Conditional BERT Contextual Augmentation

Xing Wu[1,2], Shangwen Lv[1,2], Liangjun Zang[1(✉)], Jizhong Han[1,2],
and Songlin Hu[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{wuxing,lvshangwen,zangliangjun,hanjizhong,husonglin}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences,
Beijing, China

Abstract. Data augmentation methods are often applied to prevent overfitting and improve generalization of deep neural network models. Recently proposed contextual augmentation augments labeled sentences by randomly replacing words with more varied substitutions predicted by language model. Bidirectional Encoder Representations from Transformers (BERT) demonstrates that a deep bidirectional language model is more powerful than either an unidirectional language model or the shallow concatenation of a forward and backward model. We propose a novel data augmentation method for labeled sentences called conditional BERT contextual augmentation. We retrofit BERT to conditional BERT by introducing a new conditional masked language model (The term "conditional masked language model" appeared once in original BERT paper, which indicates context-conditional, is equivalent to term "masked language model". In our paper, "conditional masked language model" indicates we apply extra label-conditional constraint to the "masked language model".) task. The well trained conditional BERT can be applied to enhance contextual augmentation. Experiments on six various different text classification tasks show that our method can be easily applied to both convolutional or recurrent neural networks classifier to obtain improvement.

1 Introduction

Deep neural network-based models are easy to overfit and result in losing their generalization due to limited size of training data. In order to address the issue, data augmentation methods are often applied to generate more training samples. Recent years have witnessed great success in applying data augmentation in the field of speech area [10,14] and computer vision [17,24,27]. Data augmentation in these areas can be easily performed by transformations like resizing, mirroring,

Supported by the National Key Research and Development Program of China (No. 2017YFB1010000) and the National Natural Science Foundation of China (No. 61702500).

J. M. F. Rodrigues et al. (Eds.): ICCS 2019, LNCS 11539, pp. 84–95, 2019.
https://doi.org/10.1007/978-3-030-22747-0_7

random cropping, and color shifting. However, applying these universal transformations to texts is largely randomized and uncontrollable, which makes it impossible to ensure the semantic invariance and label correctness. For example, given a movie review "The actors is good", by mirroring we get "doog si srotca ehT", or by random cropping we get "actors is", both of which are meaningless.

Existing data augmentation methods for text are often loss of generality, which are developed with handcrafted rules or pipelines for specific domains. A general approach for text data augmentation is replacement-based method, which generates new sentences by replacing the words in the sentences with relevant words (e.g. synonyms). However, words with synonyms from a handcrafted lexical database likes WordNet [19] are very limited, and the replacement-based augmentation with synonyms can only produce limited diverse patterns from the original texts. To address the limitation of replacement-based methods, Kobayashi [15] proposed contextual augmentation for labeled sentences by offering a wide range of substitute words, which are predicted by a label-conditional bidirectional language model according to the context. But contextual augmentation suffers from two shortages: the bidirectional language model is simply shallow concatenation of a forward and backward model, and the usage of LSTM models restricts their prediction ability to a short range.

BERT, which stands for Bidirectional Encoder Representations from Transformers, pre-trained deep bidirectional representations by jointly conditioning on both left and right context in all layers. BERT addressed the unidirectional constraint by proposing a "masked language model" (MLM) objective by masking some percentage of the input tokens at random, and predicting the masked words based on its context. This is very similar to how contextual augmentation predict the replacement words. But BERT was proposed to pre-train text representations, so MLM task is performed in an unsupervised way, taking no label variance into consideration.

This paper focuses on the replacement-based methods, by proposing a novel data augmentation method called conditional BERT contextual augmentation. The method applies contextual augmentation by conditional BERT, which is fine-tuned on BERT. We adopt BERT as our pre-trained language model with two reasons. First, BERT is based on Transformer. Transformer provides us with a more structured memory for handling long-term dependencies in text. Second, BERT, as a deep bidirectional model, is strictly more powerful than the shallow concatenation of a left-to-right and right-to left model. So we apply BERT to contextual augmentation for labeled sentences, by offering a wider range of substitute words predicted by the masked language model task. However, the masked language model predicts the masked word based only on its context, so the predicted word maybe incompatible with the annotated labels of the original sentences. In order to address this issue, we introduce a new fine-tuning objective: the "conditional masked language model" (C-MLM). The conditional masked language model randomly masks some of the tokens from an input, and the objective is to predict a label-compatible word based on both its context and sentence label. Unlike Kobayashi's work [15], the C-MLM objective allows a deep bidirectional representations by jointly conditioning on both left and right context in all layers. In order to evaluate how well our augmentation method improves performance of deep neural network models, following Kobayashi, we experiment it on two most common neural network structures, LSTM-RNN and CNN, on text classification tasks. Through the experiments on six various different text classification tasks, we demonstrate that the proposed conditional BERT model augments sentence better than baselines, and conditional BERT contextual augmentation method can be easily applied to both convolutional or recurrent neural networks classifier. We further explore our conditional MLM task's connection with style transfer task and demonstrate that our conditional BERT can also be applied to style transfer too.

Our contributions are concluded as follows:

- We propose a conditional BERT contextual augmentation method. The method allows BERT to augment sentences without breaking the label-compatibility. Our conditional BERT can further be applied to style transfer task.
- Experimental results show that our approach obviously outperforms existing text data augmentation approaches.

To our best knowledge, this is the first attempt to alter BERT to a conditional BERT or apply BERT on text generation tasks.

2 Related Work

2.1 Fine-Tuning on Pre-trained Language Model

Language model pre-training has attracted wide attention and fine-tuning on pre-trained language model has shown to be effective for improving many downstream natural language processing tasks. Dai [2] pre-trained unlabeled data to improve Sequence Learning with recurrent networks. Howard [8] proposed a general transfer learning method, Universal Language Model Fine-tuning (ULMFiT), with the key techniques for fine-tuning a language model. Radford [23] proposed that by generative pre-training of a language model on a diverse corpus of unlabeled text, large gains on a diverse range of tasks could be realized. Radford [23] achieved large improvements on many sentence-level tasks from the GLUE benchmark [29]. BERT [4] obtained new state-of-the-art results on a broad range of diverse tasks. BERT pre-trained deep bidirectional representations which jointly conditioned on both left and right context in all layers, following by discriminative fine-tuning on each specific task. Unlike previous works fine-tuning pre-trained language model to perform discriminative tasks, we aim to apply pre-trained BERT on generative tasks by perform the masked language model (MLM) task. To generate sentences that are compatible with given labels, we retrofit BERT to conditional BERT, by introducing a conditional masked language model task and fine-tuning BERT on the task.

2.2 Text Data Augmentation

Text data augmentation has been extensively studied in natural language processing. Sample-based methods includes downsampling from the majority classes and oversampling from the minority class, both of which perform weakly in practice. Generation-based methods employ deep generative models such as GANs [7] or VAEs [1,9], trying to generate sentences from a continuous space with desired attributes of sentiment and tense. However, sentences generated in these methods are very hard to guarantee the quality both in label compatibility and sentence readability. In some specific areas [5,11,32]. word replacement augmentation was applied. Wang [30] proposed the use of neighboring words in continuous representations to create new instances for every word in a tweet to augment the training dataset. Zhang [34] extracted all replaceable words from the given text and randomly choose r of them to be replaced, then substituted the replaceable words with synonyms from WordNet [19]. Kolomiyets [16] replaced only the headwords under a task-specific assumption that temporal trigger words usually occur as headwords. Kolomiyets [16] selected substitute words with top-K scores given by the Latent Words LM [3], which is a LM based on fixed length contexts. Fadaee [6] focused on the rare word problem in machine translation, replacing words in a source sentence with only rare words. A word in the translated sentence is also replaced using a word alignment method and a rightward LM. The work most similar to our research is Kobayashi [15]. Kobayashi used a fill-in-the-blank context for data augmentation by replacing every words in the sentence with language model. In order to prevent the generated words from reversing the information related to the labels of the sentences, Kobayashi [15] introduced a conditional constraint to control the replacement of words. Unlike previous works, we adopt a deep bidirectional language model to apply replacement, and the attention mechanism within our model allows a more structured memory for handling long-term dependencies in text, which resulting in more general and robust improvement on various downstream tasks.

3 Conditional BERT Contextual Augmentation

3.1 Preliminary: Masked Language Model Task

Bidirectional Language Model. In general, the language model (LM) models the probability of generating natural language sentences or documents. Given a sequence $\boldsymbol{S}$ of N tokens, $<t_1, t_2, ..., t_N>$, a forward language model allows us to predict the probability of the sequence as:

$$p(t_1, t_2, ..., t_N) = \prod_{i=1}^{N} p(t_i | t_1, t_2, ..., t_{i-1}). \tag{1}$$

Similarly, a backward language model allows us to predict the probability of the sentence as:

$$p(t_1, t_2, ..., t_N) = \prod_{i=1}^{N} p(t_i | t_{i+1}, t_{i+2}, ..., t_N). \tag{2}$$

Traditionally, a bidirectional language model a shallow concatenation of independently trained forward and backward LMs.

Masked Language Model Task. In order to train a deep bidirectional language model, BERT proposed Masked Language Model (MLM) task, which was also referred to Cloze Task [28]. MLM task randomly masks some percentage of the input tokens, and then predicts only those masked tokens according to their context. Given a masked token t_i, the context is the tokens surrounding token t_i in the sequence $\boldsymbol{S}$, i.e., cloze sentence $\boldsymbol{S}\backslash\{t_i\}$. The final hidden vectors corresponding to the mask tokens are fed into an output softmax over the vocabulary to produce words with a probability distribution $p(\cdot|\boldsymbol{S}\backslash\{t_i\})$. MLM task only predicts the masked words rather than reconstructing the entire input, which suggests that more pre-training steps are required for the model to converge. Pre-trained BERT can augment sentences through MLM task, by predicting new words in masked positions according to their context.

3.2 Conditional BERT

As shown in Fig. 1, our conditional BERT shares the same model architecture with the original BERT. The differences are the input representation and training procedure.

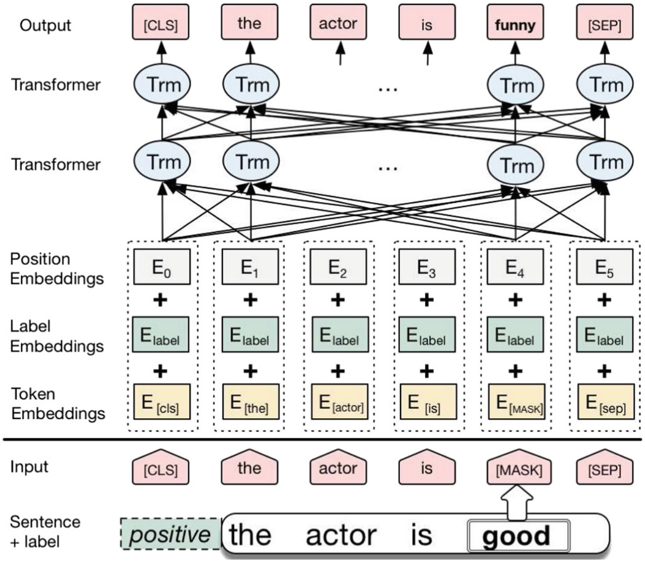

Fig. 1. Model architecture of conditional BERT. The label embeddings in conditional BERT corresponding to segmentation embeddings in BERT, but their functions are different.

The input embeddings of BERT are the sum of the token embeddings, the segmentation embeddings and the position embeddings. For the segmentation embeddings in BERT, a learned sentence A embedding is added to every token of the first sentence, and if a second sentence exists, a sentence B embedding will be added to every token of the second sentence. However, the segmentation embeddings has no connection to the actual annotated labels of a sentence, like sense, sentiment or subjectivity, so predicted word is not always compatible with annotated labels. For example, given a positive movie remark "this actor is good", we have the word"good" masked. Through the Masked Language Model task by BERT, the predicted word in the masked position has potential to be negative word likes "bad" or "boring". Such new generated sentences by substituting masked words are implausible with respect to their original labels, which will be harmful if added to the corpus to apply augmentation. In order to address this issue, we propose a new task: "conditional masked language model".

Conditional Masked Language Model. The conditional masked language model randomly masks some of the tokens from the labeled sentence, and the objective is to predict the original vocabulary index of the masked word based on both its context and its label. Given a masked token t_i, the context $\boldsymbol{S}\backslash\{t_i\}$ and label y are both considered, aiming to calculate $p(\cdot|y, \boldsymbol{S}\backslash\{t_i\})$, instead of calculating $p(\cdot|\boldsymbol{S}\backslash\{t_i\})$. Unlike MLM pre-training, the conditional MLM objective allows the representation to fuse the context information and the label information, which allows us to further train a label-conditional deep bidirectional representations.

To perform conditional MLM task, we fine-tune on pre-trained BERT. We alter the segmentation embeddings to label embeddings, which are learned corresponding to their annotated labels on labeled datasets. Note that the BERT are designed with segmentation embedding being embedding A or embedding B, so when a downstream task dataset with more than two labels, we have to adapt the size of embedding to label size compatible. We train conditional BERT using conditional MLM task on labeled dataset. After the model has converged, it is expected to be able to predict words in masked position both considering the context and the label.

3.3 Conditional BERT Contextual Augmentation

After the conditional BERT is well-trained, we utilize it to augment sentences. Given a labeled sentence from the corpus, we randomly mask a few words in the sentence. Through conditional BERT, various words compatibly with the label of the sentence are predicted by conditional BERT. After substituting the masked words with predicted words, a new sentences is generated, which shares similar context and same label with original sentence. Then new sentences are added to original corpus. We elaborate the entire process in Algorithm 1.

Algorithm 1. Conditional BERT contextual augmentation algorithm. Fine-tuning on the pre-trained BERT, we retrofit BERT to conditional BERT using conditional MLM task on labeled dataset. After the model converged, we utilize it to augment sentences. New sentences are added into dataset to augment the dataset.

```
Alter the segmentation embeddings to label embeddings
Fine-tune the pre-trained BERT using conditional MLM task on labeled dataset D until convergence
for each iteration i=1,2,...,M do
    Sample a sentence $s$ from D
    Randomly mask $k$ words
    Using fine-tuned conditional BERT to predict label-compatible words on masked positions to generate a new sentence $S'$
end for
Add new sentences into dataset $D$ to get augmented dataset $D'$
Perform downstream task on augmented dataset $D'$
```

4 Experiment

In this section, we present conditional BERT parameter settings and, following Kobayashi [15], we apply different augmentation methods on two types of neural models through six text classification tasks. The pre-trained BERT model we used in our experiment is BERT$_{BASE}$, with number of layers (i.e., Transformer blocks) $L = 12$, the hidden size $H = 768$, and the number of self-attention heads $A = 12$, total parameters $= 110M$. Detailed pre-train parameters setting can be found in original paper [4]. For each task, we perform the following steps independently. First, we evaluate the augmentation ability of original BERT model pre-trained on MLM task. We use pre-trained BERT to augment dataset, by predicted masked words only condition on context for each sentence. Second, we fine-tune the original BERT model to a conditional BERT. Well-trained conditional BERT augments each sentence in dataset by predicted masked words condition on both context and label. Third, we compare the performance of the two methods with Kobayashi's [15] contextual augmentation results.

4.1 Datasets

Six benchmark classification datasets are listed in Table 1. Following Kim [12], for a dataset without validation data, we use 10% of its training set for the validation set. Summary statistics of six classification datasets are shown in Table 1.

SST [25] SST (Stanford Sentiment Treebank) is a dataset for sentiment classification on movie reviews, which are annotated with five labels (SST5: very positive, positive, neutral, negative, or very negative) or two labels (SST2: positive or negative).

Table 1. Summary statistics for the datasets after tokenization. c: Number of target classes. l: Average sentence length. N: Dataset size. $|V|$: Vocabulary size. $Test$: Test set size (CV means there was no standard train/test split and thus 10-fold cross-validation was used).

Data	c	l	N	$\|V\|$	$Test$
SST5	5	18	11855	17836	2210
SST2	2	19	9613	16185	1821
Subj	2	23	10000	21323	CV
TREC	6	10	5952	9592	500
MPQA	2	3	10606	6246	CV
RT	2	21	10662	20287	CV

Subj [20] Subj (Subjectivity dataset) is annotated with whether a sentence is subjective or objective.
MPQA [31] MPQA Opinion Corpus is an opinion polarity detection dataset of short phrases rather than sentences, which contains news articles from a wide variety of news sources manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.).
RT [21] RT is another movie review sentiment dataset contains a collection of short review excerpts from Rotten Tomatoes collected by Bo Pang and Lillian Lee.
TREC [18] TREC is a dataset for classification of the six question types (whether the question is about person, location, numeric information, etc.).

4.2 Text Classification

Sentence Classifier Structure. We evaluate the performance improvement brought by conditional BERT contextual augmentation on sentence classification tasks, so we need to prepare two common sentence classifiers beforehand. For comparison, following Kobayashi [15], we adopt two typical classifier architectures: CNN or LSTM-RNN. The CNN-based classifier [12] has convolutional filters of size 3, 4, 5 and word embeddings. All outputs of each filter are concatenated before applied with a max-pooling over time, then fed into a two-layer feed-forward network with ReLU, followed by the softmax function. An RNN-based classifier has a single layer LSTM and word embeddings, whose output is fed into an output affine layer with the softmax function. For both the architectures, dropout [26] and Adam optimization [13] are applied during training. The train process is finish by early stopping with validation at each epoch.

Hyper-parameters Setting. Sentence classifier hyper-parameters including learning rate, embedding dimension, unit or filter size, and dropout ratio, are selected using grid-search for each task-specific dataset. We refer to Kobayashi's implementation in original paper. For BERT, all hyper-parameters are kept the

Table 2. Accuracies of different methods for various benchmarks on two classifier architectures. C-BERT, which represents conditional BERT, performs best on two classifier structures over six datasets. "w/" represents "with", lines marked with "*" are experiments results from Kobayashi [15].

Model	SST5	SST2	Subj	MPQA	RT	TREC	Avg.
CNN*	41.3	79.5	92.4	86.1	75.9	90.0	77.53
w/synonym*	40.7	80.0	92.4	86.3	76.0	89.6	77.50
w/context*	41.9	80.9	92.7	86.7	75.9	90.0	78.02
w/context+label*	42.1	80.8	93.0	86.7	76.1	90.5	78.20
w/BERT	41.5	81.9	92.9	87.7	78.2	91.8	79.00
w/C-BERT	**42.3**	**82.1**	**93.4**	**88.2**	**79.0**	**92.6**	**79.60**
RNN*	40.2	80.3	92.4	86.0	76.7	89.0	77.43
w/synonym*	40.5	80.2	92.8	86.4	76.6	87.9	77.40
w/context*	40.9	79.3	92.8	86.4	77.0	89.3	77.62
w/context+label*	41.1	80.1	92.8	86.4	77.4	89.2	77.83
w/BERT	41.3	81.4	93.5	87.3	78.3	89.8	78.60
w/C-BERT	**42.6**	**81.9**	**93.9**	**88.0**	**78.9**	**91.0**	**79.38**

same as Devlin [4]. The number of conditional BERT training epochs ranges in [1–50] and number of masked words ranges in [1–2].

Baselines. We compare the performance improvements obtained by our proposed method with the following baseline methods, "w/" means "with":

- w/synonym: Words are randomly replaced with synonyms from WordNet [19].
- w/context: Proposed by Kobayashi [15], which used a bidirectional language model to apply contextual augmentation, each word was replaced with a probability.
- w/context+label: Kobayashi's contextual augmentation method [15] in a label-conditional LM architecture.

Experiment Results. Table 2 lists the accuracies of the all methods on two classifier architectures. The results show that, for various datasets on different classifier architectures, our conditional BERT contextual augmentation improves the model performances most. BERT can also augments sentences to some extent, but not as much as conditional BERT does. For we masked words randomly, the masked words may be label-sensitive or label-insensitive. If label-insensitive words are masked, words predicted through BERT may not be compatible with original labels. The improvement over all benchmark datasets also shows that conditional BERT is a general augmentation method for multi-labels sentence classification tasks.

Effect of Number of Fine-Tuning Steps. We also explore the effect of number of training steps to the performance of conditional BERT data augmentation. The fine-tuning epoch setting ranges in [1–50], we list the fine-tuning epoch of conditional BERT to outperform BERT for various benchmarks in Table 3. The results show that our conditional BERT contextual augmentation can achieve obvious performance improvement after only a few fine-tuning epochs, which is very convenient to apply to downstream tasks.

Table 3. Fine-tuning epochs of conditional BERT to outperform BERT for various benchmarks

Model	SST5	SST2	Subj	MPQA	RT	TREC
CNN	4	3	1	2	2	1
RNN	6	2	2	2	1	1

5 Connection to Style Transfer

In this section, we further dip into the connection to style transfer and apply our well trained conditional BERT to style transfer task. Style transfer is defined as the task of rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context [22]. Our conditional MLM task changes words in the text condition on given label without changing the context. View from this point, the two tasks are very close. So in order to apply conditional BERT to style transfer task, given a specific stylistic sentence, we break it into two steps: first, we find the words relevant to the style; second, we mask the style-relevant words, then use conditional BERT to predict new substitutes with sentence context and target style property. In order to find style-relevant words in a sentence, we refer to Xu [33], which proposed an attention-based method to extract the contribution of each word to the sentence sentimental label. For example, given a positive movie remark "This movie is funny and interesting", we filter out the words that contribute largely to the label and mask them. Then

Table 4. Examples generated by conditional BERT on the SST2 dataset. To perform style transfer, we reverse the original label of a sentence, and conditional BERT output a new label compatible sentence.

Original:	there's no disguising this as one of the worst films of the summer
Generated:	there's no disguising this as one of the best films of the summer
Original:	it's probably not easy to make such a worthless film ...
Generated:	it's probably not easy to make such a stunning film ...
Original:	woody allen has really found his groove these days
Generated:	woody allen has really lost his groove these days

through our conditional BERT contextual augmentation method, we fill in the masked positions by predicting words conditioning on opposite label and sentence context, resulting in "This movie is boring and dull". The words "boring" and "dull" contribute to the new sentence being labeled as negative style. We sample some sentences from dataset SST2, transferring them to the opposite label, as listed in Table 4.

6 Conclusions and Future Work

In this paper, we fine-tune BERT to conditional BERT by introducing a novel conditional MLM task. After being well trained, the conditional BERT can be applied to data augmentation for sentence classification tasks. Experiment results show that our model outperforms several baseline methods obviously. Furthermore, we demonstrate that our conditional BERT can also be applied to style transfer task. In the future, (1) We will explore how to perform text data augmentation on imbalanced datasets with pre-trained language model, (2) we believe the idea of conditional BERT contextual augmentation is universal and will be applied to paragraph or document level data augmentation.

References

1. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349 (2015)
2. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning, pp. 3079–3087 (2015)
3. Deschacht, K., Moens, M.F.: Semi-supervised semantic role labeling using the latent words language model, pp. 21–29 (2009)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Ebrahimi, J., Rao, A., Lowd, D., Dou, D.: Hotflip: White-box adversarial examples for NLP. arXiv preprint arXiv:1712.06751 (2017)
6. Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. arXiv preprint arXiv:1705.00440 (2017)
7. Goodfellow, I., et al.: Generative adversarial nets, pp. 2672–2680 (2014)
8. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification, vol. 1, pp. 328–339 (2018)
9. Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., Xing, E.P.: Toward controlled generation of text. arXiv preprint arXiv:1703.00955 (2017)
10. Jaitly, N., Hinton, G.E.: Vocal tract length perturbation (VTLP) improves speech recognition, vol. 117 (2013)
11. Jia, R., Liang, P.: Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328 (2017)
12. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

14. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio augmentation for speech recognition (2015)
15. Kobayashi, S.: Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201 (2018)
16. Kolomiyets, O., Bethard, S., Moens, M.F.: Model-portability experiments for textual temporal analysis, pp. 271–276 (2011)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks, pp. 1097–1105 (2012)
18. Li, X., Roth, D.: Learning question classifiers, pp. 1–7 (2002)
19. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
20. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, p. 271 (2004)
21. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales, pp. 115–124 (2005)
22. Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., Black, A.W.: Style transfer through back-translation. arXiv preprint arXiv:1804.09000 (2018)
23. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018). https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
24. Simard, P.Y., LeCun, Y.A., Denker, J.S., Victorri, B.: Transformation invariance in pattern recognition—tangent distance and tangent propagation. In: Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 1524, pp. 239–274. Springer, Heidelberg (1998). https://doi.org/10.1007/3-540-49430-8_13
25. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment treebank, pp. 1631–1642 (2013)
26. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
27. Szegedy, C., et al.: Going deeper with convolutions, pp. 1–9 (2015)
28. Taylor, W.L.: "cloze procedure": a new tool for measuring readability. Journ. Bull. **30**(4), 415–433 (1953)
29. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
30. Wang, W.Y., Yang, D.: That's so annoying!!!: a lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets, pp. 2557–2563 (2015)
31. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Lang. Resour. Eval. **39**(2–3), 165–210 (2005)
32. Xie, Z., et al.: Data noising as smoothing in neural network language models. arXiv preprint arXiv:1703.02573 (2017)
33. Xu, J., et al.: Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. arXiv preprint arXiv:1805.05181 (2018)
34. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification, pp. 649–657 (2015)