DATA-DEPENDENT GAUSSIAN PRIOR OBJECTIVE FOR LANGUAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

For typical sequence prediction problems like language generation, maximum likelihood estimation (MLE) has been commonly adopted as it encourages the predicted sequence most consistent with the ground-truth sequence to have the highest probability of occurring. However, MLE focuses on a once-for-all matching between the predicted sequence and gold-standard consequently, treating all incorrect predictions as being equally incorrect. We call such a drawback *negative diversity ignorance* in this paper. Treating all incorrect predictions as equal unfairly downplays the nuance of these sequences' detailed token-wise structure. To counteract this, we augment the MLE loss by introducing an extra KL divergence term which is derived from comparing a data-dependent Gaussian prior and the detailed training prediction. The proposed data-dependent Gaussian prior objective (D2GPo) is defined over a prior topological order of tokens, poles apart from the data-independent Gaussian prior (L2 regularization) commonly adopted for smoothing the training of MLE. Experimental results show that the proposed method can effectively make use of more detailed prior in the data and significantly improve the performance of typical language generation tasks, including supervised and unsupervised machine translation, text summarization, storytelling, and image caption.

1 INTRODUCTION

Language understanding is the crown jewel of artificial intelligence. As the well-known dictum by Richard Feynman states, "what I cannot create, I do not understand;" language generation reflects the level of development of language understanding. Language generation models have seen remarkable advances in recent years, especially under the rapid development of deep neural networks (DNNs). There are several typical models for language generation such as sequence-to-sequence (seq2seq) models (Kalchbrenner & Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017), generative adversarial networks (GANs) (Goodfellow et al., 2014), variational autoencoders (VAEs) (Kingma & Welling, 2013), and auto-regressive networks (Larochelle & Murray, 2011; Van Oord et al., 2016).

Language generation is usually modeled as a sequence prediction task, which adopts maximum likelihood estimation (MLE) as standard training criterion (i.e., objective). MLE has had much success owing to its intuitiveness and flexibility. However, sequence prediction has a series of problems due to MLE:

- Exposure bias: the model is not exposed to the full range of errors during training;
- Loss mismatch: during training, we maximize the log-likelihood, whereas, during inference, the model is evaluated by a different metric such as BLEU or ROUGE;
- Generation diversity: the generations are dull, generic (Sordoni et al., 2015; Serban et al., 2016; Li et al., 2016a), repetitive, and short-sighted (Li et al., 2016b);
- Negative diversity ignorance: MLE fails to assign proper scores to different incorrect model outputs, which means that all incorrect outputs are treated equally during training.

There has been a variety of work alleviating the above MLE training shortcomings apart from negative diversity ignorance. Negative diversity ignorance is a result of unfairly downplaying the nuance

of sequences' detailed token-wise structure. When the MLE objective compares its predicted and ground-truth sequences, it takes a once-for-all matching strategy; the predicted sequence is given a binary label, either correct or incorrect. However, these incorrect training predictions may be quite diverse and letting the model be aware of which incorrect predictions are more incorrect or less incorrect than others may more effectively guide model training. For instance, an *armchair* might be mistaken with a *deckchair*, but it should usually not be mistaken for a *mushroom*.

To alleviate the issue of the negative diversity ignorance, we add an extra Gaussian prior objective to augment the current MLE training with an extra Kullback-Leibler (KL) divergence loss term. The extra loss is computed by comparing two probability distributions, the first of which comes from the detailed model training prediction, and the second of which is from a ground-truth token-wise distribution and is defined as a kind of data-dependent Gaussian prior distribution. The proposed data-dependent Gaussian prior objective (D2GPo) is then injected into the final loss through a KL divergence term. The D2GPo is poles apart from commonly adopted data-independent Gaussian prior (L2 regularization) for the purpose of smoothing the training of MLE, which is also directly added into the MLE loss.

Experimental results show that the proposed method can effectively make use of a more detailed prior in the data and significantly improve the performance of typical language generation tasks, including supervised and unsupervised machine translation, text summarization, storytelling, and image caption.

2 Related Work

Natural language generation (NLG) has long been considered the most challenging natural language processing (NLP) task (Murty & Kabadi, 1987). NLG techniques have been widely adopted as the critical module in various tasks, including control-free sentence or poem generation (Zhang & Lapata, 2014) and input-conditioned language generation such as machine translation, image caption, text summarization, storytelling (Vaswani et al., 2017; Lample et al., 2018; Karpathy & Fei-Fei, 2015; Fan et al., 2018), and sentiment/tense controlled sentence generation (Hu et al., 2017). In this work, we focus on input-conditioned language generation tasks, though, our proposed method can also be applied to other language generation fields.

Input-conditioned language generation tasks are challenging because there is an information imbalance between the input and output in these tasks, especially for cases with non-text input (Shapiro, 1992). Reiter & Dale (2000) discussed different ways of building complicated knowledge-based systems for NLG. In recent years, neural networks (NNs), especially DNNs, have shown promising results in many NLP tasks. Bengio et al. (2003) first proposed the NN language model (NNLM) to exploit the advantages of NNs for language generation tasks. In an NNLM, the n-gram paradigm is extended by the generalization ability of NNs. Mikolov et al. (2010) developed a more general implementation for a language model (called the recurrent NN language model (RNNLM) by integrating a Markov property using a recurrent NN (RNN) to address NNLMs' theoretical inability to capture long-term dependencies. RNNLM is an effective solution because it is designed to capture long-term dependencies. Because of the vanishing gradient problem in RNNs, however, its long-term dependency processing capability is limited. Radford et al. (2018) proposed a Transformer language model called GPT, which uses a left-to-right architecture, where every token can pay attention to previous tokens in the self-attention layers of the Transformer.

The generators of the most current language generation model use the RNNLM or Transformer LM structure. However, as pointed out by Bengio et al. (2015), fitting the distribution of observed data does not mean that satisfactory text will be generated, because the model is not exposed to the full range of errors during training. This is called the *exposure bias* problem. Reinforcement learning, GANs (Goodfellow et al., 2014; Yu et al., 2017), and end-to-end re-parameterization (Kusner & Hernández-Lobato, 2016) techniques have been proposed to solve this problem. The *exposure bias* is no longer an issue in reinforcement learning models because the training sequences are generated by the model itself.

Using MLE for the training objective leads to the problem of *loss mismatch*. Ranzato et al. (2015) incorporated the evaluation metric into the training of sequence-to-sequence (seq2seq) models and proposed the mixed incremental cross-entropy reinforce (MIXER) training strategy, which is similar

to the idea of minimum risk training (Smith & Eisner, 2006; Li & Eisner, 2009; Ayana et al., 2016; Shen et al., 2016). MIXER uses decoder hidden states to predict the bias term and hence reduce the variance, while minimum risk training renormalizes the predicted probabilities. Zhang & Zhao (2018) introduced a new training criterion based on the Hellinger distance for the seq2seq model and empirically compared the models of two optimization categories: *minimum divergence* and *maximum margin*.

For the *generation diversity* problem, Serban et al. (2017) applied a latent variable hierarchical encoder–decoder dialog model to introduce utterance-level variations and facilitate longer responses. Zhao et al. (2017) presented a novel framework based on conditional variational autoencoders that improves generation diversity by sampling a latent variable z and optionally adding linguistic features to constrain the style further.

There is an increasing interest in incorporating problem field knowledge in machine learning approaches (Taskar et al., 2004; Ganchev et al., 2010; Hu et al., 2016). One common way is to design specialized network architectures or features for specific knowledge (e.g., Liang et al. (2017; 2018)). In contrast, for structured probabilistic models, posterior regularization (PR) and related frameworks (Ganchev et al., 2010; Liang et al., 2009; Bellare et al., 2009) provide a general means to impose knowledge constraints during model estimation. Hu et al. (2018) established a mathematical correspondence between posterior regularization and reinforcement learning, and, based on this connection, expanded posterior regularization to learn knowledge constraints as the extrinsic reward in reinforcement learning. Our approach can be seen as incorporating a prior knowledge of the language field into language generation learning.

Additionally, Welleck et al. (2019) proposed a new objective, unlikelihood training, which forces unlikely generations to be assigned lower probability by the model. The difference is that (Welleck et al., 2019) focuses on low-frequency words, while our model focuses on negative tokens.

3 BACKGROUND

Consider a conditional probability model for sequence predictions $\boldsymbol{y} \sim p_\theta(\boldsymbol{x})$ with parameters $\boldsymbol{\theta}$. The target sequence $\boldsymbol{y}$ can be conditioned on any type of source $\boldsymbol{x}$ (e.g., phrase, sentence, and passage in human languages or even image), which are omitted for simplicity of notation. For the sequence $\boldsymbol{y} = \langle y_1, y_2, ..., y_l \rangle$, the probability $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ is

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}) = p_\theta(y_1|\boldsymbol{x})p_\theta(y_2|\boldsymbol{x}, y_1)...p_\theta(y_l|\boldsymbol{x}, y_{1:l-1}). \tag{1}$$

Commonly, sequence prediction models are trained using MLE (also known as teacher forcing) (Williams & Zipser, 1989). MLE minimizes the negative log-likelihood of $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ as follows:

$$\mathcal{L}_{\textbf{MLE}}(\theta) = -\log p_\theta(\boldsymbol{y}|\boldsymbol{x}) = -\sum_{i=1}^{l} \log p_\theta(y_i|\boldsymbol{x}, \boldsymbol{y}_{<i}). \tag{2}$$

Optimizing the MLE objective $\mathcal{L}_{\textbf{MLE}}(\boldsymbol{\theta})$ is straightforward and meets the principle of empirical risk minimization while focusing on only minimizing losses of the correct target on the training data set.

However, there may be noise in the training data, and forcibly learning the distribution of a training set cannot enable the obtained model to reach good generalization. Additionally, for sequence prediction, models trained subject to MLE cursorily evaluate all predictions as either correct or incorrect and ignores the similarity between the correct and "less incorrect" predictions. Incorrect predictions might range from nearly perfect (i.e., one token is mistaken with a synonym) to completely wrong, having nothing in common with the gold sequence. However, MLE training treats all incorrect training predictions equally, which implies that MLE actually fails to accurately assign scores to diverse (especially negative) model predictions.

4 D2GPO: DATA-DEPENDENT GAUSSIAN PRIOR OBJECTIVE

To capture the diversity of negative training predictions, we augment the MLE objective of model with an additional objective $\mathcal{O}$ which more accurately models such a negative diversity. Without loss of generality, supposing $\tilde{\boldsymbol{y}}$ is the prediction candidate, we introduce a general evaluation

function $f(\tilde{\boldsymbol{y}}, \boldsymbol{y}) \in \mathbb{R}$ independent of model prediction, such that with a golden target token y^*, a higher $f(\tilde{y}, y^*)$ value indicates a better $p_\theta(\tilde{y}|\boldsymbol{x})$ for a target candidate $\tilde{y} \in V$ (where V is the target candidate set). Note that $f(\tilde{\boldsymbol{y}}, \boldsymbol{y})$ can also involve other factors such as latent variables and extra supervisions.

There are two main methods to learn $f(\tilde{\boldsymbol{y}}, \boldsymbol{y})$ in the model. If p_θ is a GAN-like implicit generative model or an explicit distribution that can be efficiently reparametrized (e.g., Gaussian) (Kingma & Welling, 2013), then one effective method is maximizing $\mathbb{E}_{p_\theta}\left[f(\tilde{\boldsymbol{y}}, \boldsymbol{y})\right]$. The other method is computing the gradient $\nabla_\theta \mathbb{E}_{p_\theta}\left[f(\tilde{\boldsymbol{y}}, \boldsymbol{y})\right]$ using the *log-derivative* trick which can suffer from high variance but is often used for the large set of non-parameterizable explicit distributions.

Corresponding to the probability distribution of model predictions $p_\theta(\cdot)$, we define a prior distribution $q(\boldsymbol{y})$ (for each target y_i, it has its own unique distribution of $q_i = q(y_i)$) which is extracted and derived from the ground-truth data (e.g., language text in language generation tasks). To guide the probability distribution of model predictions $p_\theta(\cdot)$ to match the prior probability distributions $q(\cdot)$, we adopt Kullback–Leibler (KL) divergence. Considered with the learning of the evaluation function $f(\tilde{\boldsymbol{y}}, \boldsymbol{y})$, the loss for objective $\mathcal{O}$ is calculated as follows:

$$\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, q) = KL(q(\boldsymbol{y}) \| p_\theta(\boldsymbol{y}|\boldsymbol{x})) - \alpha \mathbb{E}_q\left[f(\tilde{\boldsymbol{y}}, \boldsymbol{y})\right], \tag{3}$$

where α is a weight for the evaluation function learning term. Since we derive the prior distribution $q(\boldsymbol{y})$ from the ground-truth data (which is independent of model parameters θ), so that $\mathbb{E}_q\left[f(\tilde{\boldsymbol{y}}, \boldsymbol{y})\right]$=0. Hence, Eq. (3) becomes

$$\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, q) = KL(q(\boldsymbol{y}) \| p_\theta(\boldsymbol{y}|\boldsymbol{x})), \tag{4}$$

in which KL-divergence can be expanded as

$$KL(q \| p_\theta) = \mathbb{E}_p(\log(\frac{q}{p})) = \sum_i q_i * \log(q_i) - \sum_i q_i * \log(p_i). \tag{5}$$

The final objective for learning the model is written as follows:

$$\min_\theta \mathcal{L}_{\mathbf{MLE}}(\boldsymbol{\theta}) + \lambda \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, q), \tag{6}$$

where λ is the balancing hyperparameter. Because optimizing the original model objective $\mathcal{L}_{\mathbf{MLE}}(\boldsymbol{\theta})$ is straightforward, in the following, we omit the discussion of $\mathcal{L}_{\mathbf{MLE}}(\boldsymbol{\theta})$ and focus on the proposed $\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, q)$.

The prior probability distribution $q(y^*)$ on y^* can be obtained from the evaluation function $f(\cdot, \cdot)$ with a softmax operation. To expose the mass of the distribution over the classes, Hinton et al. (2015) introduced a softmax temperature mechanism, therefore, the relationship between q and $f(\tilde{\boldsymbol{y}}, \boldsymbol{y})$ is:

$$q(y^*) = \frac{exp(f(\tilde{\boldsymbol{y}}, y^*)/T)}{\sum_j exp(f(\tilde{y}_j, y^*)/T)}, \tag{7}$$

where T is a temperature parameter. When $T \to 0$, the distribution becomes a Kronecker distribution (and is equivalent to a one-hot target vector); when $T \to +\infty$, it becomes a uniform distribution. The softmax operation always turns any evaluation function $f(\cdot, \cdot)$ into a form of probability distribution no matter what the form of the original $f(\cdot, \cdot)$ is, thus then we will only focus on $f(\cdot, \cdot)$.

To find a good evaluation function, we have to mine token-wise diversity about every y^*. Considering all token types $\tilde{y}_j$ in a vocabulary, with respect to each y^*, there exists a prior topological order *ORDER*(y^*) among all the known tokens, in which y^* is always ranked top priority. Then the $f(\tilde{y}_j, y^*)$ can be defined as a monotonic function over the corresponding topological order so that it gives maximal value only when the input is y^* itself. Note that defining $f(\cdot, \cdot)$ in this way leads to the resulting q also monotonic over the corresponding topological order. Considering that q is *a priori*, it will be fixed throughout the learning process.

The remaining questions are about how to find a meaningful evaluation function $f(\cdot, \cdot)$ for the distribution q. In language generation tasks, we may conveniently take word embedding as the token representation, and let embedding distance determine such an order *ORDER*(y^*) for each y^*. In this work, we adopt the cosine similarity of pre-trained embeddings to sort the token (word / subword) order.

Discussion For the evaluation function $f(\cdot,\cdot)$ of q, we adopt the Gaussian probability density function (PDF), though later we also present experimental results from other types of functions in the ablation study. As the adopted Gaussian prior used in the training objective is derived from data-dependent token-wise distribution, we thus call it data-dependent Gaussian prior objective (D2GPo), a big departure from the Gaussian prior commonly adopted for smoothing in MLE training (we call it data-independent Gaussian prior). The following briefly explains why we chose the Gaussian PDF and how our D2GPo mathematically differs from data-independent Gaussian prior.

The central limit theorem indicates that suitably standardized sums of independent random variables have an approximately normal distribution. Thus, any random variable that arises as the sum of a sufficiently large number of small random components can be modeled accurately by a normal distribution. Embedding has a linear additive property (e.g., *king* - *man* + *woman* $\approx$ *queen*). The additive property of embedding can be explained by inspecting the training objective (Mikolov et al., 2013). Each dimension of an embedding represents a potential feature of the token. Considering each potential feature as an independent random variable, the sum follows a Gaussian distribution centered on the correct vocabulary unit y^* according to the linear additive property. Therefore, we can use a Gaussian distribution for the embedding distance determined order to effectively model distribution $q(y^*)$. The overview of the concepts underlying D2GPo is illustrated in Appendix A.1.

The D2GPo in this paper is different from the data-independent Gaussian prior in machine learning optimization theory. We hypothesize and experimentally verify that the embedding feature extracted from the data obeys the Gaussian distribution. The distribution from the prior knowledge of language data is used as a soft target to guide the model language generation process using knowledge distillation. The Gaussian prior in the machine learning optimization theory assumes that each component in the parameter θ is subject to a zero-mean Gaussian prior distribution, which is equivalent to L2 regularization. In general, our Gaussian prior objective is to act on the guiding target probability, while the Gaussian prior in machine learning is applied to the selection of model parameters.

5 Experiments and Results

In this section, we describe the experimental evaluation of the D2GPo on a variety of typical language generation tasks: neural machine translation (NMT), text summarization, storytelling, and image caption. The hyperparameters in D2GPo and effect analysis are shown in Appendix A.7.

5.1 Embedding pre-training

Our proposed D2GPo approach for experimental tasks require either word embeddings or byte-pair-encoding (BPE) (Sennrich et al., 2016b) subword embeddings. We generated the pretrained embeddings using fastText (Bojanowski et al., 2017) with an embedding dimension of 512, a context window of size 5 and 10 negative samples. For NMT, fastText was applied to the concatenation of source and target language monolingual corpora, which results in cross-lingual BPE subword embedding. For text summarization, we generated the BPE subword embedding only on the English monolingual corpora, while for the storytelling, and image caption, we obtained the word embedding also on the English monolingual corpora.

5.2 Supervised NMT

We evaluated the model on several widely used translation tasks: WMT14 English-to-German (EN–DE), English-to-French (EN–FR), and WMT16 English-to-Romanian (EN–RO)[1] tasks, which all have standard large-scale corpora for NMT evaluation. Due to the space limit, the data details are provided in Appendix A.3. The sentences were encoded using sub-word types based on BPE, which has a shared vocabulary of 40K sub-word units for all three tasks. We chose the Transformer NMT (Vaswani et al., 2017) model as our baseline. For the hyperparameters of the Transformer (base/big) models, we followed the settings used in Vaswani et al. (2017)'s work. The BLEU (Papineni et al., 2002) score with multi-bleu.pl was calculated during the evaluation.

In Table 1, we report the performance of our full model, the baseline, and existing systems. Our baseline model obtains similar results to Vaswani et al. (2017), the existing strong model on these

[1]The results for EN–RO are evaluated on the dataset with diacritics removed in the reference text.

System	EN–DE	EN–FR	EN–RO	EN–RO + STD
Vaswani et al. (2017) (base)	27.30	38.10	-	-
Vaswani et al. (2017) (big)	28.40	41.00	-	-
Transformer (base)	27.35	38.44	33.22	36.68
+ **D2GPo**	**27.93++**	**39.23++**	**34.00+**	**37.11+**
Transformer (big)	28.51	41.05	33.45	37.55
+ **D2GPo**	**29.10+**	**41.77++**	**34.13+**	**37.92+**

Table 1: Comparison with the baseline and existing systems on the supervised translation tasks. Here, "++/+" after the BLEU score indicate that the proposed method was significantly better than the corresponding baseline Transformer (base or big) at significance levels p <0.01/0.05. "STD" represents synthetic training data from (Sennrich et al., 2016b).

tasks. The results indicate that our method obtained significant improvements over strong baselines in all language pairs. Our model has not only improved in the translation model of large-scale training sets but also achieved considerable improvement in small-scale training sets. Besides, please refer to Appendix A.8 and A.9 for analysis on low resource scenario and generation diversity.

5.3 UNSUPERVISED NMT

For unsupervised machine translation, we also used the three language pairs EN–DE, EN–FR, and EN–RO as our evaluation targets. Note that the evaluation performed on EN–DE uses *newstest2016* instead of *newstest2014* to keep the results comparable to other works, unlike supervised machine translation. We used the masked sequence to sequence pre-training (MASS) model (Song et al., 2019) as our baseline. Following the practice of Song et al. (2019), we pretrained our model with a masked sequence-to-sequence pre-training (MASS) objective (without D2GPo) on EN, FR, DE, and RO monolingual data samples from the WMT 2007–2018 News Crawl datasets which cover 190M, 60M, 270M, and 10M sentences, respectively. Then, we fine-tuned the models on the same monolingual data using back-translation cross-entropy loss (Lample et al., 2018) and our D2GPo loss. For the training dataset, we filtered out sentences over 175 words long and also jointly learned 60K BPE sub-word units for every language pair.

Method	EN–FR	FR–EN	EN–DE	DE–EN	EN–RO	RO–EN
Artetxe et al. (2017)	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	25.14	24.18	17.16	21.00	21.18	19.44
XLM (Lample & Conneau, 2019)	33.40	33.30	27.00	34.30	33.30	31.80
MASS (Song et al., 2019)	37.50	34.90	28.30	35.20	35.20	33.10
MASS + **D2GPo**	**37.92**	**34.94**	**28.42**	**35.62**	**36.31**	**33.41**

Table 2: BLEU score comparisons between MASS and previous methods on unsupervised NMT.

As shown in Table 2, D2GPo achieved consistent improvement over MASS (the state-of-the-art baseline) on all the unsupervised translation pairs. While MASS, XLM, etc. systems leverage large scale monolingual pre-training, the decoder (generator, LM) can still be improved by our D2GPo loss in the fine-tuning phase. This indicates the efficiency of the proposed method.

5.4 TEXT SUMMARIZATION

Text summarization is a typical language generation task which creates a short and fluent summary of the given long text document. Song et al. (2019) fine-tuned the MASS pretrained model on the text summarization task and achieved the state-of-the-art results. We chose this model as our baseline, keeping the pre-training consistent with it, and using D2GPo loss for enhancements in the fine-tuning phase. The Annotated Gigaword corpus is used as the benchmark, which is detailed in Appendix A.4. During the evaluation, ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) are reported.

	Model	ROUGE-1	ROUGE-2	ROUGE-L
Supervised	RNN-based seq2seq	35.50	15.54	32.45
	Nallapati et al. (2016)	34.97	17.17	32.70
Semi-supervised	MLM pre-training (Song et al., 2019)	37.75	18.45	34.85
	DAE pre-training (Song et al., 2019)	35.97	17.17	33.14
	MASS pre-training (Song et al., 2019)	38.73	19.71	35.96
	MASS + **D2GPo**	**39.23**	**20.11**	**36.48**

Table 3: Performance on the text summarization task

Our results on text summarization is listed in Table 3. We compared our +D2GPo with our baseline MASS which is the current state-of-the-art model; it consistently outperformed the baseline on all evaluation metrics. The models with a semi-supervised setting yielded a large-margin improvement compared to the model without any pre-training, which demonstrates that the supervised pre-training is effective for the text summarization task.

5.5 STORYTELLING

Storytelling is at the frontier of current language generation technologies: stories must maintain a consistent theme throughout the document and require very long-distance dependency modeling. Additionally, stories require creativity and a high-level plot with planning ahead rather than word-by-word generation (Wiseman et al., 2017).

We used the hierarchical story generation model (Fan et al., 2018) (which is introduced in Appendix A.5) as our baseline to test the improvements of D2GPo over the storytelling task. In order to guarantee the single variable principle, we only added the D2GPo loss to the story generation model. The prompt generation model is consistent with Fan et al. (2018).

Model	Params	Valid Perplexity	Test Perplexity
LSTM seq2seq	110.3 M	46.83	46.79
Conv seq2seq	113.0 M	45.27	45.54
Conv seq2seq + self-attention	134.7 M	37.37	37.94
Ensemble: Conv seq2seq + self-attention	270.3 M	36.63	36.93
Fusion: Conv seq2seq + self-attention	255.4 M	36.08	36.56
Conv seq2seq + self-attention + **D2GPo**	134.7 M	**35.56**	**35.74**
Fusion: Conv seq2seq + self-attention + **D2GPo**	255.4 M	**33.82**	**33.90**

Table 4: Perplexity on WRITINGPROMPTS.

For automatic evaluation, we measured the model perplexity on the valid and test set. Table 4 shows the effect of the D2GPo. Results show that adding our D2GPo, Conv seq2seq + self-attention model substantially improved the likelihood of human-generated stories and even outperformed the ensemble or fusion models without increasing the parameters. With the fusion mechanism added, the perplexity was further reduced. These results suggest that the D2GPo can improve the quality of language generation greatly, especially in settings where there are fewer restrictions on such story generation tasks.

5.6 IMAGE CAPTION

Image caption is a task which combines image understanding and language generation. It continues to inspire considerable research at the boundary of computer vision and natural language processing. In order to verify the performance of D2GPo on the language generation model of diverse types of input, we elected to experiment with image captioning.

In our experiments, we evaluated our model on an ablated baseline (Top-down, detailed in Appendix A.6) (Anderson et al., 2018) against prior work on MSCOCO 2014 captions dataset (Lin et al., 2014), which has became the standard benchmark for image caption. For validation of model hyperparame-

ters and offline testing, we used the 'Karpathy' splits (Karpathy & Fei-Fei, 2015) that have been used extensively for reporting results in prior work. SPICE (Anderson et al., 2016), CIDEr (Vedantam et al., 2015), METEOR (Denkowski & Lavie, 2014), ROUGE-L, and BLEU were used to evaluate the caption quality.

	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Att2in (Rennie et al., 2017)	-	31.3	26.0	54.3	101.3	-
Att2all (Rennie et al., 2017)	-	30.0	25.9	53.4	99.4	-
Baseline: Top-down	74.5	33.4	26.1	54.4	105.4	19.2
Baseline + **D2GPo**	**75.2**	**33.6**	**26.3**	**55.1**	**106.6**	**19.7**
Baseline + SCST	77.8	34.4	26.6	56.1	114.3	19.9
Baseline + SCST + **D2GPo**	**78.0**	**34.7**	**26.8**	**56.3**	**116.8**	**20.2**

Table 5: Image caption performance on the MSCOCO Karpathy test split.

In Table 5, we report the performance of our full model and the ResNet Top-down baseline in comparison to the existing strong Self-critical Sequence Training (SCST) (Rennie et al., 2017) approach on the test portion of the Karpathy splits. For a fair comparison, results are only reported for models trained with standard cross-entropy loss (MLE). All results are reported for a single model with no fine-tuning of the input ResNet model. Relative to the SCST models, our ResNet baseline obtained slightly better performance. After incorporating our proposed D2GPo loss, our model shows further improvements across all metrics.

6 Evaluation Function

According to the analysis in Section 4, for the embedding, we used the Gaussian PDF as our evaluation function $f(\cdot)$; however, to evaluate the effectiveness of different evaluation functions, we changed it and tested the performance changes on supervised NMT EN-DE task. We followed the same experiment settings as described in Section 5.2 and compare the BLEU score changes on the test set, as listed in Table 6.

Evaluation Function	BLEU	$\triangle$
Baseline	27.35	
Gaussian	27.93	0.58 $\uparrow$
Random	26.34	1.01 $\downarrow$
Linear	27.45	0.10 $\uparrow$
Cosine	27.62	0.27 $\uparrow$

Table 6: Ablation study on our proposed D2GPo with different evaluation function on supervised NMT WMT14 EN-DE task, Transformer base model.

We observe that the performance of Gaussian density, linear, and cosine functions increased, while the random one decreased. This shows that the distance information obtained from embedding can effectively guide the generation process. Among these functions, the Gaussian density function obtained the greatest improvement, which agrees with our analysis of the embedding features obeying the Gaussian distribution. For the linear and cosine functions, we postulate that because these two functions are a rough approximation of the Gaussian density function, and they, therefore, function as well as Gaussian.

7 Conclusion

This work proposes data-dependent Gaussian prior objective (D2GPo) for language generation tasks with the hope of alleviating the difficulty of the *negative diversity ignorance*. D2GPo imposes the prior from (linguistic) data over the sequence prediction models. Through experiments on classic language generation tasks, i.e., neural machine translation, text summarization, storytelling, and image caption tasks, D2GPo achieved significant improvement over strong baselines.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398. Springer, 2016.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

Shiqi Shen Ayana, Zhiyuan Liu, and Maosong Sun. Neural headline generation with minimum risk training. *arXiv preprint arXiv:1604.01904*, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Kedar Bellare, Gregory Druck, and Andrew McCallum. Alternating projections for learning with expectation constraints. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 43–50. AUAI Press, 2009.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pp. 1171–1179, 2015.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 933–941. JMLR. org, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 376–380, 2014.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, 2018.

Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(Jul):2001–2049, 2010.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pp. 1243–1252, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2410–2420, 2016.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Controllable text generation. *arXiv preprint arXiv:1703.00955*, 4, 2017.

Zhiting Hu, Zichao Yang, Ruslan R Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, and Eric P Xing. Deep generative models with learnable knowledge constraints. In *Advances in Neural Information Processing Systems*, pp. 10501–10512, 2018.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709, 2013.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Matt J Kusner and José Miguel Hernández-Lobato. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*, 2016.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, 2018.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 29–37, 2011.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, 2016a.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, 2016b.

Zhifei Li and Jason Eisner. First-and second-order expectation semirings with applications to minimum-risk training on translation forests. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 40–51. Association for Computational Linguistics, 2009.

Percy Liang, Michael I Jordan, and Dan Klein. Learning from measurements in exponential families. In *Proceedings of the 26th annual international conference on machine learning*, pp. 641–648. ACM, 2009.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3362–3371, 2017.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P Xing. Symbolic graph reasoning meets convolutions. In *Advances in Neural Information Processing Systems*, pp. 1853–1863, 2018.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, 2016.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pp. 95–100. Association for Computational Linguistics, 2012.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Ehud Reiter and Robert Dale. *Building natural language generation systems*. Cambridge university press, 2000.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008–7024, 2017.

Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, 2015.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 371–376, 2016a.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016b.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Stuart C Shapiro. *Encyclopedia of artificial intelligence second edition*. John, 1992.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1683–1692, 2016.

David A Smith and Jason Eisner. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 787–794. Association for Computational Linguistics, 2006.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pp. 5926–5936, 2019.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 196–205, 2015.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. In *Advances in neural information processing systems*, pp. 25–32, 2004.

Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pp. 1747–1756, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. 2019.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2253–2263, 2017.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46–55, 2018.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Huan Zhang and Hai Zhao. Minimum divergence vs. maximum margin: an empirical comparison on seq2seq models. In *ICLR*, 2018.

Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–680, 2014.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 654–664, 2017.

A APPENDIX

A.1 CONCEPTS UNDERLYING D2GPO

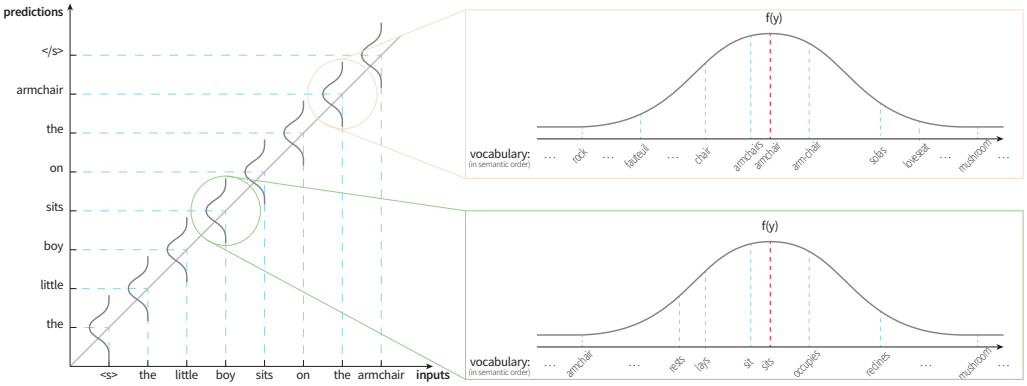

Figure 1: Overview of the concepts underlying D2GPo with the example of sentence *the little boy sits on the armchair*.

A.2 TOPOLOGICAL ORDER

Specifically, for target y_i, we calculate the embedding cosine similarity as the distance $dist(i, j)$ of y_i and all other token types in the vocabulary $\tilde{y}_j$, which are used to show the distance as follows:

$$dist_{i,j} = cosine_similarity(emb(y_i), emb(\tilde{y}_j)). \tag{8}$$

Sorting by distance from small to large to obtain a topological order for each token types yields

$$\textit{ORDER}(y_i) = sort([dist_{i,1}, dist_{i,2}, ..., dist_{i,N}]). \tag{9}$$

A.3 SUPERVISED NMT DATA

For the EN–DE translation task, 4.43M bilingual sentence pairs from the WMT'14 dataset, which includes the Common Crawl, News Commentary, and Europarl v7 datasets, were used as training data. The *newstest2013* and *newstest2014* datasets were used as the dev set and test set, respectively.

For the EN–FR translation task, 36M bilingual sentence pairs from the WMT'14 dataset were used as training data. The *newstest2012* and *newstest2013* datasets were combined for validation and *newstest2014* was used as the test set, following the configuration of Gehring et al. (2017).

For the EN–RO task, we tested two settings, one uses only the officially provided parallel corpus: Europarl v7 and SETIMES2, which yields 600K sentence pairs for a low-resource supervised machine translation study. Alternatively, following the work of Sennrich et al. (2016a), we used the synthetic training data (STD) provided by Sennrich et al. (2016a), which obtains 2.8M sentence pairs for training. We used *newsdev2016* as the dev set and *newstest2016* as the test set. The results on EN-RO we reported is evaluated on the reference which removed the diacritics from the Romanian.

A.4 TEXT SUMMARIZATION DATA

The Annotated Gigaword corpus (Napoles et al., 2012) was used as the benchmark (Rush et al., 2015). This data set is derived from news articles and consists of pairs of the main sentences in the article (longer), and the headline (shorter). The article and the headline were used as the source input sentence and reference, respectively. The data includes approximately 3.8M training samples, 400K validation samples, and 2K test samples.

A.5 HIERARCHICAL STORY GENERATION MODEL

Hierarchical story generation model (Fan et al., 2018) was proposed to tackle the challenges which first generates a sentence called *prompt* describing the theme topic for the upcoming story generation, and then conditions on the *prompt* when generating the story. Specifically, Fan et al. (2018) used a self-attention gated convolutional language model (GCNN) (Dauphin et al., 2017) as the sequence-to-sequence *prompt* generation model with the top-k random sampling. For the *prompt*-to-story generation, they collected a dataset from Reddit's WRITINGPROMPTS forum in which each *prompt* have multiple story responses. With the dataset, they trained a story generation model which gain further improvements with a novel form of model fusion that improved the relevance of the story to the prompt and adding a new gated multi-scale self-attention mechanism to model the long-range context.

A.6 TOP-DOWN IMAGE CAPTION MODEL

Top-down image caption model uses a ResNet (He et al., 2016) CNN pretrained on ImageNet (Deng et al., 2009) to encode each image. Similarly to previous work (Rennie et al., 2017), they encoded the full-sized input image with the final convolutional layer of Resnet-101 and used bilinear interpolation to resize the output to a fixed size spatial representation of 10×10. This is equivalent to the maximum number of spatial regions used in our full model.

A.7 HYPERPARAMETERS IN D2GPO

During training with our D2GPo, the value of the standard deviation of the KL diversity item λ was set to 0.1, and the softmax temperature was $T = 2.0$ in all experiments.

In order to study the effect of hyperparameters, i.e., the standard deviation λ and softmax temperature T in D2GPo, on the experimental results, we carried out experiments on WMT14 EN-DE with Transformer-base model as baseline [2], and set λ as $[0, 0.1, 0.2, 0.5, 1.0]$, T as $[1.0, 2.0, 5.0, 10.0]$.

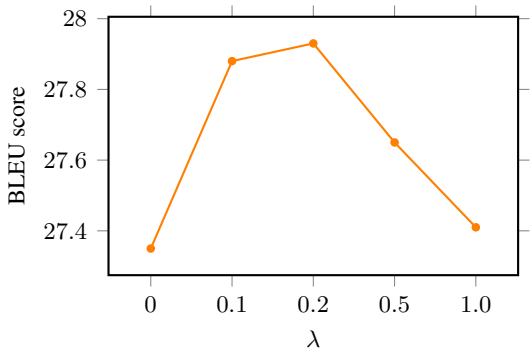

Figure 3: Performances on WMT14 EN-DE with different λ values.

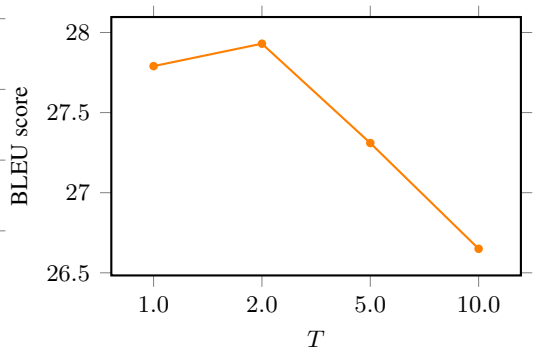

Figure 4: Performances on WMT14 EN-DE with different T values.

[2]Due to the limited experimental resources and time, we only consider the situation that λ and T change separately, that is, when λ changes, T remains unchanged at 2.0, and when T changes, λ remains unchanged at 0.1.

It could be seen from the experimental results that λ has an impact on the model training process. We guess the reason is that: a small λ makes the model unable to make full use of the prior knowledge (distribution), while a larger λ will make the model more uncertain because of the probability of incorrect or even opposite words whose fastText embeddings are similar is increased.

In addition, from the experiment results of T, when T is small, it can improve the model to some extent, but a large T will seriously decrease the performance of the model. Theoretically, when T approaches infinity, distribution q becomes uniform distribution, and there is no prior knowledge to guide the model. The loss penalty is applied to any model prediction, so excessive T is harmful to training.

A.8 D2GPo under Low-Resource Setting

Priors are generally more helpful in low-data regimes. To verify the effectiveness of D2GPo under low-resource setting, we respectively sampled 10K, 100K, and 600K paired sentence from the bilingual training data of WMT16 EN-RO, to explore the performance of our method in different low-resource scenarios. We used the same BPE codes, and fastText embeddings learned in the all WMT16 EN-RO training data.

Method	10K	100K	600K
Baseline	1.01	17.80	33.22
+ **D2GPo**	4.33	20.48	34.00

Table 7: The comparisons between our baseline and our D2GPo method under different training data scale in terms of BLEU on WMT16 EN-RO test set.

As shown in Table 7, D2GPo outperforms the baseline model demonstrating the effectiveness of our method in the low-resource scenarios. At the same time, it can be found from the results that the less the training data, the higher the improvement of D2GPo obtained. It shows that prior knowledge can substantially improve the performance of the model when the training data is scarce. The possible reason is that too little training data is not enough to train a robust model. In this case, the injection of prior knowledge can help to train the parameters of the model and substantially improve the translation performance. However, with the increase of training data, the model itself can be optimized well, and the improvement by introducing prior knowledge is not as substantial as before.

A.9 Analysis on Generation Diversity

Compared with traditional MLE training, D2GPo encourages negative diversity. In order to compare the differences between D2GPo and MLE models, we counted the high- and low-frequency words based on the training set and compared the frequency of low-frequency words prediction of the two models together with the golden reference on the test set.

	#GOLD	Baseline	**+D2GPo**
#LF	4915	3900	3998
#SUM	63086	55234	56129
#RATIO	7.79%	7.06%	7.12%

Table 8: The statics of low frequency words in the reference and generations.

The experiment was carried out on WMT14 EN-DE, and the baseline model is Transformer-base, and the statistics were based on the word level. We chose the words with a frequency less than or equal to 100 in the training set as low-frequency words. We used the golden reference (#GOLD), baseline model prediction output, and +D2GPo model prediction output to count the total number of tokens (#SUM) and the number of low-frequency words (#LF). The results are shown in Table 8.

The results show that compared with baseline, D2GPo optimized model can generate more low-frequency words, and the ratio of low-frequency words has been increased, but it is still far less than the golden reference. It is demonstrated that D2GPo can increase the diversity of model output.

A.10 EXAMPLES OF STORYTELLING

Example Prompt: The scientists have discovered something terrible .
Generation with Conv seq2seq + Self-Attention (PPL 37.37): The discovery of the discovery was a long and boring process . It was a small , ⟨unk⟩ , and a simple and elegant solution . The scientists , scientists in charge , and the scientific community , was able to decipher the meaning of the universe . The theory that the universe was not a very real one , was that the universe itself was not a planet , but a very , very real planet . The discovery of a particular object in the universe that the universe was not . We found something that resembled something that could not be described , but something that could not be described . We found a way to make the object a lot more specific . We found that it was a very , very large thing , a small , very small and very large . It was the first thing we noticed that we found that it was a massive , large , enormous , enormous , giant , giant . We found a way to make the whole thing. We found that the entire
Generation with Conv seq2seq + Self-Attention + D2GPo (PPL 35.74): " I don 't understand . Why did you tell us about it ? " The scientist 's assistant , Dr. harry , looked up from his tablet and back at the scientist . " It 's ... you see. " He looked back up at Dr. harry . " What 's it mean ? " " It means that the experiment is on the cusp of time . It 's not the same as it looks . There 's no explanation . It 's just that. . . " " How is it possible that the experiment is going to fail ? " The scientist looked ut the window and then back at Dr. harry , who was still looking out the window , at the vast blackness of the room . " You think they 'll do it ? " " Well , no , we can 't be sure , Dr. harry . They 'll do it. " Dr. harry turned to the scientist in front of him . " We can 't risk losing our best friend in the world
Generation with Fusion of Conv seq2seq + Self-Attention (PPL 36.08): We thought that we should be able to see our entire species . The scientists were n't sure what we would find . But they had to be prepared for it . They could n't be ready . The scientists had never seen anything like this , and yet they had n't seen anything like this ⟨unk⟩ in months . The scientists were just about to be in their own business , when they saw the results . They were n't ready for anything . We thought that we should take a deep breath before we go . But it was n't . There was n't any sign of any other intelligent life in the universe . It was nothing . The scientists had no idea what was going on . They just watched , waiting for a signal . They had no idea what it would mean . But they could n't see us . The scientists just knew . They were all afraid of what they would see . They could n't see the future . They could n't see the future .
Generation with Fusion of Conv seq2seq + Self-Attention + D2GPo (PPL 33.82): " I 'm afraid I wo n't be able to find out why my experiment is working . " "Well , we 've been working on the project for about a month now . " " It 's been a month and a half since I last saw it . " " We 're all looking at the results . " " You 've already been working on it for months now . You think we 've found that ? " " I do n't know, but we do have a lot of research to do . " " But it 's not like it was working , is it ? " " We do n't know . We 're not looking for a breakthrough , it 's just an experiment . " " It 's just an experiment ? People will die and the world may be destroyed . The disaster is about to happen, we have to act." " What do you mean , it 'll not . It 's just an experiment . " " No , no , no , it is something terrible we cannot ignore ."

Table 9: Example stories generated by the baselines and our full models.

A.11 EXAMPLES OF IMAGE CAPTION

	Top-down: a woman holding an umbrella in her hand **+ D2GPo:** a woman is holding an umbrella **+ SCST:** a woman holding an umbrella in a street **+ SCST+ D2GPo:** a woman is holding an umbrella in the street
	Top-down: a large airplane sitting on top of an airport runway **+ D2GPo:** an airplane is sitting on top of an airport runway **+ SCST:** a large jetliner sitting on top of an airport runway **+ SCST+ D2GPo:** a large jetliner is sitting on top of an airport runway
	Top-down: a woman holding a surf board in the ocean **+ D2GPo:** a woman is standing on the beach with a surfboard **+ SCST:** a woman holding a surfboard on the beach **+ SCST+ D2GPo:** a woman is standing on the beach with a surfboard
	Top-down: a traffic light with a traffic light on it **+ D2GPo:** a traffic light on the side of a traffic light **+ SCST:** a yellow traffic light on the side of a street **+ SCST+ D2GPo:** yellow traffic lights on the side of a street

Table 10: Captions generated for the left image by the various models described in the paper. The models trained with SCST return a more accurate and more detailed summary of the image. The models trained with D2GPo return a more grammatically complete sentence.