WAE_RN: Integrating Wasserstein Autoencoder and Relational Network for Text Sequence

Xinxin Zhang
Zhongyuan University of Technology
Zhengzhou, China
2018007088@zut.edu.cn

Xiaoming Liu
Zhongyuan University of Technology
Henan Key Laboratory on Public Opinion Intelligent Analysis ,Zhengzhou, China
ming616@zut.edu.cn

Guan Yang*
Zhongyuan University of Technology
Zhengzhou, China
yangguan@zut.edu.cn

Fangfang Li
oOh! Media
North Sydney, Australia
Fangfang.Li@oohmedia.com.au

Abstract

One challenge in Natural Language Processing (NLP) area is to learn semantic representation in different contexts. Recent works on pre-trained language model have received great attentions and have been proven as an effective technique. In spite of the success of pre-trained language model in many NLP tasks, the learned text representation only contains the correlation among the words in the sentence itself and ignores the implicit relationship between arbitrary tokens in the sequence. To address this problem, we focus on how to make our model effectively learn word representations that contain the relational information between any tokens of text sequences. In this paper, we propose to integrate the relational network(RN) into a Wasserstein autoencoder(WAE). Specifically, WAE and RN are used to better keep the semantic structurse and capture the relational information, respectively. Extensive experiments demonstrate that our proposed model achieves significant improvements over the traditional Seq2Seq baselines.

1 Introduction

Sequence problems are common in daily life that involves DNA sequencing in bioinformatics, time series prediction in Information science, and so on. NLP tasks, such as word segmentation, named entity recognition(NER), machine translation(MT), etc, are actually text sequence problems. For text sequence tasks, it is required to predict or generate target sequences based on the understanding of input source sequence, so it plays a pivotal role in NLP to deeply understand the generic knowledge representation in different context.

To learn the features of input sequences, probabilistic graphical models, such as Hidden Markov Models(HMM) and Conditional Random Field (CRF), can use manually defined feature functions to transform raw data into features, but the quality of the feature functions directly determines the quality of the data presentation.

Because deep learning can automatically learn the useful and highly abstract features of the data via artificial neural network(ANN), many researchers devoted themselves to using Neural Networks(NNs) to obtain low dimensional distributed representations of input data, especially in language modeling, using AutoEncoder(AE) (Rumelhart et al., 1988) to retain the text sequence semantic information in different context has shown promising results. These language models are pretrained on large-scale corpus and complex models to obtain the data representation which contains global information and has strong generalization ability, then the latent representation can be adapted to several contexts by fine-tuning them on various tasks. However, these models simply make use of word order information or position information and ignore the implicit relationship between arbitrary tokens in the sequence, resulting in learning inadequately hidden feature representations and obtaining only superficial semantic representation. More recently, studies on attention (Bahdanau et al., 2015; Luong et al., 2015) and self-attention(Klein and Nabi, 2019; Tan et al., 2018) mechanism demonstrate that it can effectively improve the performance of several NLP tasks by exchanging information between sentences. However, it only

*corresponding author

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1173-1183, Hainan, China, October 31 - Novermber 1, 2020.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

calculates the contribution between vectors by means of weighted sum without exploring and taking advantage of the implicit structural relationships among tokens.

In this work, we propose add relational networks(RN) (Santoro et al., 2017) to the Wasserstein AutoEncoder(WAE)(Kingma and Welling, 2014) on the basis of the Seq2Seq architecture to collect the complex relationship between objects and retain the semantic structure in sentences. Specifically, to keep the relational information and structural knowledge we add RN layer to encoder since RN integrates the relational reasoning structure that can constrain the functional form of neural network and capture the core common attributes of relational reasoning. To better capture the complex relationships and preserve the semantic structure we use WAE as our encoder because WAE maps input sequences into the wasserstein space that allows various other metric spaces to be embedded in it while preserving their original distance measurements.

The main contributions of our work can be summarized as follows:

1. We put forward an innovative idea to learn more meaningful and structural word representations in text sequneces. We consider relations between objects entail good flexibility and robustness, which are informative and helpful.

2. We propose a WAE_RN model, which integrates WAE and RN to obtain useful and generalized internal latent representations and the implicit relationships in the text sequencse.

3. We conducts experimental verification on two text sequence tasks named entity recognition and EN-GE machine translation. The experimental results demonstrate our proposed model can achive better semantic representation.

2 Related Work

2.1 AutoEncoder

Traditional AutoEncoder(AE) maps the high level characteristics of input data distribution in high dimension to the low(latent vector), and the decoder absorbs this low level representation and outputs the high level representation of the same data. Many researchers have been working on how to get better semantic representations of input sequences, methods using AE such as ELMo(Peters et al., 2018), BERT(Devlin et al., 2019), ALBERT(Lan et al., 2020), ERNIE(Zhang et al., 2019; Sun et al., 2020), XLNet(Yang et al., 2019), etc have been proven as effective techniques. Each model achieves the optimal effect at that time due to its own advantages, and their corresponding pre-trained word vector can still facilitate many downstream tasks even now. However, the latent representation learned by AE is encoded and decoded just in a deterministic way and with no constraint in the hidden space, resulting in a lack of diversity in encoding results, it was later followed by approaches based on VAE(Kingma and Welling, 2014; Bowman et al., 2016) and WAE(Tolstikhin et al., 2018).

VAE converts the potential representation obtained by the encoder into a probabilistic random variable and learn a smooth potential space representation, then the decoder reconstructs the input data and outputs the reconstructed original data. The results have shown that VAE performs competitively compared to traditional AutoEncoder, for example, (Zhang et al., 2016) attempts to use VAE for machine translation, which incorporate a continuous latent variable to model the underlying semantics of sentence pairs. (Shah and Barber, 2018) specifies the prior as a Gaussian mixture model and further develop a topic-guided variational autoencoder (TGVAE) model that is able to generate semantically-meaningful latent representation while generating sentences. However, training on VAE often leads to the disappearance of the KL term. In addition, VAE assumes that the latent variables follow a gaussian distribution, so only a gaussian encoder can be used. To solve these problems, VAE is replaced with WAE by researchers.

Wasserstein Autoencoder (WAE) use the Wasserstein distance that measures the distance between two distributions to replace the KL divergence in VAE to prevent the KL term from disappearing and help the encoder capture useful information during training. Besides, the goal of WAE is to minimize the direct distance between the marginal and the prior distribution and does not force the posterior of each sample to match the prior. In this way, different samples can keep a distance from other samples, which makes

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1173-1183, Hainan, China, October 31 - Novermber 1, 2020.

the results generated are more diverse. For instance, (Bahuleyan et al., 2019) propose a WAE variant that use an auxiliary loss to encourage the encoder more stochastic, their studies verified the WAE model achieves much better reconstruction performance. Moreover, (Wang and Wang, 2019) pointed out that the latent space is so complex that we only use standard Gaussian to assume the prior is not enough, and then they proposed to supplement some geometric properties of input space with Riemannian metric tensor to the latent space to learn more flexible latent distribution.

Furthermore, Warstam space is more flexible than Euclidean space, which is helpful for capturing the complex relationships and retaining the semantic structure. Since we focus on capturing the universal semantic representation, we choose WAE as our encoder to generate more meaningful and more flexible latent representation while maintaining the original semantic structure.

2.2 Relational Network

Because Recurrent Neural Network(RNN) gives an output for the input at each moment combined with the current model state, RNN-based model can only learn the sequence relation. While Convolutional Neural Network(CNN) continuously extracts local and overall features through a series of filters, so CNN-based model has poor ability to learn some transformation or relationship. To address this issue, there is a simple solution, that is adding some specific learning modules such as RN to help the model express and learn. RN is a neural network integrated with Relational reasoning structure, which aims to constrain the functional form of the neural network to capture the core common attributes of Relational reasoning. Almost all recent methods focus on using RN to capture the relationships between objects. For example, (Zhang et al., 2018) introduce RN to learn better representations of the input data and experiments on machine translation demonstrate RN can help retain relationships between words. (Chen et al., 2019) also use RN to capture the dependencies within a sentence between any two words and verify the effectiveness of their proposed method on two benchmark NER datasets, which all support that the RN can model relations between the input sequences.

Inspired by the success of the RN in learning the relationships between elements, in this paper, we directly incorporate RN into the WAE models, thus to fully learn the semantic representation and keep the relational information and structural knowledge between sequences to the greatest extent.

3 Preliminary

Since the purpose of the proposed method is to better obtain the semantic representation of text sequences, we will focus on the following two issues.

3.1 The Problem of Sequence Prediction

Sequence prediction is the most basic and widely used task, such as word segmentation, part-of-speech(POS) tagging, named entity recognition(NER), dependency analysis, etc. Essentially, it can be viewed as a matter of classifying each element in a linear sequence according to its context representation. That is, after understanding the input sequence and extracting its useful information, the optimal mark is made for each sequence, and then a set of globally optimal marks is selected for a given sequence at one time.

Suppose we have an input sequence $\vec{x}$ of L elements, and a tag sequence $\vec{y}$ of the same length, i.e. $\vec{x} = (x_1, x_2, \ldots, x_L)^T$, $\vec{y} = (y_1, y_2, \ldots, y_L)^T$, where x_i represents the i-th sequence and y_j represents the j-th tag, it's also requires that the value of y_j is taken from a predefined set of finite tags and i equals j, the final goal is to assign a globally optimal label y_j for each input sequence x_i. End-to-end learning is directly modeling conditional probabilities $p(y|x)$ and then map the input sequence $x_1, x_2, \ldots, x_L$ to the output sequence $y_1, y_2, \ldots, y_L$, i.e.(1).

$$Y = (y_1, y_2, \ldots y_L) = \underset{y}{\operatorname{argmax}}\, p(y|x, \theta) \tag{1}$$

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1173-1183, Hainan, China, October 31 - Novermber 1, 2020.

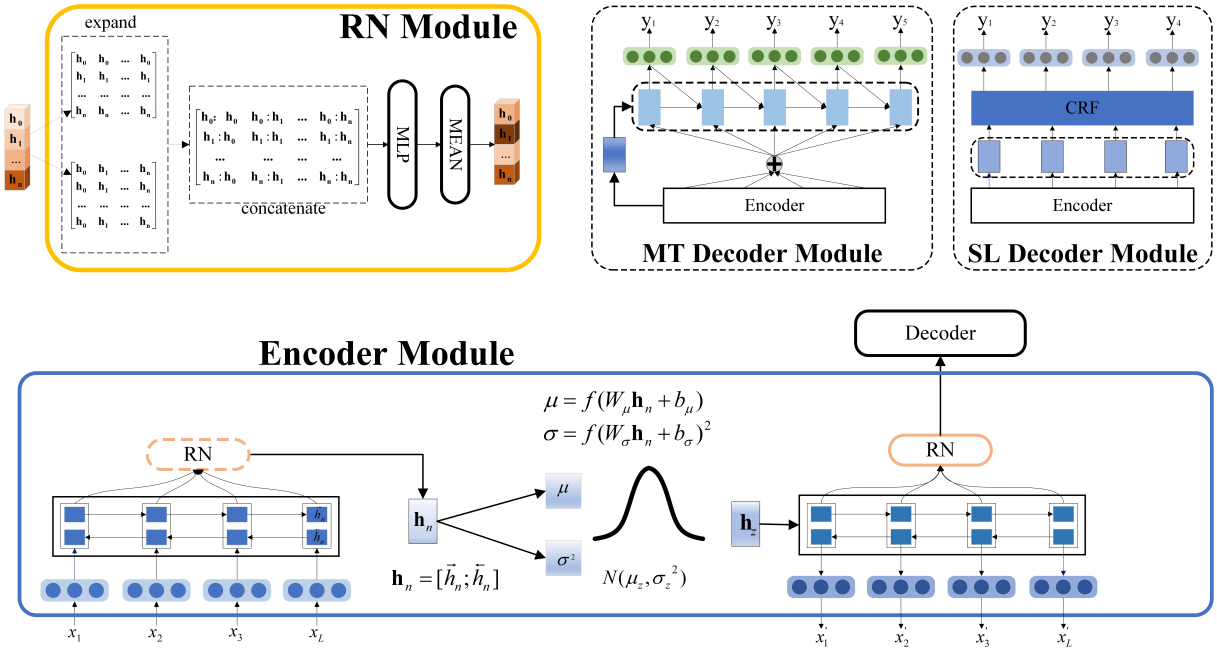

Figure 1: Model architecture

3.2 The Problem of Sequence Generation

Sequence generation is translating the dataset into a clear narrative of human understanding based on the real understanding of text content, such as machine translation, dialogue generation, abstract generation and so on. We usually decompose the generation probability into the product of the generation probability of context-related subsequence, and then use the method of auto-regression to get the text in the form of natural language that human can understand.

Suppose the input sequence is $\vec{x}$, the goal is to understand the input sequence and generate the corresponding output sequence $\vec{y}$, i.e. $\vec{x} = \left(x_1, x_2, \ldots x_{|X|}\right)^T$, $\vec{y} = \left(y_1, y_2, \ldots y_{|Y|}\right)^T$, where $|X|$ and $|Y|$ correspond to the length of input sequence and output sequence respectively. Different from sequence prediction, the purpose of sequence-to-sequence learning is to model the conditional probability $p(y|x)$ with all the sequences before the current sequence as the condition, and then map the input sequence to an output sequence, i.e. (2).

$$Y = \left(y_1, y_2, \ldots y_{|Y|}\right) = \underset{y}{\operatorname{argmax}}\, p(y|x;\theta) = y \arg\max \left(\prod_{i=1}^{|Y|} p\left(y_i|x, y_{<i};\theta\right) \right) \tag{2}$$

4 Relational Network based WAE Model

4.1 Architecture of Proposed

In order to obtain universal semantic representations that contain structured knowledge, we propose a Relational Network based Wasserstein AutoEncoder (WAE_RN) model, which have the ability to embed the potential structural information contained in sequence into semantic representation. Specifically, a relation network layer is employed to quantify the potential relationships between any two elements in the input sequence, and then these relationships are embedded into the input sequence by WAE to get semantic representation that contains relational information. Finally, the generic representation is sent to different decoders to perform different downstream tasks. Next, we will elaborate our proposed model in detail.

4.2 The Wasserstein AutoEncoder Layer

As shown in the bottom of Fig. 1, the first encoder of WAE collects the semantic information of the data, and the RN module learns the relational information between the outputs of RNNs, then the context representation is mapped to the Wasserstein space. Compared with embedding data into Euclidean space, which is the most common method, WAE embeds the input data into the Wasserstein space as a probability distribution to can help us capture the complex relationship and retain the semantic structure, so we can obtain the distribution $\mathbf{h}_n = \left[\vec{h}_n; \vec{h}_n\right]$ that covers both semantic and relational information of input data $x_1, x_2, \ldots, x_L$. Note that the relational network module can be placed either in front of or behind the first encoder, our experiments showed that it is better for the named entity recognition task to put it in the front while for the machine translation task to put it in the back.

After reparameterizing, the reconstructed hidden state $\mathbf{h}_z = N\left(\mu_z, \sigma_z^2\right)$(where $\mu_z = f(W_\mu \mathbf{h}_n + b_\mu), \sigma_z^2 = f(W_\sigma \mathbf{h}_n + b_\sigma^2)$) is sent to the second encoder in WAE as its initial state, after that this encoder relearns the latent representation of input data under the guidance of the hidden state obtained in the previous step, so as to obtain the semantic representation that both follows the source semantic information and retains the structured information. To fully exploit the relational information, we send the representation learned by the second encoder into the relationship network again.

Different from VAE, WAE can use both Gaussian encoder and deterministic encoder. Besides, the goal of Wasserstein distance is to minimize the direct distance between the marginal distribution and the prior, without forcing the posterior of each sample to match the prior, so that different samples can keep a distance from other samples to produce more diverse results.

4.3 The Relational Layer

The architecture of our RN module is shown in the upper left corner of Fig. 1, different from (Zhang et al., 2018), our RN doesn't use the CNN layer. Besides, to keep the original information of the input sequence to the great extent, we don't use any nonlinear transformations, keeping the dimensions the same. To learn the implicit internal relation between any two elements, we use some transformation between tensors to make objects fully connected and associated with each other, which means, for any vector $C = (\vec{c}_1, \vec{c}_2, \ldots \vec{c}_n)$, after concatenating, its each element $\mathbf{c}_{i,j} = [\vec{c}_i; \vec{c}_j]$. Then we directly calculate the relationships between any objects: $RN(\mathbf{o}_{i,j}) = f_\phi\left(W_{MLP}\mathbf{c}_{i,j} + b_{MLP}\right)$. Here, a multi-layer perceptron is used for f_ϕ to find the relationship between all pairwise objects and judge whether and how they are related.

4.4 The Prediction Layer

There is no difference between the decoder used in our model and the traditional decoder. As shown in the upper right corner of Fig. 1, for machine translation tasks, the decoder is the ordinary RNNs with beam search layer, which generates target sequences one by one in an auto-regressive way, while for the named entity recognition task, the decoder is the RNN network with the CRF layer.

4.5 The Objective

For AE, the training objective is the cross-entropy loss or the reconstruction loss, given by $J_{rec}(\theta, \phi, x) = E_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]$. In order to compute the loss of our model, we use MMD (given as $MMD = \left\|\int k(z,;)dp(z) - \int k(z,\cdot)dq(z)\right\|_{\mathrm{H}_k}$) to approximate Wasserstein distance, where H_k refers to the Hilbert space defined by the kernel k, for high dimensional Gaussian function, k was usually chosen as the inverse quadratic kernel : $k(x, y) = \frac{C}{C + \|x-y\|_2^2}$.

$$L(\theta; \phi; x) = E_{q(x)}\left[J_{rec}(\theta, x) + \alpha J_{task}(\Phi, x)\right] + \beta MMD \tag{3}$$

Thus the loss function(3) of our model consists of three terms: the first is the reconstruction loss, which encourages the encoder to learn to reconstruct data; the second is the Wasserstein distance between the distribution of the encoder $q_\theta(z|x)$ and prior $p(z)$ (usually p is $N(0, 1)$), which measures how much information is lost when q is represented by p; the third is the task loss between the source input $x_1, x_2, \ldots x_{|X|}$ and the generated target sequences $y_1, y_2, \ldots y_{|Y|}$. However, in the experiment we

observe that the reconstruct loss has a great influence on the results of our model, resulting in poor performance. To address this problem, we impose a weight α(here α is 2) on the translation loss to balance the influence between the task loss and the reconstruct loss. To achieve better performance, we also give another weight β(here β is 0.0001) onMMD. To the end, our model can be trained in an end-to-end manner by minimizing (3).

5 Experiments

In this section, we aim to investigate our model's performance over NER and MT, where NER belongs to the problem of sequence prediction and MT belongs to the problem of sequence generation. We first present our experimental set up, then compare our method to other baseline systems, finally we give some analyses about our method.

5.1 Datasets

We use two benchmark datasets: OntoNotes5.0 Chinese NER dataset(OntoNotes5.0 Ch-NER) and IWSLT2014 German-English dataset(IWSLT14en-de) for evaluation, the details about these corpora are shown in Table1.

Table 1: Statistics of OntoNotes5.0 Ch-NER and IWSLT2014en-de

Dataset	Type	Train	Valid	Test
OntoNotes5.0 Ch-NER	Sentences	53.5k	12.8k	4.5k
	Chars	750k	110k	90k
	Entities	62.5k	9.1k	7.5k
IWSLT2014en-de	Sentences	150k	6.9k	6.7k

5.1.1 OntoNotes5.0 Ch-NER

OntoNotes5.0 Ch-NER contains eleven different entity name types(such as PERSON, NORP, GPE, etc.) and seven different value types(DATE, TIME, MONEY, etc.). We use the same OntoNotes data split used for co-reference resolution in the CoNLL-2012 shared task (Pradhan et al., 2012) and convert the IOB boundary encoding to BIO tagging scheme(B, I, O). We preprocess by filtering out char-level sentences longer than 150 words and replacing all words that appear less than three times with an $< unk >$ token, but for testing data, we use the original dataset.

5.1.2 IWSLT14en-de

IWSLT14en-de contains transcripts of TED talks and translate between German and English in both directions. Following previous works, we use the same data cleanup as (Ranzato et al., 2016). We apply the same tokenization and truecasing using standard Moses scriptsto both our model and baseline. For training data, sentences longer than 50 tokens were chopped and rared words were replaced by a special $< unk >$ token, for testing data, we also use the original version of testing files.

5.2 Experimental Setting

For NER task, we use strong bidirectional Long Short Term Memory with CRF(Bi-LSTM-CRF) baseline, but for MT the baseline is a standard implementation of Bi-LSTM seq2seq model with dot-product attention (Bahdanau et al., 2015; Luong et al., 2015) and for decoding we use a beam width of 10 and limit the max sequence length to 100. Detail hyper-parameters can be found in Table2.

For NER task, we use the entity level accuracy rate, recall rate and F1 value to calculate the score and report standard F1-score for CoNLL NER tasks (Pradhan et al., 2012). For MT task, we adopt BLEU for translation quality evaluation and calculate the BLEU scores on test set using Moses *multi-bleu.perl* script.

Table 2: Hyper-Parameter Settings

Learning rate	$1e^{-3}$
Learning rate decay	0.5
Batch size	64
Clip norm	5.0
Embedding dim	256
Hidden dim	256
Latent dim	32
Dropout	0.3
Uniform init	0.1
Patience	20

Table 3: Corpus BLEU scores (%) on IWSLT14en-de translation tasks

	IWSLT14Ge − En(BLEU)	IWSLT14En − Ge(BLEU)
2017RaphaelShu	29.56	-
2018PoSenHuang	30.08	25.36
2019BryanEikema	28.0	23.4
Ours		
RNN_attn(baseline)	27.84	23.74
RNN_attn_RN	28.18(0.3 ↑)	23.95(0.2 ↑)
WAE(d)_attn	28.55(0.7 ↑)	24.24(0.5 ↑)
WAE(d)_attn_RN	**28.87**(**0.9** ↑)	**24.46**(**0.7** ↑)

5.3 Results and Analysis

In order to enhance the fairness of the comparisons and verify the solidity of our improvement, we train 5 times with random uniform distribution initialization and report average results of our proposed model as well as our re-implemented baselines. Note that we just use simple Seq2Seq architecture as our baseline and don't add any other methods(such as label smoothing, tied embedding, BPE, pre-trained word vector, etc) to the baseline, because our goal is to demonstrate that our proposed method can yield a more general semantic representation, rather than further boost performance.

5.3.1 Results on Machine Translation

For IWSLT14en-de translation tasks, we use deterministic encoder rather than Gaussian encoder for largely alleviating the training difficulties. We show the test results of different models in Table3.

The former lines in the table list the performance of previous methods. (Shu and Nakayama, 2018) propose compress word embedding to directly learn the discrete codes via deep compositional code learning, improving the BLEU scores from 29.45% to 29.56%. Using SleepWAke Networks (SWAN) that is a segmentation-based sequence modeling method to explicitly model the phrase structure in output sequences, (Huang et al., 2018) achieves the state-of-the-art results at that time. (Eikema and Aziz, 2019) use Auto-Encoding Variational NMT model to generate source and target sentences jointly from a shared latent representation, achieving de→en and en→de BLEU scores of 28.0% and 23.4% respectively.

The latter lines show the performance of ours, we can see that our proposed WAE_RN model achieves significant improvement over the baseline system. It demonstrates that our model can capture more useful information and improve the performance of NMT system. In particular, our proposed model outperforms the baseline by 0.9% BLEU points, while only use RN and DAE improves the baseline 0.3% and 0.7% respectively, which effectively illustrate that the combine of RN and WAE can both collect the complex relationship and retain the semantic structure between objects.

Table 4: The evaluation results on OntoNotes5.0 Chinese NER task

method	P(%)	R(%)	F
CoNLL2012	78.20	66.45	71.85
Ours			
BiLSTM_CRF (baseline)	73.08	69.20	71.08
BiLSTM_CRF_SelfAttn	70.93	67.10	68.96
BiLSTM_CRF_LM	72.69	69.59	71.11
BiLSTM_CRF_RN	73.43	69.71	71.52
WAE_CRF	72.79	69.61	71.17
WAE_CRF_RN (best)	73.09	**70.76**	**71.90(0.8** ↑)

Figure 2: Performance of our model and baseline on each category.

5.3.2 Results on Sequence Labeling

For OntoNotes5.0 Chinese NER task, we use Gaussian encoder. As shown in Table4, the first results is from the CoNLL-2012 Shared Task (Pradhan et al., 2013) and the others are ours, we can observe that WAE_RN can significantly outperforms our re-implemented baseline by 0.8, which demonstrates the robustness of our models. As depicted in Fig. 2, we can see that our method performs well on most categories, such as 'ORDINAL', 'NORP', 'LANGUAGE', etc, and slightly below baseline on the categories of 'PERSON', 'ORG' and 'TIME'. It also should be noted that our model can't find the entity named 'PRODUCT', which is the smallest number of entities in the training dataset. From the results, we can observe that our proposed model does have a positive impact on learning word representation.

Besides, we also conduct experiments using different models to explain the the performance promotion of each module, experimental results on NER task confirm the effectiveness of our proposed model, similar as shown in MT tasks.

6 Conclusion

This paper presents a WAE_RN model for text sequence tasks, which aims at learning word representatiosn containing structured knowledge. To be specific, to preserve the semantic structure between objects, we propose use WAE as the model's encoder. To capture the core common attributes of relational reasoning, we introduce RN. Both of which combine well to learn the generic representation that contains relational information. Experimental results on MT and NER tasks demonstrate that the proposed model leads to significant improvements. In the future, we plan to extend the general representation to transfer

learning.

Acknowledgements

This work was supported by the Science and Technology Planning Project of Henan Province of China(Grant No. 182102210513 and 182102310945) and the National Natural Science Foundation of China(Grant No.61672361 and 61772020).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hareesh Bahuleyan, Lili Mou, Hao Zhou, and Olga Vechtomova. 2019. Stochastic wasserstein autoencoder for probabilistic sentence generation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4068–4076. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In Yoav Goldberg and Stefan Riezler, editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21. ACL.

Hui Chen, Zijia Lin, Guiguang Ding, Jianguang Lou, Yusen Zhang, and Börje Karlsson. 2019. GRN: gated relation network to enhance convolutional neural network for named entity recognition. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6236–6243. AAAI Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2019. Auto-encoding variational neural machine translation. In Isabelle Augenstein, Spandana Gella, Sebastian Ruder, Katharina Kann, Burcu Can, Johannes Welbl, Alexis Conneau, Xiang Ren, and Marek Rei, editors, *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019*, pages 124–141. Association for Computational Linguistics.

Po-Sen Huang, Chong Wang, Sitao Huang, Dengyong Zhou, and Li Deng. 2018. Towards neural phrase-based machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4831–4836. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue, editors, *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012, July 13, 2012, Jeju Island, Korea*, pages 1–40. ACL.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In Julia Hockenmaier and Sebastian Riedel, editors, *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 143–152. ACL.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1988. Learning representations by back-propagating errors. *Nature*, 323(6088):696–699.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4967–4976.

Harshil Shah and David Barber. 2018. Generative neural machine translation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 1353–1362.

Raphael Shu and Hideki Nakayama. 2018. Compressing word embeddings via deep compositional code learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8968–8975. AAAI Press.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4929–4936. AAAI Press.

Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. 2018. Wasserstein auto-encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Prince Zizhuang Wang and William Yang Wang. 2019. Riemannian normalizing flow on variational wasserstein autoencoder for text modeling. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 284–294. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 521–530. The Association for Computational Linguistics.

Wen Zhang, Jiawei Hu, Yang Feng, and Qun Liu. 2018. Refining source representations with relation networks for neural machine translation. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1292–1303. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.