An Informative Adaptive Receptive Field Model for Recurrent Neural Networks

Abstract

Recently, the development of Information Bottleneck Theory has significantly promoted understandability of deep neural network. However, few works talks about recurrent neural network from this perspective. We introduce a mechanism for RNN called adaptive receptive field, which focuses on the preliminary treatment for input subsequence by adaptively adjusting the size of receptive field and processing the input subsequence in the field. Analysis and experiments show that the size of receptive field is related to both sufficiency and minimum principle of information bottleneck theory, which thus can be uesd as an auxiliary parts to promote comprehension of RNN from information bottleneck perspective. Experiments also reveal that our architecture for RNN can obviously improve the performance of baseline models.

Introduction

Recently, deep learning has achieved state-of-the-art results in many machine learning tasks (Lecun *et al.* 2015). However, the theory about deep learning is far behind its achievement. The general design principles of deep networks and the internal process are not well understood.

Recently, the Information Bottleneck Theory has drawn lots of attention, and helps to explain deep learning theoretically. It not only helps to understand the representation in the neural network, but also explains the optimization process and predicts the generalization ability.

Information bottleneck theory, which is introduced by (Tishby *et al.* 2000a), proposes that the representation should not only contain as much information of label data as possible, but also contain as little information of input data as possible. The two principles are referred to as sufficiency and minimum respectively. This theory formulates a variational principle for efficient representation of relevant information.

Several works have discussed the relationship between this principle and optimization of neural network. For example, In (Achille and Soatto 2018), it's shown that information bottleneck Lagrange can be done implicitly by several common trainning methods such as SGD and regularization, which indicates universality of information bottleneck theory.

It's also inspiring to consider training process in the view of information bottleneck theory. (Shwartz-Ziv and Tishby 2017) shows that in the initial phase the sufficiency principle will be followed mainly, and minimum principle will be dominant later on.

However, studies related to information bottleneck theory mainly concentrate on feedforward network. There are few works talk about recurrent neural network from this perspective.

In recent years, recurrent neural networks (RNN) have become more widely-used in various tasks and show the robust and outstanding performance, and much work has been done on improving RNNs. There are two of the most common ways. One is introducing gate mechanism (S.Hochreiter and J.Schmidhuber 1997; Xu *et al.* 2016; Kalchbrenner *et al.* 2015), and the other is dynamically adjusting the way of reading input sequence (Yu *et al.* 2017; Chang *et al.* 2017). However, theoretical analysis methods is deficient, and it's enlightening to consider information bottleneck theory.

Generally, RNN can be analyzed in the same way as feedforward network by information bottleneck theory, which considers the network as Marcov processes and utilizes Data Processing Inequalities. However, the structure of RNNs enables distinctive tricks in the analysis. The intrinsic property of RNN is the repetition of the same model unit on the sequential data. The length of data sequence is closely related to the amount of information, which inspired us to design a model that can reflect information bottleneck theory by this property.

It is worth mentioning that, evidence from cognitive neuroscience have indicated that real neurons' temporal receptive fields are actually dynamic and adaptive (Atiani *et al.* 2009), and are closely related to information (Weinberger *et al.* 2014).

Inspired by these ideas, we proposed a RNN model with mechanism called adaptive receptive field. The size of receptive field can dynamically adapt to the input, which is also a new kind of Attention mechanism (Vaswani *et al.* 2017). Further analysis shows that this mechanism is closely related to information bottleneck theory, where the receptive field size can reflect the trade-off between sufficiency and minimum. We emphasize this property because it significantly promotes

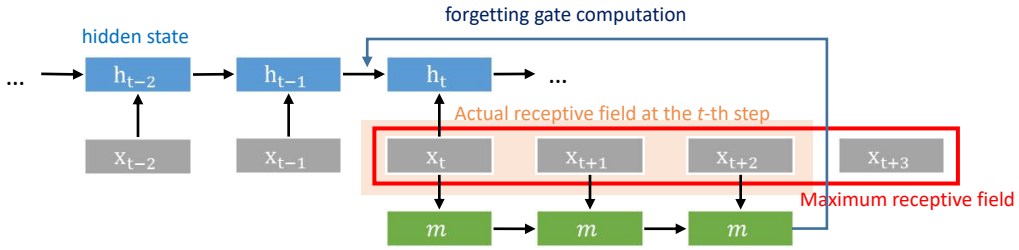


Figure 1: The Architecture of our model.

the comprehensibility of RNN.

The Proposed Model

The receptive field mechanism is actually a serial of functions that obey Multinoulli distribution. These functions share the same RNN structure, while the input sizes of them are different. We call this mechanism adaptive receptive field. The details will be described in this section.

Architecture

Our model (Figure 1) consists of three stages: receptive field RNN, forgetting gate, and standard RNN. The definitions of all these three stages are as follows:

Receptive Field RNN In this stage, our model collects and processes new unseen subsequence of appropriate amount of information by a RNN for post-treatment.

For this purpose, we need to firstly calculate the length of the subsequence to be processed dynamically. We implement it by:

$$\begin{aligned} D_{s_t} &= \text{Multinoulli}(\text{Softmax}(W_{sx}x_t + W_{sh}h_{t-1})) \\ s_t &\sim D_{s_t} \end{aligned} \tag{1}$$

where D_{s_t} denotes the distribution of the size of receptive field (the length of the subsequence to be processed) at the t-th step. *Softmax* denotes softmax function which normalizes a vector into a categorical distribution. W_{sx} and W_{sh} are parameter matrices. The size of receptive field at t-th step s_t is then sampled from D_{s_t}. We assume that the minimum length is 1, and the maximum length is noted by S.

Next, we process subsequence in the receptive field by a vanilla RNN:

$$\begin{aligned} m_{t,i+1} &= g(W_m x_{t+i} + U_m m_{t,i} + b_m) \\ i &\in [0, s_t - 1] \end{aligned} \tag{2}$$

where m denotes the hidden state of RNN. g is a nonlinear function. W_m and U_m are parameter matrices, and b_m is bias, which are all shared throughout the time steps.

The reason why we adopt the calculation above mainly lies in its simplicity. Since the length of subsequence is usually short, using LSTM unit or other relatively complicated recurrent units is unnecessary and wasteful.

It's remarkable that the receptive field is a feed forward form computation, which can utilize new information. This form allows the size of receptive field to be an index for the mutual information between input and hidden state, which will be discussed latter.

Forgetting Gate In this stage, model combines the information coming from receptive field with the last hidden state. The function we use is a forgetting gate, because empirically forgetting gate is the most important gate in recurrent network (van der Westhuizen and Lasenby 2018). However, we should note that forgetting gate doesn't mean decrease of information. It's just a form of computation. We implement it as:

$$F_t = \sigma(W_F m_{s_t} + b_F) \tag{3}$$

$$h^f_{t-1} = F_t \odot h_{t-1} \tag{4}$$

where m_t can either be the last hidden state m_{t,s_t} or the mean of $m_{t,i}, i \in [0, s_t]$. Empirically, the latter makes the optimization easier. σ is the logistic sigmoid function which limits the output between 0 and 1, W_F is parameter matrix and b_F is bias. And they are also shared throughout the time steps. F_t can be seen as a *forgetting weight vector*. h^f_{t-1} denotes the $(t-1)$-th forgotten long-term store. Each value in forgetting weight vector indicates the reservation rate of the information of corresponding dimension in h_{t-1}. $\odot$ denotes element-wise multiplication operator.

Standard RNN This stage finally computes the t-th hidden state h_t. We implement it as:

$$h_t = Unit(x_t, h^f_{t-1}) \tag{5}$$

where $Unit$ denotes any kinds of RNNs' units, such as vanilla recurrent unit, LSTM unit, GRU and some other forms like (van der Westhuizen and Lasenby 2018), which shows the extendibility of our model. Here we take vanilla recurrent unit and LSTM unit as examples.

Training

Since s_t is sampled from D_{s_t} in the stage of receptive field, one alternative approach is to using the REINFORCE (Sutton *et al.* 1999) algorithm and to train the model like (Yu *et al.* 2017). However, the optimization is difficult when the length of total sequence is long.

Another way is to use continuous approximation which makes the back-propagation possible. First, we notice that the size control of receptive field can be realized by mask

form:

$$m_{t,i+1} = (1 - mask_{t,i+1}) * g(W_m x_{t+i} + U_m m_{t,i} + b_m) + mask_{t,i+1} * m_{t,i}, \quad i = 0, 1, ..., S-1 \tag{6}$$

where $mask_t$ is a vector of mask. $mask_{t,i}, i \in [1, s_t]$ equals to 0, and $mask_{t,i}, i \in [s_t + 1, S]$ equals to 1. Note that in the position i where $mask_{t,i}$ equals to 0 the unit computes normally, while in position i where $mask_{t,i}$ equals to 1 the model just delivers the previous hidden state.

With this form, we can then use Gumbel softmax trick to get a soft approximation to the mask, which is a reparameterization method for discrete variable (Jang *et al.* 2016). Suppose the parameter of Multinoulli distribution D_{s_t} is vector d_t. We compute the vector Is_t as following:

$$\begin{aligned} Is_{t,i} &= \frac{\exp\left(\left(\log d_{t,i} + g_i\right)/\tau\right)}{\sum_{j=1}^{S} \exp\left(\left(\log d_{t,j} + g_j\right)/\tau\right)} \\ g_i &\sim Gumbel(0, 1) \\ i &= 1, 2, \cdots, S \end{aligned} \tag{7}$$

which is a soft index vector for s_t. τ is the temperature parameter. Then the mask vector can be computed as following:

$$mask_{t,i} = \sum_{j=1}^{i-1} Is_{t,j} \tag{8}$$

which means $mask_t = [0, Is_1, Is_1 + Is_2, \cdots, Is_1 + Is_2 + \cdot + Is_{S-1}]$. It's remarkable that as the temperature parameter τ becomes close to 0, Is_t can be hard enough and the the same for the mask.

With the reparameterization method, the receptive field model can be trained through gradient based method.

Complexity

The computational complexity is $O(S \cdot T \cdot d^2)$, where S is the maximal receptive field size, T is the length of inputs, and d is the dimension of state. This can be deduced from the fact that in each time step of the main RNN, the receptive field RNN will compute for S steps. The computational complexity increases linearly by S-fold compared to standard RNN, which is $O(T \cdot d^2)$. Because S is usually a small constant even if T changes, the computational expense is acceptable. And we use vanilla RNNs as receptive field RNN, which also decreases the computational cost.

The storage costs is $O(d^2)$, which is independent of T and S because the RNN shares parameters across steps.

Informational Receptive Field

In this section, we introduce Information Bottleneck Theory briefly, and show the relationship between Information Bottleneck Theory and our model.

Information Bottleneck Theory

(Tishby *et al.* 2000b) introduced information bottleneck theory as a general principle for machine learning, especially for representation learning. Several works have further discussed this principle, especially in the context of deep learning (Tishby and Zaslavsky 2015; Shwartz-Ziv and Tishby 2017; Achille and Soatto 2018).

The main idea of information bottleneck theory is to consider learning as a coding problem. The representation follows two principles: sufficient and minimum. Being sufficient means that the representation should contain sufficient information related to the output, while minimum means that the representation should contain as little information of the input data as possible. However, according to the rate distortion theory, it's impossible to minimize the two terms concurrently. There is a trade-of between the two objectives. The two terms can be combined in the Lagrange form, which is called information bottleneck Lagrange. Let's denote the input and output data by x and y, while the representation of x is h, then the information bottleneck Lagrange objective function is

$$\begin{aligned} & L[p(H|X), p(Y|H)] \\ =& \beta I(X; H) + E[KL(p(y|x)||p(y|h))] \\ =& \beta I(X; H) - I(H; Y) + C_1 \\ =& \beta I(X; H) + H(H|Y) + C_2 \end{aligned} \tag{9}$$

where the terms not related to $p(H|X)$ or $p(Y|H)$ are all represented by constant. In the final form, the conditional entropy $H(H|Y)$ is an uper bound of cross entropy loss, which is compatible to traditional objective function. As for the $I(X; H)$ term, (Achille and Soatto 2018) proposed that it can be minimized implicitly by SGD or regularization method such as dropout.

Although the information bottleneck theory provides a powerful theoretical view to understand deep learning, it's difficult to inspect model from this point of view because of the difficulty of mutual information estimation. Previous works usually track tiny networks only (Shwartz-Ziv and Tishby 2017). There are several works which proposed ingenious method to estimate mutual information in certain context (Sønderby *et al.* 2016), but other techniques are still required.

The receptive field mechanism can reflect the trade-off between two principles explicitly, which is a convenient model for analysis of information bottleneck theory.

IB theory and receptive field

In this subsection we will show that the trade-off between $I(X; H)$ and $I(Y; H)$ will be reflected in the receptive field size.

Receptive Field and $I(X; H)$ In the context of recurrent neural network, $I(X; H)$ can be expanded as following:

$$\begin{aligned} I(X; H) = & I(X; h_1) + I\left(X; h_2|h_1\right) \\ & + I(X; h_3|h_{1:2}) + \cdots + I(X; h_T|h_{1:T-1}) \end{aligned} \tag{10}$$

where T is the length of the input sequence. Now consider the t_{th} term. Notice that we can't simply eliminate $h_{1:t-2}$ in the condition side because of the dependency between $h_{1:t-2}$ and x even if conditioned by h_{t-1}. We can expand the term

into

$$\begin{aligned} & I(X; h_t|h_{1:t-1}) \\ =& I(x_{t:T}; h_t|h_{1:t-1}) + I(x_{1:t-1}; h_t|h_{1:t-1}, x_{t:T}) \\ =& I(x_{t:T}; h_t|h_{1:t-1}) \end{aligned} \tag{11}$$

The second equation is because that $x_{1:t-1}$ and h_t will be independent if conditioned by both $h_{1:t-1}$ and $x_{t:T}$.

Let's consider $I(x_{t:T}; h_t|h_{1:t-1}, s_t)$, which is a close approximation to $I(x_{t:T}; h_t|h_{1:t-1})$. In fact,

$$\begin{aligned} & I(x_{t:T}; h_t|h_{1:t-1}, s_t) \\ =& I(x_{t:T}; h_t|h_{1:t-1}) + I(x_{t:T}; s_t|h_{1:t}) \\ & - I(x_{t:T}; s_t|h_{1:t-1}) \end{aligned} \tag{12}$$

The last two terms is upper bounded by $H(s_t)$. Because s_t is one dimension discrete variable which only has a few different values, $H(s_t)$ can be ignored comparing to the first term $I(x_{t:T}; h_t|h_{1:t-1})$.

We then will show that the upper bound of $I(x_{t:T}; h_t|h_{1:t-1}, s_t)$ is related to $E(s_t)$. In fact, we have

$$\begin{aligned} & I(x_{t:T}; h_t|h_{1:t-1}, s_t) \\ =& \sum_{i=1}^{S} p(s_t = i) I(x_{t:T}; h_t|h_{1:t-1}, s_t = i) \end{aligned} \tag{13}$$

and

$$\begin{aligned} & I(x_{t:T}; h_t|h_{1:t-1}, s_t = i) \\ =& I(x_{t:t+i-1}; h_t|h_{1:t-1}, s_t = i) \end{aligned} \tag{14}$$

where the second equation is satisfied because $x_{t+i:T}$ and h_t will be independent if conditioned by $h_{1:t-1}$ and $x_{t:t+i-1}$.

We assume the model satisfies that for $i > j$,

$$\begin{aligned} & I(x_{t:t+j-1}; h_t|h_{1:t-1}, s_t = i) \\ \approx& I(x_{t:t+j-1}; h_t|h_{1:t-1}, s_t = j) \end{aligned} \tag{15}$$

which means in t_{th} step, the new input in the receptive field model will not cause loss of information of previous input . We advocate that this assumption is reasonable because we use the mean of all the hidden state of RF RNN in t_{th} step.

With assumption above, for $i > j$, we have

$$\begin{aligned} & I(x_{t:t+i-1}; h_t|h_{1:t-1}, s_t = i) \\ \geq& I(x_{t:t+j-1}; h_t|h_{1:t-1}, s_t = i) \\ \approx& I(x_{t:t+j-1}; h_t|h_{1:t-1}, s_t = j) \end{aligned} \tag{16}$$

Thus the receptive field size is positively related to $I(X; H)$. So we can predict that the minimization of $I(X; H)$ will cause the descent of $E(s)$. What's more, because the convergence of receptive field size function is slower than the main function, after the initial phase when the main function convergence (Shwartz-Ziv and Tishby 2017), the main field size E(s) can be used for approximating the mutual information.

Receptive Field and $I(Y; H)$ The receptive field size is related not only to $I(X; H)$, but also to $I(Y; H)$. Similar to the previous subsection, $I(Y; H)$ can be expanded as following:

$$\begin{aligned} & I(Y; H) \\ =& I(Y; h_1) + I(Y; h_2|h_1) + \cdots + I(Y; h_T|h_{1:T-1}) \end{aligned} \tag{17}$$

And we consider $I(Y; h_t|h_{t-1}, s_t)$ instead of $I(Y; h_t|h_{1:t-1})$, similar to previous subsection. We will show that the upper bound of $I(Y; h_t|h_{t-1}, s_t)$ is positively related to $E(s_t)$. In fact,

$$\begin{aligned} & I(Y; h_t|h_{1:t-1}, s_t) \\ =& \sum_{i=1}^{S} p(s_t = i) I(Y; h_t|h_{1:t-1}, s_t = i) \end{aligned} \tag{18}$$

and

$$\begin{aligned} & I(Y; h_t|h_{1:t-1}, s_t = i) \\ =& I(Y; x_{t:t+i-1}|h_{1:t-1}, s_t = i) \\ & + I(Y; h_t|h_{1:t-1}, x_{t:t+i-1}, s_t = i) \\ & - I(Y; x_{t:t+i-1}|h_{1:t}, s_t = i) \\ \leq& I(Y; x_{t:t+i-1}|h_{1:t-1}, s_t = i) \end{aligned} \tag{19}$$

So if we ignore the condition effect of $s_t = i$, the upper bound will be positively related to $E(s)$. Thus, the optimization of $I(Y; H)$ will increase the mean receptive field size.

In conclusion, both $I(X; H)$ and $I(Y; H)$ are positively related to receptive field size. Thus mean receptive field size can reflect the trade-off and relative intensity between $I(X; H)$ and $I(Y; H)$ timely. This property enables us to track the trainning process timely with few computing costs, which can be applied to larger networks compared to previous work (Shwartz-Ziv and Tishby 2017).

Experiments

We have shown that receptive field size is positively related to $I(X; H)$ and $I(Y; H)$, while the two terms have trade-off relationship according to Information Bottleneck theory. The exploratory experiment in subsection "Informational Receptive Field Experiments" shows the positive relationship between two terms and receptive field size, and that the receptive field size can reflect the trade-off relationship . Before these exploratory experiment, subsection "Common Task" shows that there are empirical promotes of performance of our model compared to basic models.

Common Task

We evaluate our model by applying it to 2 common tasks corresponding to text classification and sequence labeling. But we should note that our model isn't designed elaborately for excellent performance.

The datasets for text classification includes MR, TREC, SST, AG and IMDB. For sequence labeling, we specifically choose Chinese word segmentation task and PKU and MSR dataset are used. Each dataset is briefly described as follows.

- **MR:** Movie reviews with one sentence per review (Pang and Lee 2005). Classification involves detecting positive/negative reviews.

- **TREC:** TREC question dataset (Li and Roth 2002), in which the objective is to classify each question into 6 question types.
- **SST:** Stanford Sentiment Treebank, which labeled by (Socher *et al.* 2013), involves detecting very positive, positive, neutral, negative, very negative 5 fine-grained labels.
- **AG:** The dataset contains 4 classes of topics (World, Sports, Business, Sci/Tech) from the AG's news corpus.[1]
- **IMDB:** Dataset of 25,000 movies reviews from IMDB, labeled by sentiment (positive/negative). (Maas *et al.* 2011)
- **PKU:** PKU corpus can be obtained from the second International Chinese Word Segmentation Bakeoff (Emerson 2005).
- **MSR:** MSR corpus is also from the second International Chinese Word Segmentation Bakeoff (Emerson 2005).

We use 300-d pretrained word embeddings, and fine-tune the word embeddings during training to improve the performance. The dimension of recurrent states is also set to 300. The number of recurrent layers is 1. The maximum length of the subsequence S is set to 4 empirically.

Training is done through stochastic gradient descent (SGD) over shuffled mini-batches with the Adam update rule (Kingma and Ba 2014). And for regularization, L_2 parameter norm penalty and Dropout (Srivastava *et al.* 2014) are employed.

Performance in Text Classification Table 1 shows the text classification accuracies of the baseline models along with our 6 models RNN+RF, RNN+RF$^-$, RNN+RF(RL), LSTM+RF, LSTM+RF$^-$ and LSTM+RF(RL). RNN/LSTM+RF denotes the model where vanilla recurrent unit or LSTM unit is used. RF$^-$ denotes that the maximum size of the RF S is set to 1, which means the RF always receive input x_t. RL means we train the model with REINFORCE method.

As we can see, for the models based on the vanilla recurrent unit, RNN+RF achieves higher accuracies than RNN on all five datasets significantly. Besides, both RNN+RF$^-$ and RNN+RF(RL) also get higher accuracies than RNN except on SST. Similarly, for the models based on LSTM unit, our model LSTM+RF together with LSTM+RF$^-$ and LSTM+RF(RL) also surpass LSTM definitely on all five datasets. These results demonstrate the advantage of our model which is insensitive to what recurrent unit is used.

Then, we compare the results between two different training method. As is shown in Table 1, models using reparameterization method mostly achieve higher accuracies except RNN+RF(RL) on MR and LSTM+RF(RL) on TREC and IMDB. And we also found that RNN+RF and LSTM+RF have higher stability than RNN+RF(RL)and LSTM+RF(RL) in our experiments.

RNN+RF$^-$ and LSTM+RF$^-$ are simplified versions of RNN+RF and LSTM+RF respectively. Note that though the receptive field only receives x_t, RNN+RF$^-$ and LSTM+RF$^-$ both achieve higher accuracy than RNN and LSTM on all five datasets except RNN+RF$^-$ on SST. We ascribed this effect to the greater network depth.

Performance in Chinese Word Segmentation Chinese word segmentation is a typical sequence labeling task. Labels B/I/E/S are used to mark a character as the beginning, internal (neither beginning nor end), end and only-character (both beginning and end) of a word, respectively.

And in order to evaluate the model without being influenced by some interfering factors, we don't use a window of context characters as input and also don't add post-inference with a Viterbi search or beam search like (Chen *et al.* 2015).

The results of the F-Score on PKU and MSR are shown in Table 2. As we can see, on both PKU and MSR corpus, RF and RF(RL) models achieve higher F-score, especially the models that based on RNN. And similar to text classification task, RF$^-$ models generally slightly inferior than RF and RF(RL) models.

Informational Receptive Field Experiments

Influence of $I(X;H)$ and $I(Y;H)$ Bias As we analyzed previously, the size of receptive field is positively related to $I(Y;H)$ and $I(X;H)$, which allows us to inspect the trade-off in the information bottleneck Lagrange. To demonstrate this property, we inspected how receptive field size will be influenced by the trade off between $I(Y;H)$ and $I(X;H)$.

We chose three datasets, (MR, IMDB and AG) to inspect the influence of $I(Y;H)$. The previous two datasets are binary classification tasks while the last one is multicategory task. Thus, the information bottleneck Lagrange is more biased to $I(Y;H)$ in AG comparing to MR and IMDB.

As (Achille and Soatto 2018) proposed, the stochastic optimization method, like SGD, or the regularization method such as dropout or simply add noise can implicitly minimum $I(X;H)$. For simplicity, in the experiment we apply dropout or not to illustrate the influence of $I(X;H)$ regularization.

Figure 2 shows the mean receptive field size $E(s)$ versus the training step in both dropout regularization and no regularization condition on three dataset.

It's obvious that the receptive field size is sensitively to the regularization of $I(X;H)$. In the no $I(X;H)$ regularization condition, mean receptive field size grows to the maximum directly. However, in the $I(X;H)$ regularization condition, mean receptive field size decreases a lot.

And the influence of $I(Y;H)$ is also significant in $I(X;H)$ regularization condition. In MR and IMDB, the mean receptive field size decreases directly, while in AG, receptive field size grows at the beginning and then reduces. The difference is consistent with our prediction that there is greater dominance of $I(Y;H)$ in AG. and also consistent with (Shwartz-Ziv and Tishby 2017)'s observation.

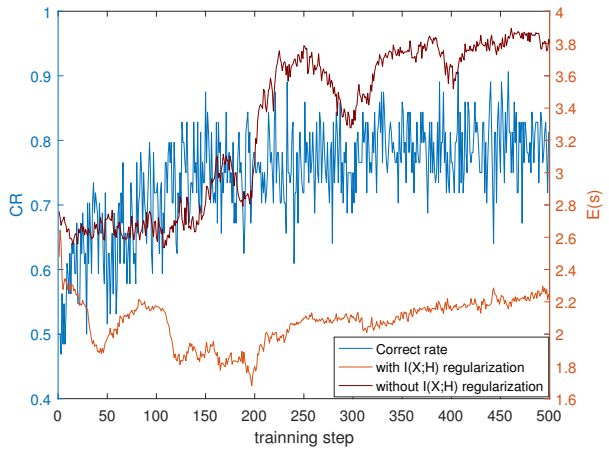
The Trade-off in Concrete Input Besides comparing over total dataset, we inspected the relationship between receptive field size and concrete input. We assume that under the constraint of $I(X;H)$, input with more information related to Y will be assigned less receptive field size.

$H(Y|X=x)$ is a measure for the information between certain input x and output. However, it's difficult to compute this term for long sequence. For simplicity, we only consider

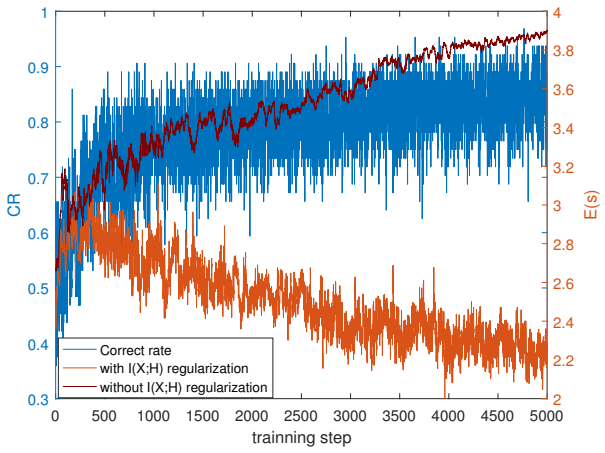
[1]the dataset is available at https://www.di.unipi.it/-gulli/AG_corpus_of_news_articles.html

Model	MR	TREC	SST	AG	IMDB
RNN	75.14%	91.96%	40.49%	89.81%	77.04%
RNN+RF$^-$	76.09%	93.75%	39.98%	90.07%	78.85%
RNN+RF(RL)	**80.68%**	92.19%	39.00%	89.55%	77.68%
RNN+RF	78.36%	**93.76%**	**40.63%**	**90.50%**	**78.92%**
LSTM	79.21%	92.41%	43.19%	92.73%	82.17%
LSTM+RF$^-$	80.44%	93.75%	43.47%	92.84%	83.27%
LSTM+RF(RL)	80.21%	**94.87%**	45.08%	92.76%	**84.30%**
LSTM+RF	**80.87%**	94.20%	**46.14%**	**93.21%**	83.47%

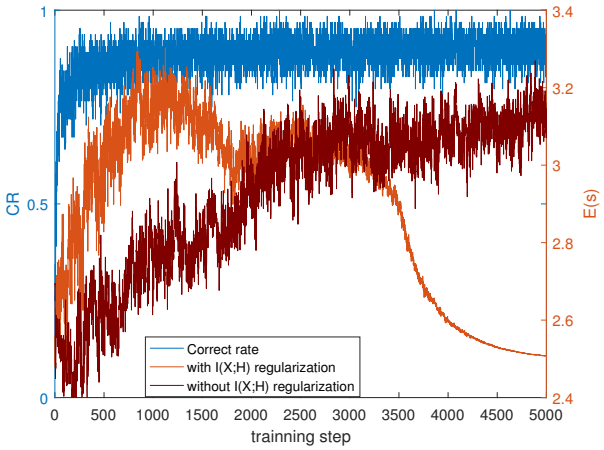
Table 1: Results of the text classification accuracies of the baseline models (RNN and LSTM) along with our models. RNN/LSTM+RF denotes the model where vanilla recurrent unit or LSTM unit is used. RF$^-$ denotes that the maximum size of the RF L is set to 1. RL means we train the model without using continuous approximation.



(a) MR

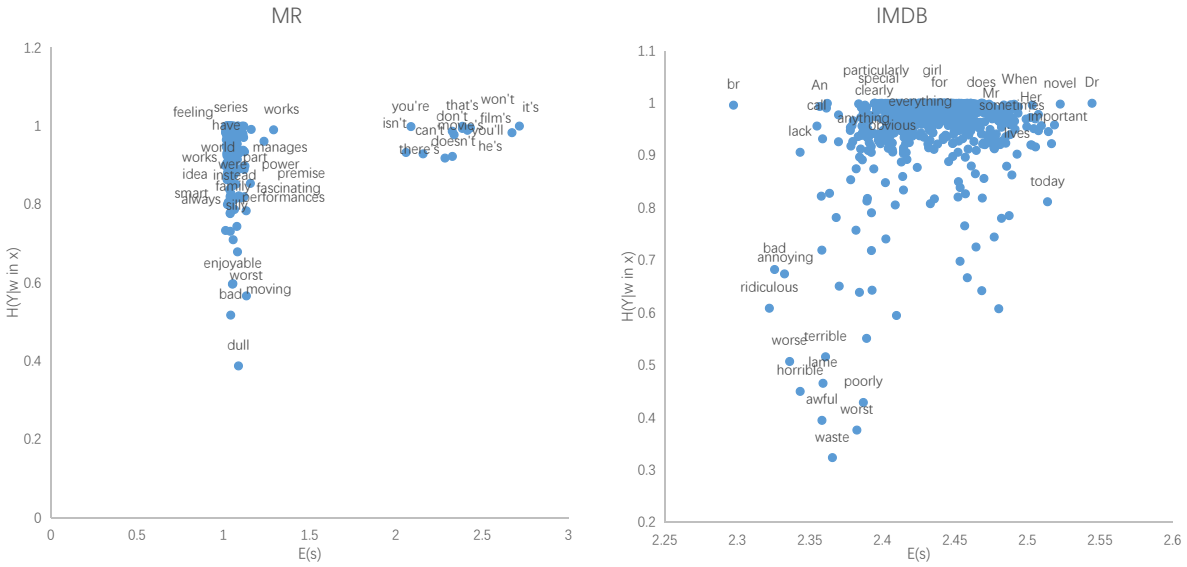


(b) IMDB



(c) AG

Figure 2: The blue line represents correct rate. The light red and deep red represent RF size $E(s_t)$ in no $I(X;H)$ and $I(X;H)$ regularization condition. In $I(X;H)$ regularization condition, $E(s_t)$ decreases a lot. And contrast between data sets reveals influence of $I(Y;H)$.



(a) MR (b) IMDB

Figure 3: The relations between $E(s_t|x_t = w)$ and $H(Y|w \in x)$ on MR (left) and IMDB (right). Horizontal axis represents $H(Y|w \in x)$ and vertical axis represents $E(s_t|x_t = w)$. It's a significant phenomenon that the words with low $H(Y|w \in x)$ always have low RF size, while the words with high RF size have high $H(Y|w \in x)$.

Model	PKU	MSR
RNN	85.27%	87.24%
RNN+RF$^-$	86.12%	87.89%
RNN+RF(RL)	86.65%	89.41%
RNN+RF	**88.19%**	**91.33%**
LSTM	91.94%	94.18%
LSTM+RF$^-$	92.06%	94.28%
LSTM+RF(RL)	**92.30%**	94.58%
LSTM+RF	92.21%	**94.60%**

Table 2: Results of the F-Score of the baseline models along with our models on PKU and MSR.

the input of t_{th} step. To inspect a certain word $x_t = w$, We intuitively measure the information of word w by $H(Y|w \in x)$, which is the entropy of conditional distribution $p(Y|w \in x)$.

Figure 3 shows the relations between $E(s_t|x_t = w)$ and $H(Y|w \in x)$ on datasets MR and IMDB. And each point in Figure 3 denotes a specific word of relatively high frequency.

We should note that because x_t is not the only variable that influences s_t, the distribution of $E(s_t|x_t = w)$ and $H(Y|w \in x)$ are not very well-conditioned. However, there is still some interesting phenomenons.

As we can see, points in Figure 3 are roughly clustered into 3 groups which is on the bottom left, upper left and upper right of the coordinate plane respectively. It's a significant phenomenon that the words with low $H(Y|w \in x)$ always have low receptive field size, while the words with high receptive field size have high $H(Y|w \in x)$. We attribute this phenomenon to the blance between $I(Y;H)$ and $I(X;H)$.

The points on the bottom left are of samller $H(Y|w \in x)$ and smaller $E(s)$. These points are mainly emotional words such as "awful", "horrible" and "enjoyable", et al., which is informative for categorizing. So the receptive field size is decreased to limit $I(X;H)$. Intuitively, models have no need to see a longer subsequence input to get extra information.

On the contrary, as for the points on the upper right, these words lack of categorization information, so models need larger receptive field size to increase $I(Y;H)$. Intuitively, further information is needed for task. Some typical words are "doesn't", "there's" and "sometimes", which have high possibility of being followed by a clause.

Conclusion and Further Research

In this paper, we proposed the adaptive receptive field mechanism, which is closely related to information bottleneck theory. As shown in analysis and experiments, the receptive field size can reflect the two principles of information bottleneck theory, especially the trade-off relationship between them.

Our receptive field model has wide applicability and can be added to all kinds of recurrent unit besides vanilla RNN and LSTM. And it's also feasible to use it in convolutional neural networks which deal with the tensor data by repetition of the same model unit. The application of this mechanism can provide a convenient method for analysis from information bottleneck perspective, which can improve explicability of models.

References

[Achille and Soatto 2018] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[Atiani *et al.* 2009] Serin Atiani, Mounya Elhilali, David Stephen V, Fritz Jonathan B, and Shamma Shihab A. Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron*, 61:467–480, 2009.

[Chang *et al.* 2017] Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael J. Witbrock, Mark A. Hasegawa-Johnson, and Thomas S. Huang. Dilated recurrent neural networks. In *Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 76–86, 2017.

[Chen *et al.* 2015] Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. Long short-term memory neural networks for chinese word segmentation. In *EMNLP 2015*, pages 1197–1206, 2015.

[Emerson 2005] Thomas Emerson. The second international chinese word segmentation bakeoff. 2005.

[Jang *et al.* 2016] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *CoRR*, abs/1611.01144, 2016.

[Kalchbrenner *et al.* 2015] Nal Kalchbrenner, Ivo Danihelka, and Alex Graves. Grid long short-term memory. *CoRR*, abs/1507.01526, 2015.

[Kingma and Ba 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[Lecun *et al.* 2015] Y Lecun, Y Bengio, and G Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[Li and Roth 2002] 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.

[Maas *et al.* 2011] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[Pang and Lee 2005] 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.

[S.Hochreiter and J.Schmidhuber 1997] S.Hochreiter and J.Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Shwartz-Ziv and Tishby 2017] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[Socher *et al.* 2013] 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. pages 1631–1642, 2013.

[Sønderby *et al.* 2016] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

[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.

[Sutton *et al.* 1999] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1057–1063, 1999.

[Tishby and Zaslavsky 2015] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

[Tishby *et al.* 2000a] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *University of Illinois*, 411(29-30):368–377, 2000.

[Tishby *et al.* 2000b] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[van der Westhuizen and Lasenby 2018] Jos van der Westhuizen and Joan Lasenby. The unreasonable effectiveness of the forget gate. *CoRR*, abs/1804.04849, 2018.

[Vaswani *et al.* 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

[Weinberger *et al.* 2014] Norman M. Weinberger, John Ashe, and Jeanmarc Edeline. *LEARNING-INDUCED RECEPTIVE FIELD PLASTICITY IN THE AUDITORY CORTEX: SPECIFICITY OF INFORMATION STORAGE*. 2014.

[Xu *et al.* 2016] Jiacheng Xu, Danlu Chen, Xipeng Qiu, and Xuanjing Huang. Cached long short-term memory neural networks for document-level sentiment classification. In *EMNLP 2016*, pages 1660–1669, 2016.

[Yu *et al.* 2017] Adams Wei Yu, Hongrae Lee, and Quoc V. Le. Learning to skim text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1880–1890, 2017.