>
<td>Saliency + Entailment reward \cite{Pasunuru2018}</td>
<td>40.43</td>
<td>18.00</td>
<td>37.10</td>
</tr>
<tr>
<td>Key information guide network \cite{Li2018}</td>
<td>38.95</td>
<td>17.12</td>
<td>35.68</td>
</tr>
<tr>
<td>Sentence Rewriting \cite{Chen2018}</td>
<td>40.88</td>
<td>17.80</td>
<td>38.54</td>
</tr>
<tr>
<td>Inconsistency loss \cite{Hsu2018}</td>
<td>40.68</td>
<td>17.97</td>
<td>37.13</td>
</tr>
<tr>
<td>ML + RL \cite{Paulus2018}</td>
<td>38.30</td>
<td>14.81</td>
<td>35.49</td>
</tr>
<tr>
<td>Pointer-Generator 36.25</td>
<td>16.17</td>
<td>33.41</td>
<td></td>
</tr>
<tr>
<td>Pointer-Generator + Coverage Penalty 39.12</td>
<td>17.35</td>
<td>36.12</td>
<td></td>
</tr>
<tr>
<td>Transformer + Copy + Coverage Penalty 39.25</td>
<td>17.54</td>
<td>36.45</td>
<td></td>
</tr>
<tr>
<td>Bottom-Up Summarization 41.22</td>
<td>18.68</td>
<td>38.34</td>
<td></td>
</tr>
<tr>
<td>Pointer-Generator + OT 39.75</td>
<td>17.87</td>
<td>38.22</td>
<td></td>
</tr>
<tr>
<td>Pointer-Generator + Coverage Penalty + OT 40.28</td>
<td>18.17</td>
<td>38.51</td>
<td></td>
</tr>
<tr>
<td>Pointer-Generator+ Coverage Penalty + OTRL 41.40</td>
<td>18.22</td>
<td>38.86</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results on the testing set on COCO-caption dataset on the model with the best evaluation loss. The first section are results from previous image caption models. The second section are results of our baselines and the proposed model.

<table>
<thead>
<tr>
<th>Method</th>
<th>CIDEr</th>
<th>BLEU-4</th>
<th>BLUE-3</th>
<th>BLEU-2</th>
<th>BLEU-1</th>
<th>ROUGE</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Attention \cite{Xu2015}</td>
<td>-</td>
<td>24.3</td>
<td>34.4</td>
<td>49.2</td>
<td>70.7</td>
<td>-</td>
<td>23.9</td>
</tr>
<tr>
<td>Hard Attention \cite{Xu2015}</td>
<td>-</td>
<td>25.0</td>
<td>35.7</td>
<td>50.4</td>
<td>71.8</td>
<td>-</td>
<td>23.0</td>
</tr>
<tr>
<td>Show &amp; Tell \cite{Vinyals2015}</td>
<td>85.5</td>
<td>27.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23.7</td>
</tr>
<tr>
<td>ATT-FCN \cite{You2016}</td>
<td>-</td>
<td>30.4</td>
<td>40.2</td>
<td>53.7</td>
<td>70.9</td>
<td>-</td>
<td>24.3</td>
</tr>
<tr>
<td>SCN-LSTM \cite{Gan2017}</td>
<td>101.2</td>
<td>33.0</td>
<td>43.3</td>
<td>56.6</td>
<td>72.8</td>
<td>-</td>
<td>25.7</td>
</tr>
<tr>
<td>Adaptive Attention \cite{Lu2017}</td>
<td>108.5</td>
<td>33.2</td>
<td>43.9</td>
<td>58.0</td>
<td>74.2</td>
<td>-</td>
<td>26.6</td>
</tr>
<tr>
<td>MLE</td>
<td>106.3</td>
<td>34.3</td>
<td>45.3</td>
<td>59.3</td>
<td>73.6</td>
<td>55.2</td>
<td>26.2</td>
</tr>
<tr>
<td>MLE+OT</td>
<td>107.9</td>
<td>34.8</td>
<td>46.1</td>
<td>60.1</td>
<td>76.2</td>
<td>55.6</td>
<td>26.5</td>
</tr>
<tr>
<td>MLE+SC</td>
<td>110.4</td>
<td>33.4</td>
<td>45.8</td>
<td>61.4</td>
<td>79.5</td>
<td>56.0</td>
<td>26.6</td>
</tr>
<tr>
<td>MLE+OTRL</td>
<td><strong>111.8</strong></td>
<td>34.4</td>
<td><strong>46.6</strong></td>
<td><strong>61.8</strong></td>
<td><strong>79.3</strong></td>
<td><strong>56.2</strong></td>
<td><strong>26.8</strong></td>
</tr>
</tbody>
</table>

The performance of our implementation is at least comparable with the reported results. Both encoder and decoder are one-layer LSTMs with 256 hidden states. The word embedding dimension is also set to 256. Dropout is also applied with rate 0.2. We use Adam optimization with learning rate 0.001 and batch size = 64 for training.

For the CNN/Daily-mail experiment, we preprocess the data using the same method proposed in \cite{See2017}. We use the most recent extractor-abstractor framework \cite{Chen2018}, \cite{Gehrmann2018}, \cite{Rush2018}, consisting of two parts, extractor and abstractor. We implement the extractor following the structure from \cite{Chen2018}, and modify the abstractor using our OTRL loss function without changing the network architecture, in comparison to the OpenNMT implementation \cite{Klein2017}. The word embedding size is set to 128. The encoder and decoder are also both one-layer LSTMs with 256 hidden states.

The image captioning experiment is implemented based on a public Pytorch implementation \cite{Paszke2017}. We pre-train the image feature extractor using the bottom-up attention model \cite{Anderson2018}. The setup of the caption decoder is the same as self-critic model \cite{Rennie2017}, where the image feature is fed into a one-layer LSTM. For the LSTM, the hidden state dimension and the dimension of word vectors are both set to 512. All experiments are performed on one NVIDIA TITAN X GPU.

**Neural machine translation**

The commonly applied metric BLEU score \cite{Papineni2002} is used to evaluate the performance of different models. To assess variation, we test each model 5 times with

https://github.com/ruotianluo/self-critical.pytorch
different random seeds and calculate the mean and standard deviation for each test run. The results in Table 3 show that our model results in consistent improvement over the baselines, across all three datasets.

**Abstractive summarization**

ROUGE-1, -2 and -L scores [Lin (2004)] are employed to evaluate the performance of the abstractive summarization model.

**English gigawords** Table 4 shows performance comparisons of seven baselines and our model. We implement all baseline models on Texar [Hu et al. (2018)] under the same setup as ours. Our model achieves the highest score on the test dataset on ROUGE-1, -2, -L metrics consistently. Generated examples are found in the Appendix.

**CNN/Daily-mail** We also run our model on the large summarization dataset CNN/Daily-mail. Table 1 shows ROUGE score results on different models. We compare our model with the most recent works, e.g., Gehrmann, Deng, and Rush (2018); Chen and Bansal (2018). Our OTRL framework achieves better ROUGE-1 and ROUGE-L score on this dataset, indicating that our model’s predictions can better capture the semantic meaning from the source paragraph.

**Image captioning**

For the image caption task, the OT loss between the generated sentences and the ground truth sentences is employed to regularize training. Image features are extracted from fast R-CNN [Ren et al. (2015)] following the same setup as [Anderson et al., 2018], which is denoted as bottom-up features.

We perform our comparisons based on two different scenarios: teacher-forcing MLE objective with attention mechanism and the self-critic RL objective. In the teacher-forcing setup, we select the checkpoint which has the minimal loss on the validation set from each model. In the self-critic setup, we select the checkpoint with the largest evaluation CIDEr score [Vedantam, Lawrence Zitnick, and Parikh (2015)] from each model. We then evaluate all models using different metrics on the testing set. The split of the training, evaluation and test set is the same as in [Karpathy and Fei-Fei (2015)], for fair comparison, and the detailed results are shown in Table 2. We report BLEU-\(k\) (\(k\) from 1 to 4) [Papineni et al. (2002)], CIDEr [Vedantam, Lawrence Zitnick, and Parikh, 2015], and METEOR [Banerjee and Lavie (2005)] scores. It is observed that by employing optimal transport, under two different setups, our OTRL framework consistently outperforms all baselines and compared methods.

**Analysis**

**Hyper-parameter setup**

For hyper-parameters \(\lambda_1, \lambda_2, \lambda_3\), we use parameter grid-search to find the best setup for different tasks. In machine translation and abstractive summarization, We set \(\lambda_1 = 0.9, \lambda_2 = 0.1, \lambda_3 = 0\) as our initialization. Then we gradually decrease \(\lambda_1\) to 0.7, and increase \(\lambda_3\) to 0.2. This annealing process starts after the 7-th epoch. Note that the annealing process is implemented as a linear decay function provided in Tensorflow (Abadi et al., 2016)/Pytorch (Paszke et al., 2017). For the image-captioning task, the training is divided into two phases. In the first phase, we set \(\lambda_1 = 0.9, \lambda_2 = 0.1, \lambda_3 = 0\) and choose the best on the evaluation set as the initializer for the second phase, where \(\lambda_1 = 0, \lambda_2 = 0.1, \lambda_3 = 1\). The annealing trace plot is provided in the Appendix.

**Ablation study**

To evaluate the stability of different frameworks, in Figure 2 we show the convergence curve of the ROUGE-1 F1 score on the testing dataset for each training epoch. We test different methods on the English gigawords dataset with the same sequence generation architecture. The num-
The help of OT, resulting in a smaller variance than the exploration space of the blue line is presumably controlled by the hyper-parameter setup. This phenomenon implies that the blue line converges faster than the red line with the same training objective. As shown in the Figure 3, we initialize the validation dataset with respect to the loss switching in age captioning, to show how the CIDEr score changes on the validation set drops significantly, indicating the exploration space becomes larger, so that the performance drops dramatically after the switching. For the blue line experiment, the objective is regularized with OT. When the MLE+OT loss switched to OTRL, the performance does not drop significantly, and the blue line converges faster than the red line with the same hyper-parameter setup. This phenomenon implies that the exploration space of the blue line is presumably controlled with the help of OT, resulting in a smaller variance than the red line, thus a more efficient training.

**Conclusions**

We have presented a new framework to control (regularize) the exploration space of policy optimization. This also provides a new way to reduce the variance of RL approaches in NLP tasks, so that RL-based sequence generation models can be further improved. We evaluate our new sequence level loss on three NLP tasks. The consistent improvement in machine translation, abstractive summarization and image captioning strongly support our motivations.

**Acknowledgments**

The authors would like to thank the anonymous reviewers for their insightful comments. This research was supported in part by DARPA, DOE, NIH, ONR, NSF and RTI Internal research & development funds.

<table>
<thead>
<tr>
<th>Model</th>
<th>EN-DE</th>
<th>EN-VI</th>
<th>EN-VI</th>
<th>EN-VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>26.44 ± 0.18</td>
<td>23.76 ± 0.17</td>
<td>26.10 ± 0.07</td>
<td></td>
</tr>
<tr>
<td>Scheduled Sampling</td>
<td>24.11 ± 0.10</td>
<td>23.71 ± 0.14</td>
<td>26.11 ± 0.12</td>
<td></td>
</tr>
<tr>
<td>RAML</td>
<td>27.22 ± 0.14</td>
<td>24.24 ± 0.16</td>
<td>26.42 ± 0.17</td>
<td></td>
</tr>
<tr>
<td>MIXER</td>
<td>26.53 ± 0.11</td>
<td>23.81 ± 0.18</td>
<td>26.33 ± 0.17</td>
<td></td>
</tr>
<tr>
<td>SPG</td>
<td>26.62 ± 0.05</td>
<td>24.27 ± 0.11</td>
<td>26.51 ± 0.09</td>
<td></td>
</tr>
<tr>
<td>ERPL</td>
<td>27.82 ± 0.11</td>
<td>24.56 ± 0.21</td>
<td>26.88 ± 0.13</td>
<td></td>
</tr>
<tr>
<td>Seq2Seq-OT</td>
<td>27.79 ± 0.12</td>
<td>24.46 ± 0.11</td>
<td>26.91 ± 0.13</td>
<td></td>
</tr>
<tr>
<td>MLE+OTRL</td>
<td>28.13 ± 0.11</td>
<td>24.85 ± 0.13</td>
<td>27.11 ± 0.18</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Results of machine translation for EN-DE and EN-VI.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>36.11 ± 0.21</td>
<td>16.39 ± 0.16</td>
<td>32.32 ± 0.19</td>
</tr>
<tr>
<td>RAML</td>
<td>36.30 ± 0.04</td>
<td>16.69 ± 0.20</td>
<td>32.49 ± 0.17</td>
</tr>
<tr>
<td>SPG</td>
<td>36.48 ± 0.24</td>
<td>16.84 ± 0.26</td>
<td>32.79 ± 0.26</td>
</tr>
<tr>
<td>MIXER</td>
<td>36.34 ± 0.23</td>
<td>16.61 ± 0.25</td>
<td>32.57 ± 0.15</td>
</tr>
<tr>
<td>Scheduled Sampling</td>
<td>36.59 ± 0.12</td>
<td>16.79 ± 0.22</td>
<td>32.77 ± 0.17</td>
</tr>
<tr>
<td>ERPL</td>
<td>36.72 ± 0.29</td>
<td>16.99 ± 0.17</td>
<td>32.95 ± 0.33</td>
</tr>
<tr>
<td>Seq2Seq-OT</td>
<td>36.82 ± 0.25</td>
<td>17.22 ± 0.16</td>
<td>33.15 ± 0.23</td>
</tr>
<tr>
<td>MLE+OTRL</td>
<td>37.2 ± 0.19</td>
<td>17.63 ± 0.15</td>
<td>33.76 ± 0.13</td>
</tr>
</tbody>
</table>

Table 4: Results of text summarization on gigawords dataset.


Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.


