Improving Goal-Oriented Visual Dialog Agents via Advanced Recurrent Nets with Tempered Policy Gradient

Rui Zhao and **Volker Tresp**
Siemens AG, Corporate Technology, Munich, Germany
Ludwig-Maximilians-Universität München, Munich, Germany
{ruizhao, volker.tresp}@siemens.com

Abstract

Learning goal-oriented dialogues by means of deep reinforcement learning has recently become a popular research topic. However, training text-generating agents *efficiently* is still a considerable challenge. Commonly used policy-based dialogue agents often end up focusing on simple utterances and suboptimal policies. To mitigate this problem, we propose a class of novel temperature-based extensions for policy gradient methods, which are referred to as Tempered Policy Gradients (TPGs). These methods encourage exploration with different temperature control strategies. We derive three variations of the TPGs and show their superior performance on a recently published AI-testbed, i.e., the GuessWhat?! game. On the testbed, we achieve significant improvements with two innovations. The first one is an extension of the state-of-the-art solutions with Seq2Seq and Memory Network structures that leads to an improvement of 9%. The second one is the application of our newly developed TPG methods, which improves the performance additionally by around 5% and, even more importantly, helps produce more convincing utterances. TPG can easily be applied to any goal-oriented dialogue systems.

1 Introduction

In recent years, deep learning has shown convincing performance in various areas such as image recognition, speech recognition, and natural language processing (NLP). Deep neural nets are capable of learning complex dependencies from huge amounts of data and its human generated annotations in a supervised way. In contrast, reinforcement learning agents [Sutton and Barto, 1998] can learn directly from their interactions with the environment without any supervision and surpass human performance in several domains, for instance in the game of GO [Silver *et al.*, 2016], as well as many computer games [Mnih *et al.*, 2015]. In this paper we are concerned with the application of both approaches to goal-oriented dialogue systems [Bordes and Weston, 2017; de Vries *et al.*, 2017; Das *et al.*, 2017; Strub *et al.*, 2017; Das *et al.*, 2017; Lewis *et al.*, 2017; Dhingra *et al.*, 2016], a problem that has recently caught the attention of machine learning researchers. De Vries et al. [2017] have proposed as AI-testbed a visual grounded object guessing game called GuessWhat?!. Das *et al.* [2017] formulated a visual dialogue system which is about two chatbots asking and answering questions to identify a specific image within a group of images. More practically, dialogue agents have been applied to negotiate a deal [Lewis *et al.*, 2017] and access certain information from knowledge bases [Dhingra *et al.*, 2016]. The essential idea in these systems is to train different dialogue agents to accomplish the tasks. In those papers, the agents have been trained with policy gradients, i.e. REINFORCE [Williams, 1992].

In order to improve the exploration quality of policy gradients, we present three instances of temperature-based methods. The first one is a single-temperature approach which is very easy to apply. The second one is a parallel approach with multiple temperature policies running concurrently. This second approach is more demanding on computational resources, but results in more stable solutions. The third one is a temperature policy approach that dynamically adjusts the temperature for each action at each time-step, based on action frequencies. This dynamic method is more sophisticated and proves more efficient in the experiments. In the experiments, all these methods demonstrate better exploration strategies in comparison to the plain policy gradient.

We demonstrate our approaches using a real-world dataset called GuessWhat?!. The GuessWhat?! game [de Vries *et al.*, 2017] is a visual object discovery game between two players, the Oracle and the Questioner. The Questioner tries to identify an object by asking the Oracle questions. The original works [de Vries *et al.*, 2017; Strub *et al.*, 2017] first proposed supervised learning to simulate and optimize the game. Strub *et al.* [2017] showed that the performance could be improved by applying plain policy gradient reinforcement learning, which maximizes the game success rate, as a second processing step. Building on these previous works, we propose two network architecture extensions. We utilize a Seq2Seq model [Sutskever *et al.*, 2014] to process the image along with the historical dialogues for question generation. For the guessing task, we develop a Memory Network [Sukhbaatar *et al.*, 2015] with Attention Mechanism [Bahdanau *et al.*, 2014] to process the generated question-answer pairs. We first train these two models using the plain policy gradient and use them

as our baselines. Subsequently, we train the models with our new TPG methods and compare the performances with the baselines. We show that the TPG method is compatible with state-of-the-art architectures such as Seq2Seq and Memory Networks and contributes orthogonally to these advanced neural architectures. To the best of our knowledge, the presented work is the first to propose temperature-based policy gradient methods to leverage exploration and exploitation in the field of goal-oriented dialogue systems. We demonstrate the superior performance of our TPG methods by applying it to the GuessWhat?! game. Our contributions are:

- We introduce Tempered Policy Gradients in the context of goal-oriented dialogue systems, a novel class of approaches to temperature control to better leverage exploration and exploitation during training.
- We extend the state-of-the-art solutions for the GuessWhat?! game by integrating Seq2Seq and Memory Networks. We show that TPGs are compatible with these advanced models and further improve the performance.

2 Preliminaries

In our notation, we use $\mathbf{x}$ to denote the input to a policy network π, and x_i to denote the i-th element of the input vector. Similarly, $\mathbf{w}$ denotes the weight vector of π, and w_i denotes the i-th element of the weight vector of that π. The output y is a multinoulli random variable with N states that follows a probability mass function, $f(y = n \mid \pi(\mathbf{x} \mid \mathbf{w}))$, where $\Sigma_{n=1}^{N} f(y = n \mid \pi(\mathbf{x} \mid \mathbf{w})) = 1$ and $f(\cdot) \geq 0$. In a nutshell, a policy network parametrizes a probabilistic unit that produces the sampled output, mathematically, $y \sim f(\pi(\mathbf{x} \mid \mathbf{w}))$.

At this point, we have defined the policy neural net and now discuss performance measures commonly used for optimizations. Typically, the expected value of the accumulated reward, i.e. return, conditioned on the policy network parameters $E(r \mid \mathbf{w})$ is used. Here, E denotes the expectation operator, r the accumulated reward signal, and $\mathbf{w}$ the network weight vector. The objective of reinforcement learning is to update the weights in a way that maximizes the expected return at each trial. In particular, the REINFORCE updating rule is: $\Delta w_i = \alpha_i(r - b_i)e_i$, where Δw_i denotes the weight adjustment of weight w_i, α_i is a non-negative learning rate factor, and b_i is a reinforcement baseline. The e_i is the *characteristic eligibility* of w_i, defined as $e_i = (\partial f/\partial w_i)/f = \partial \ln f/\partial w_i$. Williams [1992] has proved that the updating quantity, $(r - b_i)\partial \ln f/\partial w_i$, represents an unbiased estimate of $\partial E(r \mid \mathbf{w})/\partial w_i$.

3 Tempered Policy Gradient

In order to improve the exploration quality of REINFORCE in the task of optimizing policy-based dialogue agents, we attempt to find the optimal compromise between exploration and exploitation. In TPGs we introduce a parameter τ, the sampling temperature of the probabilistic output unit, which allows us to explicitly control the strengths of the exploration.

3.1 Exploration and Exploitation

The trade-off between exploration and exploitation is one of the great challenges in reinforcement learning [Sutton and Barto, 1998]. To obtain a high reward, an agent must exploit the actions that have already proved effective in getting more rewards. However, to discover such actions, the agent must try actions, which appear suboptimal, to explore the action space. In a stochastic task like text generation, each action, i.e. a word, must be tried many times to find out whether it is a reliable choice or not. The exploration-exploitation dilemma has been intensively studied over many decades [Carmel and Markovitch, 1999; Nachum *et al.*, 2016; Liu *et al.*, 2017]. Finding the balance between exploration and exploitation is considered crucial for the success of reinforcement learning [Thrun, 1992].

3.2 Temperature Sampling

In text generation, it is well-known that the simple trick of temperature adjustment is sufficient to shift the language model to be more conservative or more diversified [Karpathy and Fei-Fei, 2015]. In order to control the trade-off between exploration and exploitation, we borrow the strength of the temperature parameter $\tau \geq 0$ to control the sampling. The output probability of each word is transformed by a temperature function as:

$$f^{\tau}(y = n \mid \pi(\mathbf{x} \mid \mathbf{w})) = \frac{f(y = n \mid \pi(\mathbf{x} \mid \mathbf{w}))^{\frac{1}{\tau}}}{\Sigma_{m=1}^{N} f(y = m \mid \pi(\mathbf{x} \mid \mathbf{w}))^{\frac{1}{\tau}}}.$$

We use notation f^{τ} to denote a probability mass function f that is transferred by a temperature function with temperature τ. When the temperature is high, $\tau > 1$, the distribution becomes more uniform; when the temperature is low, $\tau < 1$, the distribution appears more spiky. TPG is defined as an extended algorithm of the Monte Carlo Policy Gradient approach. We use a higher temperature, $\tau > 1$, to encourage the model to explore in the action space, and conversely, a lower temperature, $\tau < 1$, to encourage exploitation. In the extreme case, when $\tau = 0$, we obtain greedy search.

3.3 Tempered Policy Gradient Methods

Here, we introduce three instances of TPGs in the domain of goal-oriented dialogues, including single, parallel, and dynamic tempered policy gradient methods.

Single-TPG: The Single-TPG method simply uses a global temperature τ_{global} during the whole training process, i.e., we use $\tau_{global} > 1$ to encourage exploration. The forward pass is represented mathematically as: $y^{\tau_{global}} \sim f^{\tau_{global}}(\pi(\mathbf{x} \mid \mathbf{w}))$, where $\pi(\mathbf{x} \mid \mathbf{w})$ represents a policy neural network that parametrizes a distribution $f^{\tau_{global}}$ over the vocabulary, and $y^{\tau_{global}}$ means the word sampled from this tempered word distribution. After sampling, the weight of the neural net is updated using,

$$\Delta w_i = \alpha_i(r - b_i)\partial \ln f(y^{\tau_{global}} \mid \pi(\mathbf{x} \mid \mathbf{w}))/\partial w_i.$$

Noteworthy is that the actual gradient, $\partial \ln f(y^{\tau_{global}} \mid \pi(\mathbf{x} \mid \mathbf{w}))/\partial w_i$, depends on the sampled word, $y^{\tau_{global}}$, however, does not depend directly on the temperature, τ. We prefer to find the words that lead to a reward, so that the model can learn quickly from these actions, otherwise, the neural network only learns to avoid current failure actions. With Single-TPG and $\tau > 1$, the entire vocabulary of a dialogue

agent is explored more efficiently than by REINFORCE, because nonpreferred words have a higher probability of being explored. This Single-TPG method is very easy to use and could yield a performance improvement after training because the goal-oriented dialogue optimization could benefit from increased exploration. The temperature is initialized with $\tau = 1$, then fine-tuned based on the learning curve on the validation sets, and subsequently left fixed..

Parallel-TPG: A more advanced version of Single-TPG is the Parallel-TPG that deploys several Single-TPGs concurrently with different temperatures, $\tau_1, ..., \tau_n$, and updates the weights based on all generated samples. During the forward pass, multiple copies of the neural nets parameterize multiple word distributions. The words are sampled in parallel at various temperatures, mathematically, $y^{\tau_1}, ..., y^{\tau_n} \sim f^{\tau_1,...,\tau_n}(\pi(\mathbf{x} \mid \mathbf{w}))$. After sampling, in the backward pass the weights are updated with the sum of gradients. The formula is given by

$$\Delta w_i = \Sigma_k \alpha_i (r - b_i) \partial \ln f(y^{\tau_k} \mid \pi(\mathbf{x} \mid \mathbf{w}))/\partial w_i,$$

where $k \in \{1, ..., n\}$. The combinational use of higher and lower temperatures ensures both exploration and exploitation at the same time. The sum over weight updates of parallel policies gives a more accurate Monte Carlo estimate of $\partial E(r \mid \mathbf{w})/\partial w_i$, due to the nature of Monte Carlo methods [Robert, 2004]. Thus, compared to Single-TPG, we would argue that Parallel-TPG is more robust and stable, although Parallel-TPG needs more computational power. However, these computations can be easily distributed in a parallel fashion using state-of-the-art graphics processing units.

Dynamic-TPG: As a third variant, we introduce the Dynamic-TPG, which is the most sophisticated approach in the current TPG family. The essential idea is that we use a heuristic function h to assign the temperature τ to the word distribution at each time step, t. The temperature is bounded in a predefined range $[\tau_{min}, \tau_{max}]$. The heuristic function we used here is based upon the term frequency inverse document frequency, *tf-idf* [Leskovec *et al.*, 2014]. In the context of goal-oriented dialogues, we use the counted number of each word as term frequency *tf* and the total number of generated dialogues during training as document frequency *df*. We use the word that has the highest probability to be sampled at current time-step, y_t^*, as the input to the heuristic function h. Here, y_t^* is the maximizer of the probability mass function f. Mathematically, it is defined as $y_t^* = \text{argmax}(f(\pi(\mathbf{x} \mid \mathbf{w})))$. We propose that *tf-idf*(y_t^*) approximates the concentration level of the distribution, which means that if the same word is always sampled from a distribution, then the distribution is very concentrated. Too much concentration prevents the model from exploration, so that a higher temperature is needed. In order to achieve this effect, the heuristic function is defined as

$$\begin{aligned}\tau_t^h &= h(\textit{tf-idf}(y_t^*)) \\ &= \tau_{min} + (\tau_{max} - \tau_{min}) \frac{\textit{tf-idf}(y_t^*) - \textit{tf-idf}_{min}}{\textit{tf-idf}_{max} - \textit{tf-idf}_{min}}.\end{aligned}$$

With this heuristic, words that occur very often are depressed by applying a higher temperature to those words, making them less likely to be selected in the near future. In the forward pass, a word is sampled using $y^{\tau_t^h} \sim f^{\tau_t^h}(\pi(\mathbf{x} \mid \mathbf{w}))$. In the backward pass, the weights are updated correspondingly, using

$$\Delta w_i = \alpha_i (r - b_i) \partial \ln f(y^{\tau_t^h} \mid \pi(\mathbf{x} \mid \mathbf{w}))/\partial w_i,$$

where τ_t^h is the temperature calculated from the heuristic function. Compared to Parallel-TPG, the advantage of Dynamic-TPG is that it assigns temperature more appropriately, without increasing the computational load.

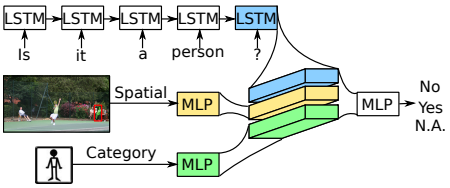


Figure 1: Oracle model

4 GuessWhat?! Game

We evaluate our concepts using a recent testbed for AI, called the GuessWhat?! game [de Vries *et al.*, 2017], available at `https://guesswhat.ai`. The dataset consists of 155 k dialogues, including 822 k question-answer pairs, each composed of around 5 k words, about 67 k images [Lin *et al.*, 2014] and 134 k objects. The game is about visual object discovery trough a multi-round QA among different players.

Formally, a GuessWhat?! game is represented by a tuple (I, D, O, o^*), where $I \in \mathbb{R}^{H \times W}$ denotes an image of height H and width W; D represents a dialogue composed of M rounds of question-answer pairs (QAs), $D = (\mathbf{q}_m, a_m)_{m=1}^M$; O stands for a list of K objects $O = (o_k)_{k=1}^K$; and o^* is the target object. Each question is a sequence of words, $\mathbf{q}_m = \{y_{m,1}, ..., y_{m,N_m}\}$ with length N_m. The words are taken from a defined vocabulary V, which consists of the words and a start token and an end token. Each answer is either yes, no, or not applicable, i.e. $a_m \in \{yes, no, n.a.\}$. For each object o_k, there is a corresponding object category $c_k \in \{1, ..., C\}$ and a pixel-wise segmentation mask $S_k \in \{0, 1\}^{H \times W}$. Finally, we use colon notation (:) to select a subset of a sequence, for instance, $(\mathbf{q}, a)_{1:m}$ refers to the first m rounds of QAs in a dialogue.

4.1 Models and Pretraining

Following [Strub *et al.*, 2017], we first train all three models in a supervised fashion.

Oracle: The task of the Oracle is to answer questions regarding to the target object. We outline here the simple neural network architecture that achieved the best performance in the study of [de Vries *et al.*, 2017], and which we also used in our experiments. The input information used here is of three modalities, namely the question $\mathbf{q}$, the spatial information $x_{spatial}^*$ and the category c^* of the target object. For encoding the question, de Vries et al. first use a lookup table to learn the embedding of words, then

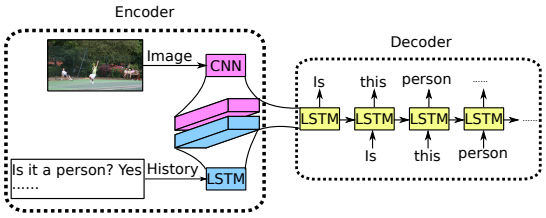


Figure 2: Question-Generator model

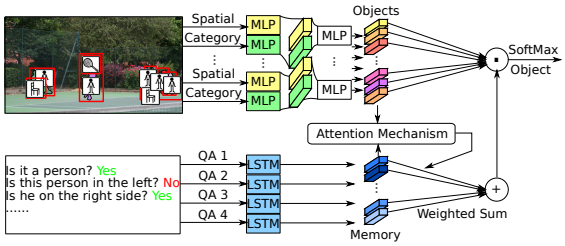


Figure 3: Guesser model

use a one layer long-short-term-memory (LSTM) [Hochreiter and Schmidhuber, 1997] to encode the whole question. For spatial information, de Vries et al. extract an 8-dimensional vector of the location of the bounding box $[x_{min},\ y_{min},\ x_{max},\ y_{max},\ x_{center},\ y_{center},\ w_{box},\ h_{box}]$, where x, y denote the coordinates and w_{box}, h_{box} denote the width and height of the bounding box, respectively. De Vries et al. normalize the image width and height so that the coordinates range from -1 to 1. The origin is at the image center. The category embedding of the object is also learned with a lookup table during training. At the last step, de Vries et al. concatenate all three embeddings into one feature vector and fed it into a one hidden layer multilayer perceptron (MLP). The softmax output layer predicts the distribution, $\text{Oracle} := p(a \mid \mathbf{q},\ c^*,\ x^*_{spatial})$, over the three classes, including no, yes, and not applicable. The model is trained using the negative log-likelihood criterion. The Oracle structure is shown in Fig. 1.

Question-Generator: The goal of the Question-Generator (QGen) is to ask the Oracle meaningful questions, $\mathbf{q}_{m+1}$, given the whole image, I, and the historical question-answer pairs, $(\mathbf{q}, a)_{1:m}$. In previous work [Strub *et al.*, 2017], the state transition function was modelled as an LSTM, which was trained using whole dialogues so that the model memorizes the historical QAs. We refer to this as dialogue level training. We develop a novel QGen architecture using a modified version of the Seq2Seq model [Sutskever *et al.*, 2014]. The modified Seq2Seq model enables *question level training*, which means that the model is fed with historical QAs, and then learns to produce a new question. Following [Strub *et al.*, 2017], we first encode the whole image into a fixed-size feature vector using the VGG-net [Simonyan and Zisserman, 2014]. The features come from the fc-8 layer of the VGG-net. For processing historical QAs, we use a lookup table to learn the word embeddings, then again use an LSTM encoder to encode the history information into a fixed-size latent representation, and concatenate it with the image representation:

$$\mathbf{s}^{enc}_{m,Nm} = \text{encoder}((\text{LSTM}(\mathbf{q}, a)_{1:m}), \text{VGG}(I)).$$

The encoder and decoder are coupled by initializing the decoder state with the last encoder state, mathematically, $\mathbf{s}^{dec}_{m+1,0} = \mathbf{s}^{enc}_{m,Nm}$. The LSTM decoder generates each word based on the concatenated representation and the previous generated word (note the first word is a start token):

$$y_{m+1,n} = \text{decoder}(\text{LSTM}((y_{m+1,n-1}, \mathbf{s}^{dec}_{m+1,n-1})).$$

The decoder shares the same lookup table weights as the encoder. The Seq2Seq model, consisting of the encoder and the decoder, is trained end-to-end to minimize the negative log-likelihood cost. During testing, the decoder gets a start token and the representation from the encoder, and then generates each word at each time step until it encounters a question mark token, $\text{QGen} := p(y_{m+1,n} \mid (\mathbf{q}, a)_{1:m},\ I)$. The output is a complete question. After several question-answer rounds, the QGen outputs an end-of-dialogue token, and stops asking questions. The overall structure of the QGen model is illustrated in Fig. 2.

Guesser: The goal of the Guesser model is to find out which object the Oracle model is referring to, given the complete history of the dialogue and a list of objects in the image, $p(o^* \mid (\mathbf{q}, a)_{1:M},\ x^O_{spatial},\ c^O)$. The Guesser model has access to the spatial, $x^O_{spatial}$, and category information, c^O, of the objects in the list. The task of the Guesser model is challenging because it needs to understand the dialogue and to focus on the important content, and then guess the object. To accomplish this task, we decided to integrate the Memory [Sukhbaatar *et al.*, 2015] and Attention [Bahdanau *et al.*, 2014] modules into the Guesser architecture used in the previous work [Strub *et al.*, 2017]. First, we use an LSTM header to process the varying lengths of question-answer pairs in parallel into multiple fixed-size vectors. Here, each QA-pair has been encoded into some facts, $\text{Fact}_m = \text{LSTM}((\mathbf{q}, a)_m)$, and stored into a memory base. Later, we use the sum of the spatial and category embeddings of all objects as a key, $\text{Key}_1 = \text{MLP}(x^O_{spatial},\ c^O)$, to query the memory and calculate an attention mask, $\text{Attention}_1(\text{Fact}_m) = \text{Fact}_m \odot \text{key}_1$, over each fact. Next, we use the sum of attended facts and the first key to calculate the second key. Further, we use the second key to query the memory base again to have a more accurate attention. These are the so called "two-hops" of attention in the literature [Sukhbaatar *et al.*, 2015]. Finally, we compare the attended facts with each object embedding in the list using a dot product. The most similar object to these facts is the prediction, $\text{Guesser} := p(o^* \mid (\mathbf{q}, a)_{1:M},\ x^O_{spatial},\ c^O)$. The intention of using the attention module here is to find out the most relevant descriptions or facts concerning the candidate objects. We train the whole Guesser network end-to-end using the negative log-likelihood criterion. A more graphical description of the Guesser model is shown in Fig. 3.

4.2 Reinforcement Learning

Now, we post-train the QGen and the Guesser model with reinforcement learning. We keep the Oracle model fixed. In each game episode, when the models find the correct object, $r = 1$, otherwise, $r = 0$.

Next, we can assign credits for each action of the QGen and the Guesser models. In the case of the QGen model, we spread the reward uniformly over the sequence of actions in the episode. The baseline function, b, used here is the running average of the game success rate. Consider that the Guesser model has only one action in each episode, i.e., taking the guess. If the Guesser finds the correct object, then it gets an immediate reward and the Guesser's parameters are updated using the REINFORCE rule without baseline. The QGen is trained using the following four methods.

REINFORCE: The baseline method used here is REINFORCE [Williams, 1992]. During training, in the forward pass the words are sampled with $\tau = 1$, $y_{m+1,n} \sim f(\text{QGen}(\mathbf{x} \mid \mathbf{w}))$. In the backward pass, the weights are updated using REINFORCE, as,

$$\mathbf{w} = \mathbf{w} + \alpha(r - b)\nabla_{\mathbf{w}} \ln f(y_{m+1,n} \mid \text{QGen}(\mathbf{x} \mid \mathbf{w})).$$

Single-TPG: We use temperature $\tau_{global} = 1.5$ during training to encourage exploration, mathematically, $y^{\tau_{global}}_{m+1,n} \sim f^{\tau_{global}}(\text{QGen}(\mathbf{x} \mid \mathbf{w}))$. In the backward pass, the weights are updated using

$$\mathbf{w} = \mathbf{w} + \alpha(r - b)\nabla_{\mathbf{w}} \ln f(y^{\tau_{global}}_{m+1,n} \mid \text{QGen}(\mathbf{x} \mid \mathbf{w})).$$

Parallel-TPG: For Parallel-TPG, we use two temperatures $\tau_1 = 1.0$ and $\tau_2 = 1.5$ to encourage the exploration. The words are sampled in the forward pass using $y^{\tau_1}_{m+1,n}$, $y^{\tau_2}_{m+1,n} \sim f^{\tau_1,\tau_2}(\text{QGen}(\mathbf{x} \mid \mathbf{w}))$. In the backward pass, the weights are updated using

$$\mathbf{w} = \mathbf{w} + \Sigma^2_{k=1}\alpha(r - b)\nabla_{\mathbf{w}} \ln f(y^{\tau_k}_{m+1,n} \mid \text{QGen}(\mathbf{x} \mid \mathbf{w})).$$

Dynamic-TPG: The last method we evaluated is Dynamic-TPG. We use a heuristic function to calculate the temperature for each word at each time step:

$$\tau^h_{m+1,n} = \tau_{min} + (\tau_{max} - \tau_{min})\frac{\textit{tf-idf}(y^*_{m+1,n}) - \textit{tf-idf}_{min}}{\textit{tf-idf}_{max} - \textit{tf-idf}_{min}},$$

where we set $\tau_{min} = 0.5$, $\tau_{max} = 1.5$, and set $\textit{tf-idf}_{min} = 0$, $\textit{tf-idf}_{max} = 8$. After the calculation of $\tau^h_{m+1,n}$, we substitute the value into the formula at each time step and sample the next word using

$$y^{\tau^h_{m+1,n}}_{m+1,n} \sim f^{\tau^h_{m+1,n}}(\text{QGen}(\mathbf{x} \mid \mathbf{w})).$$

In the backward pass, the weights are updated using

$$\mathbf{w} = \mathbf{w} + \alpha(r - b)\nabla_{\mathbf{w}} \ln f(y^{\tau^h_{m+1,n}}_{m+1,n} \mid \text{QGen}(\mathbf{x} \mid \mathbf{w})).$$

For all four methods, we use greedy search in evaluation.

5 Experiment

We first train all the networks in a supervised fashion, and then further optimize the QGen and the Guesser model using reinforcement learning. The source code is available at `https://github.com/ruizhaogit/GuessWhat-TemperedPolicyGradient`, which uses Torch7 [Collobert *et al.*, 2011].

#	Method	Accuracy
1	[Strub *et al.*, 2017]	52.30%
2	[Strub and de Vries, 2017]	60.30%
3	REINFORCE	**69.66%**
4	Single-TPG	69.76%
5	Parallel-TPG	73.86%
6	Dynamic-TPG	**74.31%**

Table 1: Performance comparison of our methods to other methods reported in literature after reinforcement learning

5.1 Pretraining

We train all three models using 0.5 dropout [Srivastava *et al.*, 2014] during training, using the ADAM optimizer [Kingma and Ba, 2014]. We use a learning rate of 0.0001 for the Oracle model and the Guesser model, and a learning rate of 0.001 for QGen. All the models are trained with at most 30 epochs and early stopped within five epochs without improvement on the validation set. We report the performance on the test set which consists of images not used in training. We report the game success rate as the performance metric for all three models, which equals to the number of succeeded games divided by the total number of all games. Compared to previous works [de Vries *et al.*, 2017; Strub *et al.*, 2017; Strub and de Vries, 2017], after supervised training, our models obtain a game success rate of 48.77%, that is 4% higher than state-of-the-art methods [Strub and de Vries, 2017], which has 44.6% accuracy.

5.2 Reinforcement Learning

We first initialize all models with pre-trained parameters from supervised learning and then post-train the QGen using either REINFORCE or TPG for 80 epochs. We update the parameters using stochastic gradient descent (SGD) with a learning rate of 0.001 and a batch size of 64. In each epoch, we sample each image in the training set once and randomly pick one of the objects as a target. We track the running average of the game success rate and use it directly as the baseline, b, in REINFORCE. We limit the maximum number of questions to 8 and the maximum number of words to 12. Simultaneously, we train the Guesser model using REINFORCE without baseline and using SGD with a learning rate of 0.0001. The performance comparison between our baseline (# 3) with methods from literature (# 1 & # 2) is shown in Tab. 1.

REINFORCE Baseline: From Tab. 1, we see that our models trained with REINFORCE (# 3) are about 9% better than the state-of-the-art methods (# 1 & # 2). The improvements are due to using advanced mechanisms and techniques such as the Seq2Seq structure in the QGen, the memory and attention mechanisms in the Guesser, and the training of the Guesser model with reinforcement learning. One important difference is that our QGen model is trained in *question level*. This means that the model first learns to query meaningfully, step by step. Eventually, it learns to conduct a meaningful dialog. Compared to directly learning to manage a strategic conversation, this bottom-up training procedure helps the model absorb knowledge, because it breaks large tasks down into smaller, more manageable pieces. This makes the learn-

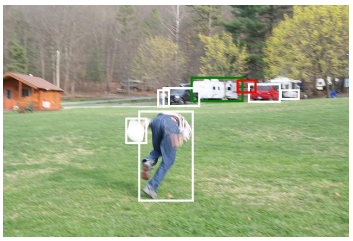
Image	Policy Gradient		Tempered Policy Gradient	
	Is it in left?	No	Is it a person?	No
	Is it in front?	No	Is it a **vehicle**?	Yes
	Is it in right?	Yes	Is it a **truck**?	Yes
	Is it in middle?	Yes	Is it in front of photo?	No
	Is it person?	No	In the left half?	No
	Is it ball?	No	In the middle of photo?	Yes
	Is it bat?	No	Is it to the right photo?	Yes
	Is it **car**?	Yes	Is it in the middle of photo?	Yes
	Status:	Failure	Status:	Success
	Is it in **left**?	No	Is it a giraffe?	Yes
	Is it in front?	Yes	In front of photo?	Yes
	Is it in right?	No	In the **left half**?	Yes
	Is it in middle?	Yes	Is it in the middle of photo?	Yes
	Is it person?	No	Is it **to the left** of photo?	Yes
	Is it giraffe?	Yes	Is it to the right photo?	No
	Is in middle?	Yes	**In the left** in photo?	No
	Is in middle?	Yes	In the middle of photo?	Yes
	Status:	Failure	Status:	Success

Table 2: Some samples generated by our improved models using REINFORCE (left column: "Policy Gradient") and Dynamic-TPG (right column: "Tempered Policy Gradient"). The green bounding boxes highlight the target objects; the red boxes highlight the wrong guesses.

ing for QGen much easier. In the remainder of the section, we use our models, boosted with memory network, attention, and Seq2Seq, trained with REINFORCE as a strong baseline and analyse the performance improvements achieved by the TPGs.

From Tab. 1, we see that compared to REINFORCE (#3), Single-TPG (#4) with $\tau_{global} = 1.5$ achieves a comparable performance. With two different temperatures $\tau_1 = 1.0$ and $\tau_2 = 1.5$, Parallel-TPG (#5) achieves an improvement of approximately 4%. Parallel-TPG requires more computational resources. Compared to Parallel-TPG, Dynamic-TPG only uses the same computational power as REINFORCE does and still gives a larger improvement by using a dynamic temperature, $\tau_t^h \in [0.5, 1.5]$. After comparison, we can see that the best model is Dynamic-TPG (#6), which gives a 4.65% improvement upon our strong baseline. Here, we have shown that our proposed methods contribute orthogonally, in the sense that they further improve the models already boosted with advanced modules such as memory network, attention, and Seq2Seq.

TPG Dialogue Samples: The generated dialogue samples in Tab. 2 can give some interesting insights in explaining why TPG methods give a better result. First of all, the sentences generated from TPG-trained models are on average longer and use slightly more complex structures, such as "Is it in the middle of photo?" instead of a simple form "Is it in middle?". Secondly, TPGs enable the models to explore better and comprehend more words. For example, in the first task (upper half of Tab. 2), both models ask about the category. The REINFORCE-trained model can only ask with the single word "car" to query about the vehicle category. In contrast, the TPG-trained model can first ask a *more general* category with the word "vehicle" and follows up querying with a *more specific* category "trucks". These two words "vehicle" and "trucks" give much more information than the single word "car", and help the Guesser model identify the truck among many cars. Lastly, similar to the category case, the models trained with TPG can first ask a *larger* spatial range of the object and follow up with a *smaller* range. In the second task (lower half of Tab. 2), we see that the TPG-trained model first asks "In the left half?", which refers to all the three giraffes in the left half, and the answer is "Yes". Then it asks "Is it to the left of photo?", which refers to the second left giraffe, and the answer is "Yes". Eventually the QGen asks "In the left in photo?", which refers to the most left giraffe, and the answer is "No". These specific questions about locations are not learned using REINFORCE. The REINFORCE-trained model can only ask a similar question with the word "left". In this task, there are many giraffes in the left part of the image. The top-down spatial questions help the Guesser model find the correct giraffe. To summarize, the TPG-trained models use longer and more informative sentences than the REINFORCE-trained models.

6 Conclusion

Our paper makes two contributions. Firstly, by extending existing models with Seq2Seq and Memory Networks we could improve the performance of a goal-oriented dialogue system by 9%. Secondly, we introduced TPG, a novel class of temperature-based policy gradient approaches. TPGs boosted the performance of the goal-oriented dialogue systems by another 4.7%. Among the three TPGs, Dynamic-TPG gave the best performance, which helped the agent comprehend more words, and produce more meaningful questions. TPG is a generic strategy to encourage word exploration on top of policy gradients and can be applied to any text-generating agents.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Bordes and Weston, 2017] Antoine Bordes and Jason Weston. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations (ICLR)*, 2017.

[Carmel and Markovitch, 1999] David Carmel and Shaul Markovitch. Exploration strategies for model-based learning in multi-agent systems: Exploration strategies. *Autonomous Agents and Multi-agent systems*, 2(2):141–172, 1999.

[Collobert *et al.*, 2011] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision (ICCV)*, 2017.

[de Vries *et al.*, 2017] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Dhingra *et al.*, 2016] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. End-to-end reinforcement learning of dialogue agents for information access. *arXiv preprint arXiv:1609.00777*, 2016.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[Karpathy and Fei-Fei, 2015] 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*, pages 3128–3137, 2015.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Leskovec *et al.*, 2014] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge university press, 2014.

[Lewis *et al.*, 2017] Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

[Lin *et al.*, 2014] 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*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2017] Yang Liu, Prajit Ramachandran, Qiang Liu, and Jian Peng. Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*, 2017.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

[Nachum *et al.*, 2016] Ofir Nachum, Mohammad Norouzi, and Dale Schuurmans. Improving policy gradient by exploring under-appreciated rewards. *arXiv preprint arXiv:1611.09321*, 2016.

[Robert, 2004] Christian P Robert. *Monte carlo methods*. Wiley Online Library, 2004.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[Strub and de Vries, 2017] Florian Strub and Harm de Vries. Guesswhat?! models. `https://github.com/GuessWhatGame/guesswhat/`, 2017.

[Strub *et al.*, 2017] Florian Strub, Harm de Vries, Jérémie Mary, Bilal Piot, Aaron C. Courville, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

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

[Sutton and Barto, 1998] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[Thrun, 1992] Sebastian B Thrun. Efficient exploration in reinforcement learning. 1992.

[Williams, 1992] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.