Bivariate Beta LSTM

Kyungwoo Song  JoonHo Jang  Seung jae Shin  Il-Chul Moon
{gtshs2,adkto8093,tmdwo0910,icmoon} @kaist.ac.kr
Department of Industrial and Systems Engineering
KAIST
Daejeon, South Korea

Abstract

Long Short-Term Memory (LSTM) infers the long term dependency through a cell state maintained by the input and the forget gate structures, which models a gate output as a value in $[0,1]$ through a sigmoid function. However, due to the graduality of the sigmoid function, the sigmoid gate is not flexible in representing multimodality or skewness. Besides, the previous models lack correlation modeling between the gates, which would be a new method to adopt domain knowledge. This paper proposes a new gate structure with the bivariate Beta distribution. The proposed gate structure enables hierarchical probabilistic modeling on the gates within the LSTM cell, so the modelers can customize the cell state flow. Also, we observed that our structured flexible gate modeling is enabled by the probability density estimation. Moreover, we theoretically show and empirically experiment that the bivariate Beta distribution gate structure alleviates the gradient vanishing problem. We demonstrate the effectiveness of bivariate Beta gate structure on the sentence classification, image classification, polyphonic music modeling, and image caption generation.

1 Introduction

One of the most commonly used Recurrent Neural Network (RNN) variant is Long Short-Term Memory (LSTM) [1], which introduces additional gate structures for controlling cell states. LSTM controls the information flow from a sequence with an input, a forget, and an output gate. The input and the forget gates decide the ratio of mixture between the current and the previous information at each time step. There has been a question on the sigmoid function used for the gates in LSTM. The sigmoid function is defined to be bounded and monotonically increasing, so the sigmoid has been a popular choice for such gate mechanisms. For instance, the confined gate value range, which is narrower than the 0-1 bound, means the majority of gate values fall into the narrower range that makes the gate values lose potential discrimination power. Some tried to use additional hyper-parameters to sharpen the sigmoid function, i.e., the sigmoid function with temperature parameter [2], but these would be limited to the support for the sigmoid function. From this perspective, there are few works to probabilistically model the flexibility of the gate structure, i.e., $G^2$LSTM [2] with the Bernoulli distribution, but the current probabilistic model missed the gradual aspect of the gate value change.

Moreover, it has been known that the gates could be correlated, and the performance can be improved by exploiting this correlation structure. One common conjecture is the correlation between the input and the forget gate values in LSTM [1]. However, the structure of LSTM does not explicitly model such correlation, so its enforcing structure was handled at the technical implementation level. For instance, CIFGLSTM [3] enforces the negative correlation between the input and the output gate values. CIFGLSTM show competitive performance with reduced parameters because of the
correlation modeling. However, CIFGLSTM enforces the strict negative correlation: -1 only, and it may not align to datasets. We can improve the correlation structure that the correlation is adaptable to the dataset and its range lies in [-1,1] flexibly.

We propose a bivariate Beta LSTM (bBLSTM) which improves the sigmoid function in the input gate, and the forget with a bivariate Beta distribution. bBLSTM has three advantages over the LSTM. First, since a Beta distribution is a generalized distribution of the uniform distribution, the power function, and the Bernoulli distribution with 0.5 probability; the Beta distribution can represent values of [0,1] flexibly. The Beta distribution can represent either symmetric or skewed shape by adjusting two shape parameters. Second, the bivariate Beta distribution can represent the covariance structure of input and forget gates because a bivariate Beta distribution shares the Gamma random variables which make the correlation between two sampled values. We utilized the property of the bivariate Beta distribution for modeling the input and the forget gates in bBLSTM. The bivariate Beta distribution covariance could be further elaborated by expanding the probabilistic model, i.e., adding a common cause prior to the input gate and the forget gate distributions. Third, the bivariate Beta distribution can alleviate the gradient vanishing problem of LSTM. Under a certain condition, we verify that the derivative of gates in bBLSTM is greater than those of LSTM, experimentally and theoretically.

2 Preliminary: Stochastic Gate Mechanism in Recurrent Neural Networks

Since RNN is a deterministic model, it is difficult to prevent overfitting and to generate diverse outputs. Therefore, multiple methods were explored to model the stochasticity in sequence learning. First, dropout methods for RNN \cite{4,6} demonstrated that a stochastic masking can improve its generalization. Second, latent variables were a good combination with the RNN structure, such as Variational RNN (VRNN) \cite{7} and latent variable hierarchical recurrent encoder decoder (VHRED) \cite{8}. Third, gate mechanisms, which is extensively used in RNN variants due to the vanishing gradient, can be substituted with probabilistic models.

When we investigate further on the gate mechanism, there have been efforts in reducing the number of gates \cite{9}, enabling a gate structure to be a complex number \cite{10}, correlating gate structure \cite{3}, modeling gates to be probabilistic \cite{11}. For instance, Gumbel Gate LSTM (\(G^2\)LSTM) \cite{2} replaces the sigmoid function of the input and the forget gates with Bernoulli distributions. The Bernoulli gates in \(G^2\)LSTM turns the continuous gate values to be the binary value of 0 or 1. This work is interesting because of its reparameterization effort to model the Bernoulli distribution. However, this work can be expanded to incorporate a continuous spectrum, a multi-modality, and stochasticity, at the same time. Furthermore, \(G^2\)LSTM uses a Gumbel-Softmax which still remains in the realm of sigmoid gates, so the limitations discussed in the Introduction are still applicable. Figure 2 enumerates the gate and the information flow in \(G^2\)LSTM, and \(G\) is the Gumbel-softmax function with a temperature parameter of \(\tau\) \cite{12}.

When we consider a stochastic expansion on gate mechanisms, it is natural to structure the random variables with conditional independence and priors. For example, the input and the forget gates are both related to the cell state in the LSTM cell, so we may conjecture its correlations through a common cause prior. To our knowledge, CIFGLSTM \cite{3} in Figure 1 is the first model to introduce a structured input and forget gate modeling by assuming \(f_t = 1 - i_t\). This hard assignment is not a flexible covariance modeling, so this can be further extended by adopting the previous probabilistic gate mechanism.

\[i_t = G(W_{xi}x_t + W_{hi}h_{t-1} + b_i, \tau)\]  
\[f_t = G(W_{xf}x_t + W_{hf}h_{t-1} + b_f, \tau)\]  
\[\tilde{c}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)\]  
\[c_t = \tilde{c}_t \odot c_{t-1} + i_t \odot \tanh(\tilde{c}_t)\]  
\[\sigma_t = \sigma(W_{zo}x_t + W_{ho}h_{t-1} + b_o)\]  
\[h_t = o_t \odot \tanh(c_t)\]

\(G\) is the Gumbel-softmax function with a temperature parameter of \(\tau\) and the common cause prior to the input gate and the forget gate distributions. Third, the bivariate Beta distribution covariance could be further elaborated by expanding the probabilistic model, i.e., adding a common cause prior to the input gate and the forget gate distributions. This, the bivariate Beta distribution can alleviate the gradient vanishing problem of LSTM. Under a certain condition, we verify that the derivative of gates in bBLSTM is greater than those of LSTM, experimentally and theoretically.
3 Methodology

Our probabilistic gate model will reside in a neural network cell, as Figure 3, the inference on
the distribution parameters should utilize the reparameterization technique, read Appendix A for
details. Our reparametrization on the Beta distribution uses a composition of reparametrized Gamma
distributions\(^2\). The followings are its technical walkthrough. We sample a random variable \(u_t^{(j)}\)
from a Gamma distribution with \(U_t^{(j)}\) shape parameter, which is parameterized by the current input
and the previous hidden state. As we follow the reparameterization technique of optimal mass
transport (OMT) gradient estimator [14] which utilize the implicit differentiation, we can compute
the stochastic gradient of random variable \(u_t^{(j)}\) with respect to \(U_t^{(j)}\) without inverse CDF.

\[
U_t^{(j)} = g_j(x_t, h_{t-1}), j = 1, ..., 4
\]

\[
u_t^{(j)} \sim \text{Gamma}(U_t^{(j)}, 1), j = 1, ..., 4
\]

\[
i_t = \frac{u_t^{(1)}}{u_t^{(1)} + u_t^{(2)}}, f_t = \frac{u_t^{(3)}}{u_t^{(3)} + u_t^{(4)}}
\]

We formulate \(U_t^{(j)}\), the shape parameter of a Gamma distribution; as a function, \(g_j\) of the current
input \(x_t\) and the previous hidden state \(h_{t-1}\). We omit the amortized inference on the rate parameter
of a Gamma distribution by setting it as a constant of 1. Each \(g_j\) can be a multi-layered perceptron
(MLP) that combines \(x_t\) and \(h_{t-1}\). Given the shape parameter’s constraint of being positive, the final
output of the MLP can be gained by either softplus or Relu.

3.1 Beta LSTM

This section proposes a Beta LSTM (BLSTM) that embeds independent Beta distributions on the
input and the forget gates, instead of the sigmoid function. We construct each Beta distribution with
two Gamma distributions.

\[
U_t^{(j)} = g_j(x_t, h_{t-1}), j = 1, ..., 4
\]

\[
u_t^{(j)} \sim \text{Gamma}(U_t^{(j)}, 1), j = 1, ..., 4
\]

\[
i_t = \frac{u_t^{(1)}}{u_t^{(1)} + u_t^{(2)}}, f_t = \frac{u_t^{(3)}}{u_t^{(3)} + u_t^{(4)}}
\]

We can approximate a random variable \(i_t \sim \text{Beta}(U_t^{(1)}, U_t^{(2)})\) by rewriting \(i_t = \frac{u_t^{(1)}}{u_t^{(1)} + u_t^{(2)}}\) where
\(u_t^{(1)} \sim \text{Gamma}(U_t^{(1)}, 1)\) and \(u_t^{(2)} \sim \text{Gamma}(U_t^{(2)}, 1)\) [13].
The bivariate Beta distribution utilizes Gamma random variables to handle the correlation between the input and the forget gate values, but bBLSTM(3G) can only model the positive correlation between 0 and 1 [15]. In practice, for example, natural language processing, the input and the forget gates might show both positive and negative correlations in cases. Sequential correlated words, i.e., idioms or phrases, would prefer a positive correlation because the cell state should include both previous and current information. On the contrary, if there is a connection that is semantically turning the information, the cell state should disregard the previous information and adapt the current information. The latter case will require a negative correlation, but bBLSTM(3G) lack this functionality. We extended the bivariate Beta distribution in bBLSTM(3G) to be bBLSTM(5G) that uses a bivariate Beta distribution with a more flexible covariance structure. bBLSTM(5G) consists of five random variables following a Gamma distribution, and bivariate Beta distribution with five random variables can handle both negative and positive correlation [16]. bBLSTM(5G) is a generalized model of CIFGLSTM with a probabilistic covariance model in terms of correlation. The formulation of \( U_t^{(j)} \) and \( u_t^{(j)} \) is same within Equation (11) for all \( j \).

\[
\begin{align*}
    i_t &= \frac{u_t^{(1)} + u_t^{(3)}}{u_t^{(1)} + u_t^{(3)} + u_t^{(4)} + u_t^{(5)}}, \quad f_t = \frac{u_t^{(2)} + u_t^{(4)}}{u_t^{(2)} + u_t^{(3)} + u_t^{(4)} + u_t^{(5)}}
\end{align*}
\tag{11}
\]

Another advantage of using a bivariate Beta distribution as an activation function is resolving the gradient vanishing problem of LSTM. We provide a proposition that the gradient value of a gate value in bBLSTM(5G) with respect to the gate parameter is larger than that of LSTM under the certain condition.

**Proposition 1.** (Given the proof in Appendix D.) Assume that \( u_t^{(j)} < 0.8 \) or \( 8 < U_t^{(j)} \) which satisfy the \( |u_t^{(j)} - U_t^{(j)}| \leq \delta \cdot U_t^{(j)} \) for all \( j \) and \( \delta > 0 \), and that \( u_{3:5} \) be 0.5. Then,

(i) for \( 0 < \delta \leq \frac{8}{117} \), the range of \( \frac{\partial i_t}{\partial U_t^{(1)}} \big|_{U_t^{(1)}=0.5} \) lies in \([S_0, S_1]\)

(ii) for \( \delta > \frac{8}{117} \), the range of \( \frac{\partial i_t}{\partial U_t^{(1)}} \big|_{U_t^{(1)}=0.5} \) lies in \([S_0, S_1]\) where \( S_0 = \frac{-(361255^2 - 1077800^2 - 214200)}{18800(4+3)^2} \) and \( S_1 = \frac{-(361255^2 - 1077800^2 - 214200)}{18800(4+3)^2} \)

### 3.3 bivariate Beta LSTM with Structured Prior Model

Hierarchical Bayesian modeling can impose uncertainty on a model as well as a mutual dependence between variables. bBLSTM(5G) has a component of probabilistic modeling, and it is easy to incorporate a prior distribution to the likelihood of the gate value. We present bBLSTM(5G) with prior, bBLSTM(5G+p), and we optimize bBLSTM(5G+p) by maximizing the log marginal likelihood of the target sequence \( y_{1:T} \) in Equation (12) see Figure 3d. However, the direct maximization of the log marginal likelihood is intractable, so we approximate the log marginal likelihood with the evidence lower bound (ELBO) [17] [18] in Equation (13) with a variational distribution, \( q \), which is a feed-forward neural network with the current input \( x_t \) and the previous hidden \( h_{t-1} \).

\[
\log p(y_{1:T}) = \log \int p(u_{1:5}^{(1:5)}) p(y_t|u_t^{(1:5)}) du_t^{(1:5)}
\tag{12}
\]

\[
\mathcal{L}_{ELBO} = \sum_{t=1}^{T} \mathbb{E}_{q(u_t^{(1:5)}|x_t, h_{t-1})}[p(y_t|u_t^{(1:5)})] - KL[q(u_t^{(1:5)}|x_t, h_{t-1}) \parallel p(u_t^{(1:5)})]
\tag{13}
\]

The prior distribution in Equation (12) [13] becomes a distribution with either constant parameters or the parameters inferred by a neural network. VRNN [7] and VHRED [8], which are the variational recurrent model for the context modeling, set up their priors as neural networks depending on the previous input. Similarly, we model a prior distribution of \( p(u_t^{(1:5)}) \), as a Gamma distribution which is a conjugate distribution of the Gamma distribution of \( u_t^{(1:5)} \) in Equation (13). A Gamma distribution takes two parameters, which are shape and rate, and our framework enables learning of the two parameters with an inference network.

The prior on the gate in bBLSTM(5G+p) is extended to incorporate other probabilistic or deep generative models, such as Latent Dirichlet Allocation (LDA) [19] or word vector, i.e. Glove [20].
Considering the proposed gate resides in a LSTM cell at a certain time, \( t \), the prior can be better off from a global context extracted from \( x_{1:T} \). To demonstrate this capability, we adapted Equation 13 to be \( \sum_{t=1}^{T} \mathbb{E}_{q(u_t^{1:5} | x_t, h_{t-1})} [p(u_t | u_t^{1:5})] - \lambda \{ \text{KL}(q(u_t^{3:5} | x_t, h_{t-1}) \parallel p(u_t^{3:5})] + \text{KL}(q(u_t^{1:2} | x_t, h_{t-1}) \parallel p(u_t^{1:2} | \beta_{t-1}, \beta_t)) \}. \) Here, \( \lambda \) is the weight of the prior regularization, to balance the likelihood and KL regularization term; and \( \beta_t \) is the topic probability of a word at time \( t \) in the sequence, which follows the definition in the original LDA. While we used a pre-trained LDA model, a modeler can alternatively learn the LDA parameter since it has its own ELBO for the gradient descent. Finally, we assume \( p(u_t^{1:2} | \beta_{t-1}, \beta_t) = \text{Gamma}(k_{RBF}(\beta_t, \beta_{t-1}), 1) \). Since a radial basis kernel function, \( k_{RBF} \), is used, its length and scale parameters is also learnable. The model on \( p(u_t) \) can be adapted to domains, and our modeling motivation is capturing the significant word semantic compositions to influence the input and the forget gate outputs. It should be noted that \( k_{RBF} \) can be replaced by MLP. Additionally, \( \beta \) can be substituted by a word vector, i.e. Glove. Figure 4 illustrates three insights. First, the strength of the prior should be limited by \( \lambda \). Second, the prior with LDA is generally better than the prior with a static parameter, bBLSTM(5G+p). Third, it is important to learn the inference model, i.e. the kernel hyperparameters used for the parameter of \( p(u_t^{1:2} | \beta_{t-1}, \beta_t) \). For the formal experiments in the next section, we utilized a constant parameter because of the computational burden.

4 Experiments

4.1 Text Classification

Text classification is one of the most frequent benchmark tasks for LSTM. To perform a sentence-level classification task, we need to model the overall semantics of a sentence by focusing on keywords. Effective modeling on input and forget gates are necessary to attend the keywords and to preserve the selective information, respectively. We compare our models on six benchmark datasets. For the comparison on a sentence-level classification task, we use customer reviews (CR) [21], sentence subjectivity (SUBJ) [22], movie reviews (MR) [23], questions type (TREC) [24], opinion polarity (MPQA) [25], and Stanford Sentiment Treebank (SST) [26] dataset. For LSTM model structure, we use a two-layer structure with 200 hidden dimensions for each layer. We set the hidden dimensions of models to have the same number of parameters across the compared models. Table 1 shows the test accuracies for each model and each dataset. bBLSTM(5G+p) performs better than other models on all datasets except TREC with a marginal difference.

| Models     | Size | CR   | SUBJ  | MR   | TREC | MPQA | SST  |
|------------|------|------|-------|------|------|------|------|
| LSTM       | 351k | 82.9±2.9 | 92.6±0.8 | 80.4±0.9 | 94.4±1.0 | 89.4±0.5 | 88.1±0.6 |
| CIFGLSTM   | 356k | 83.3±1.7 | 92.7±0.8 | 79.9±0.9 | 94.0±0.7 | 89.1±0.9 | 87.6±0.4 |
| G²LSTM     | 351k | 83.3±1.6 | 92.7±0.7 | 80.1±1.1 | 94.7±0.3 | 89.3±0.5 | 88.4±0.9 |
| BLSTM      | 361k | 84.4±1.8 | 93.3±0.8 | 81.1±0.9 | 94.4±0.6 | 89.6±0.4 | 88.6±0.6 |
| bBLSTM(3G) | 348k | 83.6±2.1 | 93.2±0.7 | 81.4±0.9 | 94.2±0.4 | 89.4±0.9 | 88.2±0.6 |
| bBLSTM(5G) | 360k | 84.4±1.9 | 92.8±0.7 | 81.0±1.0 | 94.3±0.6 | 89.4±0.6 | 88.8±0.4 |
| bBLSTM(5G+p)| 360k | **84.6±2.4** | **93.3±0.7** | **81.6±0.9** | **94.6±0.7** | **89.6±0.4** | **88.9±0.6** |

Table 1: Test accuracies on sentence classification task.

\(^3\) Appendix B provides the detailed statistical description for the datasets and the experimental settings.
We visualize the input gate, the forget gate, and the correlation of bBLSTM(5G+p) to verify our model assumption. Figure 5 shows the input and the forget gates for each model, and Figure 6 shows the correlation between the input and the forget gates. The qualitative observation meets our assumption because the input and the forget gate outputs fully utilize the range of [0, 1] in bBLSTM(5G+p), and the correlation between gates can exhibit both negative and positive values in bBLSTM(5G+p).

Figure 5: The correlation of bBLSTM(5G+p) on CR (Left) and MR (Right) dataset.

(a) LSTM  
(b) CIFGLSTM  
(c) $G^2$LSTM  
(d) bBLSTM(5G+p)

Figure 6: Histogram of input gate value (first row) and the forget gate value (second row) on CR dataset. Our proposed model bBLSTM(5G+p) shows the more flexible gate value than that of other models. CR dataset is used for the sentiment classification, and only a few words are important instead of whole words. As a result, the input gate in all models has a relatively higher portion of 0 than the portion of value 1. The bBLSTM(5G+p) is more likely to have such a tendency, and it leads to better performance of bBLSTM(5G+p) on CR dataset.

To further understand the model structure and its assumptions, we performed qualitative analysis on a sentence example in the MR dataset. Figure 7 shows the heatmap of the input gate and the forget gate for each model; and the correlation from our proposed model, bBLSTM(5G+p). bBLSTM(5G+p) model has a large input gate value on "but he loses his focus" ($t = 22 \sim 26$) and a large forget gate value on 25 and 26 timestep to propagate the "losses" information well. Because of the appropriate input and forget gate modeling, bBLSTM(5G+p) can compose the meaning of "but he loses his focus" well. This effect originates from the structured gate modeling, which handles the correlation while other models, such as LSTM, $G^2$LSTM, and BLSTM, do not model. There is a relatively large correlation in the "his focus" ($t = 25, 26$), and as a result, both input and forget gates have large values to propagate the important information efficiently. The sentiment label for the sentence is negative, and only bBLSTM(5G+p) classifies the label correctly.

4.2 Polyphonic Music Modeling

Unlike the text classification whose purpose is predicting a single label for entire timesteps, polyphonic music modeling predicts a binary vector at every timestep. Therefore, it is important to model the input and the forget gate appropriately for every timestep. We use four polyphonic music modeling benchmark datasets: JSB Chorales, Muse, Nottingham, and Piano [27, 28]. Table 2 shows the test negative log-likelihood(NLL) on four music datasets [3]. For the comparison, we use two-layered RNNs with 200 hidden dimensions for each layer with Adam optimizer [29]. For a fair comparison, all models are adjusted to have the same number of parameters. Our proposed model, bBLSTM(5G+p), performs better than other models. The number of training samples in polyphonic music dataset is relatively small, and it shows that the prior modeling to the gate improves the robustness.
Figure 7: Visualization of input and forget gates for each model and the correlation for bBLSTM(5G+p). The dataset is designed for the sentiment classification task, and the "but he losses his focus" \( t = 22 \sim 26 \) is an important part. At time step 25, both input gate and forget gate have high values to propagate the information "losses his" efficiently. This is the result of a relatively large correlation value at time step 25, and this correlation helps to propagate the information through the model.

4.3 Pixel by Pixel MNIST

Pixel-by-pixel MNIST task is predicting a category for a given 784 pixels \([30]\). There are two tasks for sequential MNIST: sMNIST, and pMNIST. sMNIST task handles each pixel with a sequential order, and pMNIST task models each pixel in a randomly permutated order. 784 timesteps are longer than the average text sentence length, so it becomes a challenging task to overcome the long term dependency.

We divide the MNIST dataset into three sets: 50,000, 10,000, 10,000 for the train, the validation, and the test dataset, respectively. For LSTM baseline, we use a single-layer model with 128 hidden dimensions with Adam optimizer\[3\]. We set an appropriate hidden dimension for other models to the fair comparison. Table 3 shows the test error rates for sMNIST and pMNIST. For the pMNIST classification task, bBLSTM(5G+p) shows the best performance, while CIFGLSTM shows the best performance on the sMNIST classification task. Left in Figure 8 shows the gradients flow for each time step and the validation error curve for each training epoch on the pMNIST dataset. From the perspective of the gradient flow, we calculate the Frobenius norm of the gradient \( \partial L / \partial \text{ELBO} \), and we average the norm over the image instance. We found that our proposed models, BLSTM, bBLSTM(5G), and bBLSTM(5G+p), propagate the information to the early timestep efficiently. Right in Figure 8 shows the validation error curve, and our proposed model bBLSTM(5G+p), which incorporates the prior, shows the relatively stable learning curve among the stochastic models.

| Models       | JSB  | Muse | Nottingham | Piano |
|--------------|------|------|------------|-------|
| LSTM         | 8.61 | 7.15 | 3.25       | 7.99  |
| CIFGLSTM     | 8.68 | 7.30 | 3.27       | 8.33  |
| G²LSTM       | 8.67 | 7.16 | 3.21       | 8.18  |
| BLSTM        | 8.62 | 7.12 | 3.29       | 8.02  |
| bBLSTM(5G)   | 8.56 | 7.14 | 3.28       | 7.83  |
| bBLSTM(5G+p) | **8.30** | **7.02** | **3.16**   | **7.55** |

Table 2: Negative log-likelihood on polyphonic music

| Models       | sMNIST | pMNIST |
|--------------|--------|--------|
| LSTM         | 3.99   | 9.81   |
| CIFGLSTM     | **1.32** | 8.83   |
| G²LSTM       | 4.46   | 9.49   |
| BLSTM        | 3.76   | 8.30   |
| bBLSTM(5G)   | 2.10   | 8.04   |
| bBLSTM(5G+p) | 1.47   | **7.89** |

Table 3: Test error rates on MNIST
Figure 8: Average gradient norm, $\| \frac{\partial L_{ELBO}}{\partial c_t} \|$ for loss $L_{ELBO}$ over each time step. BLSTM and bBLSTM(5G+p) considers long-term dependency relatively well because they have larger gradients for initial timesteps, (left). bBLSTM(5G+p) which incorporates the prior distribution shows relatively stable validation error curve between stochastic models, $G^2$LSTM, BLSTM, and bBLSTM(5G+p) (right).

4.4 Image Captioning

To verify the compatibility of our proposed model with other models, we evaluate our model on the image captioning task with Microsoft COCO dataset (MS-COCO) [31]. For the experiment, we split the dataset into 80,000, 5,000, 5,000 for the train, the validation, and the test dataset, respectively [32]. We use 512 hidden dimensions for the conditional caption generation, and we also used Resnet [33] to retrieve image feature vectors. Table 4 shows the test performance for MS-COCO dataset based on Show & Tell [34] encoder-decoder structure. bBLSTM(5G+p) shows the best performance in BLEU, ROUGE-L, and SPICE.

| Models                      | B-1 | B-2 | B-3 | B-4 | METEOR | CIDEr | ROUGE-L | SPICE |
|-----------------------------|-----|-----|-----|-----|--------|-------|---------|-------|
| DeepVS [32]                 | 62.5| 45.0| 32.1| 23.0| 19.5   | 66.0  | —       | —     |
| ATT-FCN [35]                | 70.9| 53.7| 40.2| 30.4| 24.3   | —     | —       | —     |
| Show & Tell [34]            | —   | —   | —   | 27.7| 23.7   | 85.5  | —       | —     |
| Soft Attention [36]         | 70.7| 49.2| 34.4| 24.3| 23.9   | —     | —       | —     |
| Hard Attention [36]         | 71.8| 50.4| 35.7| 25.0| 23.0   | —     | —       | —     |
| MSM [37]                    | 73.0| 56.5| 42.9| 32.5| 25.1   | 98.6  | —       | —     |
| Adaptive Attention [38]     | 74.2| 58.0| 43.9| 33.2| 26.6   | 108.5 | —       | —     |
| $h$-detach [39]             | 72.0| 54.6| 39.8| 28.8| 24.8   | 94.7  | 52.5    | 17.9  |
| **Show & Tell with Resnet152** (Our implementation) | 71.7| 54.4| 39.7| 28.8| 24.8   | 93.0  | —       | —     |

Table 4: Test performance on MS-COCO dataset for image captioning for BLEU [40], METEOR [41], CIDEr [42], ROUGE [43], SPICE [44] evaluation metric. We re-implement the Show and Tell [34] based on resnet152, and compare the performance between models.

5 Conclusion

We propose a new structured gate modeling which can improve the LSTM structure through probabilistic modeling on gates. While the current sigmoid gate would satisfy the boundedness, we improve the sigmoid function with a Beta distribution to add flexibility. Moreover, bBLSTM enables detailed modeling on the covariance structure between gates, and bBLSTM+prior guides the learning of the covariance structure. Also, our propositions state the improved characteristics of our probabilistic gate compared to the sigmoid function. This work envisions how to incorporate the neural network models with probabilistic components to improve its flexibility and stability. We demonstrated our model is a case study of incorporating the prior to the gate structure.
References

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[2] Zhuohan Li, Di He, Fei Tian, Wei Chen, Tao Qin, Liwei Wang, and Tie-Yan Liu. Towards binary-valued gates for robust lstm training. *arXiv preprint arXiv:1806.02988*, 2018.

[3] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017.

[4] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016.

[5] Konrad Zolna, Devansh Arpit, Dendi Suhubdy, and Yoshua Bengio. Fraternal dropout. In *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, 2018.

[6] Sungrae Park, Kyungwoo Song, Mingi Ji, Wonsung Lee, and Il-Chul Moon. Adversarial dropout for recurrent neural networks. *CoRR*, abs/1904.09816, 2019.

[7] Junyoung Chung, Kyle Kastner, Laurent Dinl, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015.

[8] 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.

[9] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1724–1734, 2014.

[10] Moritz Wolter and Angela Yao. Complex gated recurrent neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10536–10546. Curran Associates, Inc., 2018.

[11] Matthew Thornton, Jithendar Anumula, and Shih-Chii Liu. Gaussian-gated LSTM: Improved convergence by reducing state updates. 2018.

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, April 24-26, 2017, Conference Track Proceedings, 2017.

[13] Francisco R Ruiz, Michalis Titsias RC AUEB, and David Blei. The generalized reparameterization gradient. In *Advances in neural information processing systems*, pages 460–468, 2016.

[14] Martin Jankowiak and Fritz Obermeyer. Pathwise derivatives beyond the reparameterization trick. *arXiv preprint arXiv:1806.01851*, 2018.

[15] A bivariate beta distribution. *Statistics and Probability Letters*, 62(4):407–412, 2003.

[16] Flexible bivariate beta distributions. *Journal of Multivariate Analysis*, 102(8):1194–1202, 2011.

[17] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.

[19] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
[20] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1532–1543, 2014.

[21] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004, pages 168–177, 2004.

[22] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain., pages 271–278, 2004.

[23] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA, pages 115–124, 2005.

[24] Xin Li and Dan Roth. Learning question classifiers. In 19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002, 2002.

[25] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3):165–210, 2005.

[26] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1631–1642, 2013.

[27] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012, 2012.

[28] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. CoRR, abs/1803.01271, 2018.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[30] Quoc V. Le, Navdeep Jaitly, and Geoffrey E. Hinton. A simple way to initialize recurrent networks of rectified linear units. CoRR, abs/1504.00941, 2015.

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, pages 740–755, 2014.

[32] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 3128–3137, 2015.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778, 2016.

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 3156–3164, 2015.

[35] Quanzeng You, HaiLin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 4651–4659, 2016.
[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, pages 2048–2057, 2015.

[37] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 4904–4912, 2017.

[38] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 3242–3250, 2017.

[39] Devansh Arpit, Bhargav Kanuparthi, Giancarlo Kerg, Nan Rosemary Ke, Ioannis Mitliagkas, and Yoshua Bengio. h-detach: Modifying the LSTM gradient towards better optimization. CoRR, abs/1810.03023, 2018.

[40] 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 of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA., pages 311–318, 2002.

[41] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005, pages 65–72, 2005.

[42] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 4566–4575, 2015.

[43] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain., pages 605–612, 2004.

[44] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V, pages 382–398, 2016.