INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Latent Topic Attention for Domain Classification

Peisong Huang, Peijie Huang[*]*, Wencheng Ai, Jiande Ding, Jinchuan Zhang*

College of Mathematics and Informatics, South China Agricultural University, China

bringtree@qq.com, pjhuang@scau.edu.cn, nhnvnw@qq.com,
{bighead, jczhang}@stu.scau.edu.cn

Abstract

Attention-based bidirectional long short-term network (BiLSTM) models have recently shown promising results in text classification tasks. However, when the amount of training data is restricted, or the distribution of the test data is quite different from the training data, some potential informative words maybe hard to be captured in training. In this work, we propose a new method to learn attention mechanism for domain classification. Unlike the past attention mechanisms only guided by domain tags of training data, we explore using the latent topics in the data set to learn topic attention, and employ it for BiLSTM. Experiments on the SMP-ECDT benchmark corpus show that the proposed latent topic attention mechanism outperforms the state-of-the-art soft and hard attention mechanisms in domain classification. Moreover, experiment result shows that the proposed method can be trained with additional unlabeled data and further improve the domain classification performance.

Index Terms: domain classification, attention mechanism, latent topic, BiLSTM, BTM

1. Introduction

Spoken language understanding (SLU) applications are becoming increasingly important in our daily lives [1]. In using of human-computer dialogue based applications, human may have various intent, for example, chit-chatting, asking questions, booking air tickets, inquiring weather, etc. Therefore, after receiving an input message (text or automatic speech recognition result) from a user, the first step is to classify the user utterance into a specific domain for further processing. For example, a user asking the chatbot for "今天福州的天气怎么样? (How is the weather in Fuzhou today?)" should have his utterance classified as weather-query so that the query can be routed to the correct natural understanding subsystem. Domain classification can be treated as a semantic utterance classification problem, and popular classifiers like support vector machines (SVM) [2] and deep neural network methods [3] can be applied. Especially, the recurrent neural network (RNN) model has been successfully applied in many sequence learning problems [4], as it can be easy to deal with variable-length input utterances.

Recent approaches introduce attention mechanisms to focus the models on informative words [5–7]. However, to our knowledge, the parameter vectors of the attention in all of the existing attention approaches are trained with the domain tags. When the amount of training data is limited, or the distribution of the test data is quite different from the training data, some potential informative words maybe hard to be captured in training. Unlike the guidance of the domain tags, the focus of this work is to improve the accuracy of domain classification by exploiting a new attention mechanism using the latent topic.

Modeling an utterance as a mixture of latent topics is a valuable way to infer semantics in an unsupervised manner. Several conventional topic modeling techniques such as latent semantic analysis [8], probabilistic latent semantic analysis (pLSA), latent Dirichlet allocation (LDA) [9] and biterm topic model (BTM) [10] have been used to good success in inferring the high level meaning of documents through a set of representative word (topics) [11]. In text classification, many researches adopt topic model to mine latent topics as classification features [12-14]. Different from these studies, we use topic model to calculate the attention and employ it for bidirectional long short-term memory network (BiLSTM). The experiments on the SMP-ECDT benchmark corpus show the well performance of the proposed attention mechanism.

The remainder of the paper is organized as follows. In Section 2, we introduce the related prior work including attention mechanisms and biterm topic model. In Section 3, we describe the proposed method. Section 4 discusses the experiment setup and results on SMP-ECDT benchmarking task. Section 5 concludes the work.

2. Relation to prior work

2.1. Attention mechanisms

Recently, various attention mechanisms have been previously studied in text classification problems. Here, we consider two popular alternatives, soft attention [5-7] and hard attention [15].

Soft attention was widely used in text classification, such as intent detection [5], relation classification [6] and document classification [7]. Although there exists a little difference in the computations about the functions of score, the attention mechanism is to calculate the "soft" aligned attention weight from the output of the encoder, and then the output of the encoder is scaled according to the weight of attention:

$$\mathrm{u_i} = score(h_i), \tag{1}$$

$$\alpha_\mathrm{i} = \frac{\exp(u_i)}{\sum_i^L \exp(u_i)}, \tag{2}$$

$$h^* = \sum_i^L \alpha_i \, h_i, \tag{3}$$

where *score* is a feed-forward neural network. We first feed the LSTM outputs h_i through a multi-layer perceptron (MLP) to get u_i as a hidden representation of h_i. Then the h^* is computed as a weighted sum of the LSTM outputs $(h_1, h_2, \ldots, h_L)$ based on the weights $(a_1, a_2, \ldots, a_L)$.

Another variant of attention mechanism is hard attention [15] that aligns each output to exactly one input state but

[*] Corresponding author.

http://dx.doi.org/10.21437/Interspeech.2019-2228

requires intricate training to teach the network to choose that state:

$$p(s_i = 1|\alpha) = \alpha_i, \quad (4)$$

$$h^* = \sum_i^L s_i \alpha_i h_i, \quad (5)$$

where α_i is the probability that the LSTM output of i^{th} time step is selected. s is an one-hot vector. When the LSTM output of i^{th} time step is selected, the corresponding s_i is set to 1, and others are set to 0.

Although the abovementioned studies have shown that the existing attention mechanisms empirically perform well, all of them are trained with the domain tags. When the amount of the labeled training data is restricted, some potential informative words maybe hard to be captured in the existing attention mechanisms.

2.2. Biterm topic model

Biterm topic model (BTM) is introduced by Yan et al. [10]. Yan et al. had used BTM model to train some text features and use the text features to do text classification. Figure 1 shows the graphical representation of BTM.

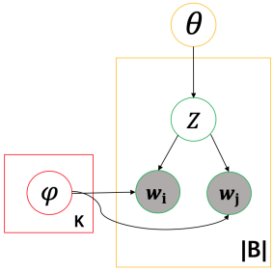


Figure 1: *Graphical representation of BTM.*

The key idea of BTM is to learn topics over short text. Suppose β *and* γ are the Dirichlet priors. The specific generative process of the corpus in BTM can be described as follows:

1. For each topic z, draw a topic-specific word distribution $\phi_z \sim Dir(\beta)$.
2. Draw a topic distribution $\theta \sim Dir(\gamma)$ for the whole collection.
3. For each biterm b in the biterm set B:
   (a) draw a topic assignment $z \sim Multi(\theta)$,
   (b) draw two words: $w_i, w_j \sim Multi(\phi_z)$.

Following the above procedure, the joint probability of a biterm $b = (w_i, w_j)$ can be written as:

$$P(b) = \sum_z P(z)P(w_i|z)P(w_j|z) = \sum_z \theta_z \phi_{i|z} \phi_{j|z}. \quad (6)$$

Thus the likelihood of the whole corpus is:

$$P(B) = \prod_{(i,j)} \sum_z \theta_z \phi_{i|z} \phi_{j|z}. \quad (7)$$

By using BTM, we can get the proportions of the word under the topic. Moreover, BTM can make full use of data that is not labeled because it is an unsupervised model. In our proposed method, we make use of the characteristics of BTM and import BTM into attention mechanism.

3. Proposed method

In this section, we first describe the approach on latent topic attention-based BiLSTM for domain classification. Following that, we describe the proposed method on inferring latent topics to affect the attention weight directly. To the best of our knowledge, using the topic model as an attention mechanism has not yet been studied.

3.1. Latent topic attention for domain classification

We modify the model for domain classification based on [6] and incorporate a new attention mechanism using BTM, namely *Latent topic att*. As shown in Figure 2, it consists of five layers.

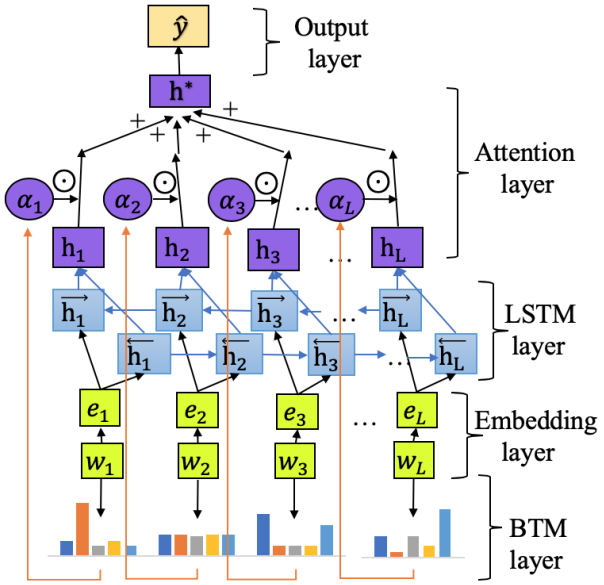


Figure 2: *BiLSTM model with latent topic attention.*

In the embedding layer, Given a utterance consisting of L words $U = \{x_1, x_2, \dots, x_L\}$, every word x_i is converted into a real-valued vector e_i. For each word in U, we first look up the embedding matrix $W^{wrd} \in \mathbb{R}^{d^w|V|}$, where V is a fixed-sized vocabulary, and d^w is the size of word embedding. The matrix W^{wrd} is a parameter to be learned, and d^w is a hyper-parameter to be chosen by users. We transform a word x_i into its word embedding e_i by using the matrix-vector product:

$$e_i = W^{wrd} v^i, \quad (8)$$

where v^i is a vector of size $|V|$ which has value 1 at index e_i and 0 in all other positions. Then the utterance is feed into the next layer as a real-valued vectors $emb_U = \{e_1, e_2, \dots, e_L\}$.

In the LSTM layer, we use a bidirectional RNN [16]. Bidirectional RNN has been successfully applied in speech recognition [17] and spoken language understanding [1]. We use LSTM [18] as the basic recurrent network unit for its ability to better model long-term dependencies comparing to simple RNN. We concatenation of the forward state $\overrightarrow{h_i}$ and backward state $\overleftarrow{h_i}$.

$$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]. \quad (9)$$

In the attention layer, we will use a function to calculate the attention weight α_i for the h_i and we introduce the calculation of α from BTM layer in Subsection 3.2. Then the representation h^* of the utterance is formed by a weighted sum of these output vectors.

$$h^* = \sum_{i=1}^{L} \alpha_i h_i. \quad (10)$$

In the output layer, we use a softmax classifier to predict label $\hat{y}$ from a discrete set of domain classes Y for an utterance U. The classifier takes the hidden state h*as input:

$$\hat{y} = argmax_y \, \text{softmax}\left(W^{(U)} h^* + b^{(U)}\right). \quad (11)$$

The cost function is negative log-likelihood of the true domain class labels $\hat{y}$:

$$\mathcal{J}(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \, log(y_i), \quad (12)$$

where $t \in \Re^m$ is the one-hot represented ground truth and $y \in \Re^m$is the estimated probability for each domain class by softmax (m is the number of target domain classes).

3.2. Latent topic attention mechanism

The idea of the latent topic attention mechanism is mining latent topic information to guide the distribution of attention weight. The words that significantly differ in various topics are captured as informative words that need more attention:

Firstly, we can easily estimate the topic-word distributions in BTM [10]:

$$\phi_{w|z} = \frac{n_{w|z}+\delta}{\sum_w n_{w|z}+M\delta}, \quad (13)$$

where M is the total number of words, n is the times of word w assigned to the topic z, and δ is the hyperparameter.

Then, by word proportions of the topic, one can get the correlation c_{ij} between topic z_j and word w_i.

$$c_{ij} = f\left(\phi_{w_i|z_j}\right). \quad (14)$$

In the experiments of this paper, we sort the words in every topic by the value of $\phi_{w_i|z_j}$ and then set the value of c_{ij} by the ranking of word w_i in the topic z_j.

Afterward, for each word, we will calculate the difference in different topics based on their ranking under different topics.

$$d_i = g\left(c_{i1}, c_{i2}, \ldots, c_{ij}\right), \quad (15)$$

where c_{ij} is the correlation value between the j^{th} topic z_j and the i^{th}word w_i. d_i means the topic difference of the i^{th} word.

In the experiments, $g\left(c_{i1}, c_{i2}, \ldots, c_{ij}\right)$ in Eq. 15 is set to $max\left(c_{i1}, c_{i2}, \ldots, c_{ij}\right) - min(c_{i1}, c_{i2}, \ldots, c_{ij})$.

Finally, the attention weight α_i of the LSTM layer output at the i^{th} time is:

$$\alpha_1, \alpha_2, \ldots, \alpha_i = softmax\left(d_1, d_2, \ldots, d_i\right). \quad (16)$$

4. Experiments

4.1. Dataset

In this paper, we are particularly interested in considering classifying the user domain in a single utterance and we chose the benchmark corpus of SMP-ECDT [19] provided by the iFLYTEK Co. Ltd. to evaluate the proposed attention mechanism. SMP-ECDT (**S**ocial **M**edia **P**rocessing - the **E**valuation of **C**hinese Human-Computer **D**ialogue **T**echnology) 2018 is the second evaluation of Chinese human-computer dialogue technology, and subtask 1 is for Chinese utterance domain classification. The benchmark corpus consists of the two top categories *chit-chat* and *task-oriented*. Meanwhile, the *task-oriented* dialogue also includes 30 sub-categories, making this a 31-category classification task. This corpus contains 3736 training data and 4528 test data items.

4.2. Experimental setup

We cut the training dataset into 10 parts for cross validation, and train the model to minimize the categorical cross-entropy loss and choose the best arguments using the grid search. The range of the LSTM num_units is [50, 100, 150, 300], the range of the input keep prob is [0.2, 0.3], and the range of the state keep prob is [0.2, 0.3]. The output keep prob is set to 0.3. The topic numbers of BTM are set as 5, 10, 20, 30, and 50 for selection. The batch training size is 32. In addition, we use the pre-train word2vec word vectors [20] and apply the Adam optimization method following the suggested parameter setup in [21]. All data shown in the following results are the average of five independent experiments.

4.3. Results and analysis

4.3.1. Topic number selection

We compare the validation of our proposed *Latent topic att* model in different number of topics, shown in Table 1. Results show that *Latent topic att* model gets the best validation performance with 20 topics, which is a moderate topic number.

Table 1: *Comparison of validation accuracy of the proposed model using different number of topics.*

Number of topics	Validation accuracy
5	92.17%
10	92.57%
20	**93.34%**
30	92.39%
50	92.28%

4.3.2. Overall performance

We then compare the proposed *Latent topic att* model with several baseline methods, including *BiLSTM* and some state-of-the-art attention mechanisms. The experimental results are shown in Table 2.

Table 2: *Performance of different models.*

Models	Accuracy
BiLSTM	76.40%
Soft att [5]	78.62%
Hard att [15]	78.25%
Latent topic att	**79.09%**

BiLSTM: *BiLSTM* is our primary baseline in domain classification, which achieves an accuracy of 76.40%.

Soft att: Liu et al. [5] employed the network for intent direction. We installed his attention mechanism on the baseline model for domain classification. This attention mechanism gets 78.62% accuracy.

Hard att: We install the hard attention mechanism [15] on the baseline model for domain classification. This attention mechanism gets 78.25% accuracy.

Our proposed *Latent topic att* model yields an accuracy of 79.09%. It outperforms all of the competing approaches.

4.3.3. Effect of the proposed model with additional unlabeled data

Furthermore, since BTM is an unsupervised model, we turn to examine if the proposed method can be trained with additional unlabeled data and further improve the domain classification performance. We cut 4/5 test data and use it as additional unlabeled data for *Latent topic att* model. The remainder 1/5 of the test data is used for testing. The experimental results are shown in Table 3.

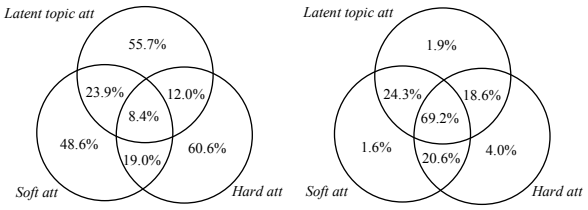
Table 3: *Performance of Latent topic att model with additional unlabeled data.*

Models	Accuracy
Latent topic att (only training data)	79.29%
Latent topic att (training data+4/5 test data)	**79.90%**

As the result shown in Table 3, trained only with training data, *Latent topic att* model gets the accuracy of 79.29% on 1/5 test data, which is approximate with that on the whole test data (79.09% in Table 2). When trained with additional unlabeled data (i.e. 4/5 test data), *Latent topic att* model gets 1% further improvement of the test accuracy on 1/5 test data.

4.4. Attention intersection

We calculate the percentage of the utterances whose top k attention words in different attention mechanisms have intersection or just belong to a certain attention mechanism. The results are shown in Figure 3 (a) and (b) for k=1 and k=3, respectively.



(a) Top 1 attention word (b) Top 3 attention words

Figure 3: *Illustration of attention intersection.*

As we can see from Figure 3 (a), when using *Latent topic att* model, the top 1 attention word of 55.7% utterances is different from *soft att* and *hard att* models, which indicates the distinctiveness of the *Latent topic att* model. And as shown in Figure 3 (b), in as much as 69.2% utterances, the top 3 attention words of the three attention mechanisms have intersection, which, in a certain sense, shows the effectiveness of the *Latent topic att* model.

4.5. Visualization of attention

The statistical results of attention intersection already show the effectiveness and distinctiveness of our models. In this case study, we further investigate the attention outputs of our model and the competing attention mechanisms. An example utterance (labeled as *calc* domain) is chosen from the test data for illustrating the effectiveness of our model in capturing latent informative keywords. We visualize the attention layer of it in Figure 4.

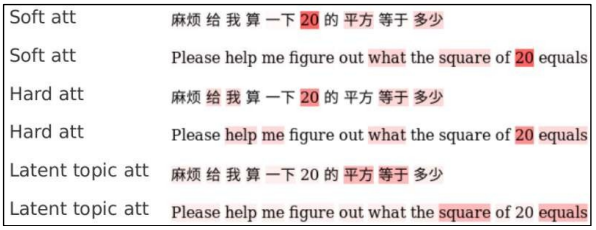


Figure 4: *Comparison of attention in different models.*

As shown in Figure 4, both of *soft att* and *hard att* models are directed to an irrelevant word "20", while our *Latent topic att* model can successfully capture the informative keywords, i.e. "平方 (the square)" and "等于 (equals)". In this example, we think the reason why the competing models cannot pay enough attention to the keywords is that the attention in their models is directed by the domain to learning. When the amount of training data is limited, it is easy for irrelevant words to help the model classification accidentally and they are convinced by the model that they are keywords. When training with the domain tags, there are not substantial data about *calc* domain to tell the model what it should pay more attention. In our model, we find that both of "平方 (the square)" and "等于 (equals)" have significant ranking differences (Eq. 15) in different latent topics, and thus are accurately assigned with high attention weight, which helps the classifier make the right prediction in this example.

5. Conclusions

In this paper, we propose a latent topic attention mechanism for classifying utterance. Our attention mechanism uses BTM to find more latent informative words and calculate the attention. Experimental results demonstrate that our attention mechanism outperforms the competing models. Moreover, experiment result shows that the proposed method can be trained with additional unlabeled data and further improve the domain classification performance. In addition, the statistical results of attention intersection show the effectiveness and distinctiveness of our model. And visualization of the attention layers illustrates that our attention mechanism is effective in capturing latent informative keywords.

6. Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 71472068), National Innovation Training Project for College Students of China (No. 201710564154), and Innovation Training Project for College Students of Guangdong Province (No. 201810564094). We also thank the SCIR Lab of Harbin Institute of Technology and the iFLYTEK Co. Ltd. for providing the SMP-ECDT benchmark corpus.

7. References

[1] P. Xu and R. Sarikaya, "Contextual domain classification in spoken language understanding systems using recurrent neural network," in *ICASSP 2014 - 39th IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, Proceedings*, 2014, pp. 136–140.

[2] P. Haffner, G. Tur, and J. H. Wright, "Optimizing SVMs for complex call classification," in *ICASSP 2003 - 28th IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, Proceedings*, 2003, pp. I-632-I–635.

[3] R. Sarikaya, G. E. Hinton, and B. Ramabhadran, "Deep belief nets for natural language call-routing," in *ICASSP 2011 - 36th IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, Proceedings,* 2011, pp. 5680–5683.

[4] S. Ravuri and A. Stoicke, "A comparative study of neural network models for lexical intent classification," in *ASRU 2015 - 5th IEEE Workshop on Automatic Speech Recognition and Understanding, Proceedings,* 2015, pp. 368–374.

[5] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling", in *INTERSPEECH 2016 - 17th Annual Conference of the International Speech Communication Association, Proceedings*, 2016, pp. 685–689.

[6] P. Zhou, W. Shi, J. Tian, et al., "Attention-based bidirectional long short-term memory networks for relation classification," in *ACL 2016 - 54th Annual Meeting of the Association for Computational Linguistics, Proceedings*, 2016, pp. 207–212.

[7] J. Ive, G. Gkotsis, R. Dutta, R. Stewart, and S. Velupillai, "Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health", in *CLPsych 2018 - 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, Proceedings*, 2018, pp. 69-77.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2012.

[10] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *WWW 2013 - 22nd International Conference on World Wide Web, Proceedings,* 2013, pp. 1445–1456.

[11] V. K. R. Sridhar, "Unsupervised topic modeling for short texts using distributed representations of words," in *VSM–NLP 2015 - 1st workshop on Vector Space Modeling for Natural Language Processing, Proceedings, 2015*, pp. 192–200.

[12] Y. Luan, S. Watanabe, and B. Harsham, "Efficient learning for spoken language understanding tasks with word embedding based pre-training," in *INTERSPEECH 2015 - 16th Annual Conference of the International Speech Communication Association, Proceedings,* 2015.pp.1398-1402.

[13] C. Li, H. Wang, Z. Zhang, et al., "Topic Modeling for Short Texts with Auxiliary Word Embeddings," in *SIGIR 2016 - 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings,* 2016, pp. 165-174.

[14] A. Budhkar and F. Rudzicz, "Augmenting word2vec with Latent Dirichlet Allocation within a Clinical Application," *in NAACL-HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings,* 2019, pp. 4095–4099.

[15] S. Shankar, S. Garg, and S. Sarawagi, "Surprisingly Easy Hard-Attention for Sequence to Sequence Learning," *in EMNLP 2018 - 2018 Conference on Empirical Methods in Natural Language Processing, Proceedings,* 2018, pp. 640–645.

[16] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[17] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *ASRU 2013 - 4th IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, Proceedings,* 2013, pp. 273–278.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[19] W. Zhang, Z. Chen, W. Che, G. Hu, and T. Liu, "The First Evaluation of Chinese Human-Computer Dialogue Technology," *arXiv:1709.10217 [cs]*, Sep. 2017.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *in ICLR 2013 - 1st International Conference on Learning Representations, Proceedings,* 2013, pp. 1–12.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *in ICLR 2015 - 3rd International Conference on Learning Representations, Proceedings,* 2015, pp. 1–13.