








# Text Augmentation in a Multi-Task View

Jason Wei  Chengyu Huang  Shiqi Xu  Soroush Vosoughi   
 ProtagoLabs, {jason, xu}@protagolabs.com  
 International Monetary Fund, huangchengyu24@gmail.com  
 Dartmouth College, soroush@dartmouth.edu

## Abstract

Traditional data augmentation aims to increase the coverage of the input distribution by generating augmented examples that strongly resemble original samples in an online fashion where augmented examples dominate training.

In this paper, we propose an alternative perspective—a *multi-task view* (MTV) of data augmentation—in which the primary task trains on original examples and the auxiliary task trains on augmented examples. In MTV data augmentation, both original and augmented samples are weighted substantively during training, relaxing the constraint that augmented examples must resemble original data and thereby allowing us to apply stronger levels of augmentation.

In empirical experiments using four common data augmentation techniques on three benchmark text classification datasets, we find that the MTV leads to higher and more robust performance improvements than traditional augmentation.

## 1 Introduction

Most data augmentation techniques aim to generate augmented examples for training that are similar to original data. In computer vision, operations such as flipping, cropping, and color jittering are both widely used and highly effective—it is self-evident that augmented examples closely resemble original data, and so we generate augmented data in an online fashion during each minibatch such that no original, unmodified examples are seen during training (Krizhevsky et al., 2012; Zagoruyko and Komodakis, 2016; Huang et al., 2017).

In language, on the other hand, even slight modifications can cause significant semantic changes, and so it is not always clear whether augmented examples resemble original data. Despite this uncertainty, many augmentation techniques in NLP still

### Traditional Data Augmentation

- **Intuition:** Increase coverage of input distribution by using augmented examples for training.
- **Guideline:** Augmented examples should be similar to original data.
- **Training:** Dominated by augmented examples that are generated stochastically.

### Multi-Task View (MTV) of Data Augmentation

- **Intuition:** Auxiliary task of classifying augmented examples acts as regularization for the primary task of classifying original examples.
- **Guideline:** It might be a good idea for augmented samples to resemble original data, but they can be anything that boosts performance.
- **Training:** Both original and augmented data receive substantive weighting during training.

Table 1: Summary of traditional data augmentation versus MTV data augmentation.

generate examples stochastically and ignore original data (Zhang et al., 2015; Sennrich et al., 2016a; Xie et al., 2017; Li et al., 2017; Kobayashi, 2018; Wang et al., 2018). When it is unclear whether augmented examples resemble original data—as is often the case—is it wise to neglect the original training data?

Our paper questions this practice by proposing to include original data during training. Specifically, we make two contributions:

1. We propose a *multi-task view of data augmentation* (MTV data augmentation), which trains on both original and augmented examples and therefore allows us to relax the constraint that augmented examples must resemble original data. The MTV facilitates augmentation using a higher strength parameter.
2. We show empirically that four common data augmentation techniques provide higher and more robust performance gains using the MTV compared with traditional augmentation.

## 2 Traditional Data Augmentation<sup>1</sup>

**Situation.** During regular training, the canonical maximum likelihood objective minimizes the cost of the original training set  $J_O$ :

$$J_O(\theta) = \mathbb{E}_{x,y \sim \hat{\mathbf{p}}(X,Y)} [-\log \mathbf{p}_\theta(y|x)],$$

where  $\hat{\mathbf{p}}(X, Y)$  is the empirical distribution of training pairs  $x, y$  and  $\mathbf{p}_\theta(y|x)$  is the parameterized model that we aim to learn (e.g., a neural network). As  $\hat{\mathbf{p}}(X, Y)$  is typically the observed data, it will likely have some mismatch with the true data distribution  $\mathbf{p}(X, Y)$ . When the mismatch is dramatic—for instance, when  $\hat{\mathbf{p}}(X, Y)$  does not sufficiently cover the training space—model performance will likely suffer.

**Remedy.** In practice, we often use data augmentation to mitigate the inadequacy of  $\hat{\mathbf{p}}(X, Y)$  by providing additional training data. We generate an augmented distribution  $\mathbf{q}(\hat{X}, \hat{Y})$  and now minimize the cost of this augmented training set  $J_{\text{aug}}$ :

$$J_{\text{aug}}(\theta) = \mathbb{E}_{x,y \sim \mathbf{q}(\hat{X}, \hat{Y})} [-\log \mathbf{p}_\theta(y|x)].$$

As we now optimize solely on  $\mathbf{q}(\hat{X}, \hat{Y})$ , our goal is to find  $(\hat{x}, \hat{y})$  pairs that are likely to fall in the true distribution  $\mathbf{p}$ . Assuming the smoothness of  $\mathbf{p}$ , similar  $(x, y)$  pairs will have similar probabilities, and therefore if an augmented example is more similar to an observed example, it is more likely to be sampled under the true distribution. In other words, good augmented examples resemble the observed data, and we aim to find them. Conversely, if an augmented example diverges too far from any observed data, it is likely invalid and thus harmful for training; we don't want to train on these examples.

The majority of prior work follows this framework of augmented examples resembling real data. As popular techniques, semantic noising substitutes tokens with synonyms (Wang and Yang, 2015; Zhang et al., 2015; Li et al., 2017); Pervasive Dropout randomly removes words from the input sequence (Sennrich et al., 2016a); and SwitchOut (for machine translation) replaces some words in both source and target sentences with other words from their corresponding vocabularies (Wang et al., 2018).

Moreover, most of these techniques perform augmentation on every training example in an online fashion, implicitly assuming that augmented examples so closely resemble original data that directly

<sup>1</sup>We closely follow the intuition and notation of Wang et al. (2018)

training on original examples is not even worth considering. As we shall see in the next section, adding in these original examples during training might actually be a worthwhile idea.

## 3 MTV Data Augmentation

Multi-task optimization jointly trains on a primary task and one or more auxiliary tasks—the intuition is that requiring an algorithm to also learn an auxiliary task can act as better regularization than penalizing all complexity uniformly. Prior work has found that multi-task models work particularly well when the tasks are similar, but can also improve performance even on unrelated tasks (Paredes et al., 2012; Hajiramezanali et al., 2018).

We propose a multi-task view of data augmentation that has a primary task that optimizes regular training on original examples and an auxiliary task that optimizes training on augmented data. This MTV jointly optimizes the primary and auxiliary task(s) using a weighted cost function so that both original and augmented data receive substantial weight during training:

$$J(\theta) = \gamma_O \cdot J_O(\theta) + \gamma_{\text{aug}} \cdot J_{\text{aug}}(\theta),$$

where  $\gamma_O$  is the weight of original data and  $\gamma_{\text{aug}}$  is the weight of augmented data, and  $\gamma_O + \gamma_{\text{aug}} = 1$ . In this context, observe that vanilla training uses  $\gamma_O = 1$  and  $\gamma_{\text{aug}} = 0$ , and traditional data augmentation uses  $\gamma_O = 0$  and  $\gamma_{\text{aug}} = 1$ .

The MTV gives us an important freedom that is not offered by the traditional data augmentation framework. Since traditional data augmentation only trains on augmented examples, performance suffers detrimentally when augmented data differs too much from the true distribution—therefore, most studies aim to generate augmented examples that resemble original data. MTV data augmentation, however, jointly trains on both original and augmented data, thereby allowing us to relax the constraint that original and augmented examples come from the same distribution. In fact, accepting that the original and augmented distributions might differ or could even be unrelated—as work in multi-task learning has done (Paredes et al., 2012; Hajiramezanali et al., 2018; Rai and Daumé, 2010)—liberates us to apply stronger levels of data augmentation, which, as we will demonstrate in the next section, leads to higher and more robust performance.

## 4 Experiments

This section compares multi-task view augmentation to traditional augmentation for various datasets and augmentation techniques.

### 4.1 Experimental Setup

**Datasets.** We conduct experiments on three text classification tasks often used as benchmarks (Kim, 2014): (1) Stanford Sentiment Treebank (SST2) (Socher et al., 2013) of movie reviews classified as positive/negative, (2) subjectivity/objectivity dataset (SUBJ) (Pang and Lee, 2004), where sentences are classified as either subjective or objective, and (3) question type dataset (TREC) (Li and Roth, 2002), in which questions ask for either a description, entity, abbreviation, human, location, or number.

**Models and Experimental Procedures.** For text classification, we use BERT (Devlin et al., 2019) (bert-base-uncased from HuggingFace) to extract features by averaging the last hidden states of the input tokens. To reduce the number of model hyperparameters and save computation time, we classify these features using a linear SVM trained for 1000 epochs.<sup>2</sup> Since training data size depends on the amount of augmented data, we adjust the number of training epochs so that all models receive the same number of updates. All experiments are run for five random seeds. Our baseline models without data augmentation achieved 84.5%, 93.1%, and 83.9% accuracy respectively on the SST2, SUBJ, and TREC tasks.

**Augmentation Techniques.** In this paper, we experiment with four simple and common data augmentation techniques studied in Wei and Zou (2019): (1) **Token Substitution** (Zhang et al., 2015) replaces words with WordNet (Miller, 1995) synonyms; (2) **Pervasive Dropout** (Sennrich et al., 2016a) applies word-level dropout; (3) **Token Injection** (Wei and Zou, 2019) insert a synonym of a random token in the sequence into a random position in that sequence; (4) **Positional Shuffling** (Wei and Zou, 2019) randomly chooses two tokens and swaps their positions. For all four techniques, a parameter  $\alpha$  indicates *augmentation strength* by dictating how many perturbations are performed. For a given  $\alpha$ , we perform  $n = \alpha l$  perturbations, where  $l$  is the sequence length.

<sup>2</sup>This setup is not state-of-the-art but allows for experiments to be performed on CPU.

Aug. Technique	MTV	Best $\alpha$	Avg. Boost	( $\Delta_{\text{MTV}}$ )
Token Substitution	✗	0.05	1.3%	-
	✓	0.3	2.1%	(+0.8%)
Pervasive Dropout	✗	0.1	1.8%	-
	✓	0.4	2.5%	(+0.7%)
Token Injection	✗	0.05	0.7%	-
	✓	0.5	2.2%	(+1.5%)
Positional Shuffling	✗	0.05	1.4%	-
	✓	0.4	2.5%	(+1.1%)

Table 2: Average performance boost on three text classification tasks for four augmentation techniques using the best-performing augmentation strength from  $\alpha \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ . Traditional data augmentation works best at low  $\alpha$ , whereas MTV data augmentation provides the strongest performance for high  $\alpha$ .  $\Delta_{\text{MTV}}$  indicates additional boost from using the MTV compared with traditional augmentation.

### 4.2 Stronger augmentation for more gains

Table 2 summarizes results for data augmentation in the MTV using  $\gamma_{\text{O}} = \gamma_{\text{aug}} = 0.5$  compared with traditional augmentation for the best-performing augmentation strength from  $\alpha \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ . In the traditional framework, pervasive dropout had the strongest performance boost of 1.8% using  $\alpha = 0.1$ . The MTV, however, allowed for stronger augmentation (i.e.,  $\alpha \geq 0.3$ ) that resulted in all four techniques to achieving boosts of more than 2.0%.

Perhaps strikingly, token injection and positional shuffling, which are less intuitive and not as commonly used as token substitution and pervasive dropout, achieve the strongest gains ( $> 1.0\%$ ) from using the MTV. One potential reason for this is that, compared with token substitution and pervasive dropout, token injection and positional shuffling are non-destructive in that they do not remove any of the original words, and so the nature of examples augmented at high  $\alpha$  could be more conducive for the MTV.

### 4.3 More-robust gains at high $\alpha$

When using data augmentation with high  $\alpha$ , high levels of noising are employed and augmented data are therefore more likely to diverge from their original examples. Figure 1 takes a closer look at how performance is affected by varying  $\alpha$ . Whereas traditional augmentation often negatively affected performance at high  $\alpha$ , the multi-task view, which jointly optimizes the original distribution, had robust performance gains at high augmentation strengths.

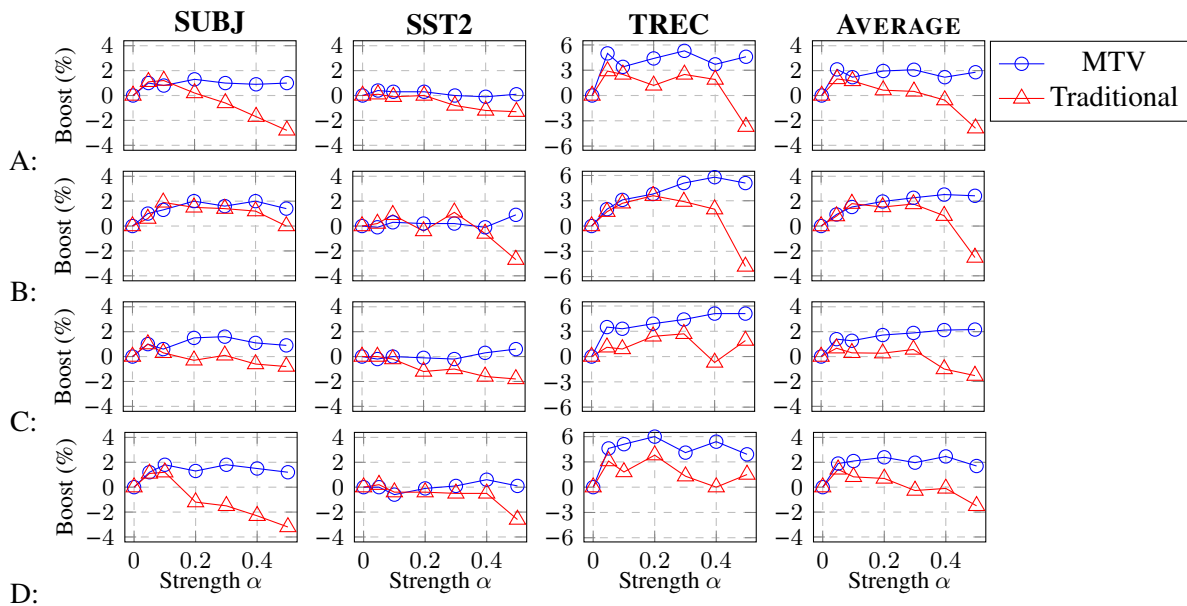


Figure 1: Performance boosts on three tasks using the traditional and multi-task view (MTV) frameworks for four data augmentation techniques: Token Substitution (Zhang et al., 2015) (A), Pervasive Dropout (Sennrich et al., 2016a) (B), Token Injection (Wei and Zou, 2019) (C), Positional Shuffling (Wei and Zou, 2019) (D). In the traditional framework, improvements are largest when augmentation strength  $\alpha$  is small, with performance deteriorating for large  $\alpha$ . The MTV, on the other hand, jointly optimizes for both original and augmented data, leveraging higher  $\alpha$  to provide higher and more robust performance gains.

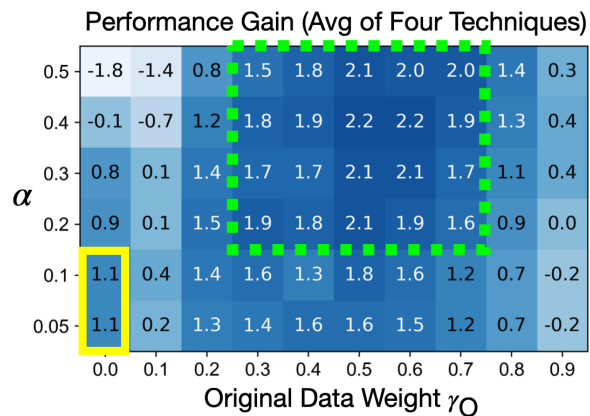


Figure 2: Performance boost (%) with varying augmentation strengths  $\alpha$  and weights of original data  $\gamma_0$  during training. Traditional data augmentation (yellow solid box in lower left) uses modest augmentation strength ( $\alpha=0.05, 1$ ) with no original examples for training ( $\gamma_0=0$ ). The MTV data augmentation approach (green dashed box) suggests substantive weighting of original examples (e.g.,  $\gamma_0=0.5$ ) which allows for much stronger augmentation (e.g.,  $\alpha \geq 0.3$ ).

#### 4.4 Choosing $\gamma_0$ and $\gamma_{\text{aug}}$ weighting

As our experiments so far have used the MTV with balanced weighting of original and augmented data ( $\gamma_0 = \gamma_{\text{aug}} = 0.5$ ), in this section we explore different weightings of  $\gamma_0$  and  $\gamma_{\text{aug}}$ . Figure 2 shows these results averaged over all three datasets and

all four augmentation techniques. Traditional data augmentation, which uses modest augmentation strength (e.g.,  $\alpha \in \{0.05, 0.1\}$ ) and does not train on original data ( $\gamma_0=0.0$ ), achieves reasonable performance gains. As expected, when stronger augmentations were applied (e.g.,  $\alpha \geq 0.4$ ), training with only augmented data hurts performance. When training on both augmented and original data, however, performance improved with stronger augmentation and remained robust for varying augmentation strengths  $0.2 \leq \alpha \leq 0.5$  and original data weights  $0.3 \leq \gamma_0 \leq 0.7$ .

## 5 Further Related Work

Prior work on data augmentation, to our knowledge, generally follows the traditional data augmentation framework. In addition to the methods mentioned in §2, Xie et al. (2017) replaced words with samples from the unigram frequency distribution; Yu et al. (2018) translated English sentences to French and back to English (backtranslation); and Kobayashi (2018) replaced words with other words based on a language model. All these methods could potentially be formulated in the MTV.

Some prior work has also drawn connections between seeing data augmentation as multiple tasks. Similar to how we optimize augmented data as a separate task, Meyerson and Miikkulainen (2018)

created fake tasks by using multiple distinct decoders to train a shared structure to solve the same problem in different ways. In machine translation, Sennrich et al. (2016b) used monolingual training examples as parallel examples with an empty source side, noting that their setup could be seen as multi-task learning with the tasks as translation with known sources and language modeling with unknown sources. Compared with these papers that create multiple tasks in very specialized scenarios, the multi-task view that we have presented here can be used for any type of text data augmentation.

To be clear, our study is not the first to mix original and augmented data in training. For instance, Wang and Yang (2015) use a ratio of 1:5 original to augmented examples, but this weight of original data is much smaller than the  $0.3 \leq \gamma_0 \leq 0.7$  that we advocate for. Sennrich et al. (2016b) also include original data when training with back-translation augmentation, but the given ratios of original and augmented data they use appear to be dictated by the speed of their back-translation models rather than an intentionally-motivated design choice. We see our work as the first to explicitly formulate the MTV, advocate for a joint optimization function, and comprehensively explore its implications on common text augmentation techniques.

As a limitation, our study has focused on label-preserving augmentation techniques, and our line of reasoning may not apply when augmentation techniques intentionally change the label. Moreover, we have only studied text classification with simple models using task-agnostic augmentation techniques. Future work in this direction could experiment with larger-scale models or study task-specific augmentation.

## 6 Conclusions

We have proposed a multi-task view that gives both original and augmented examples substantial weight during training, contrasting prior work that performs stochastic data augmentation and ignores original training data. For four common augmentation techniques, we found experimentally that this alternative view allows for stronger levels of augmentation, which in turn leads to better and more robust performance than traditional augmentation. We hope our paper inspires future work using text data augmentation to think more explicitly about how much augmented examples resemble original data and consider substantive weighting of origi-

nal data when using data augmentation to improve model performance.

To close, we leave the enthusiastic reader with one last thought. Most existing text data augmentation techniques have obediently followed the paradigm from computer vision of generating augmented examples that are similar to the original data. Who's to say that's how data augmentation ought to work in NLP? In this paper, we've shown how to search for relative freedom from this constraint, simply by taking a different view of the underlying assumptions. Now, a bigger question arises on the horizon—what new text augmentation techniques are unlocked when augmented data are not forced to resemble the original?

## References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ehsan Hajiramezani, Siamak Zamani Dadaneh, Alireza Karbalayghareh, Mingyuan Zhou, and Xiaoning Qian. 2018. [Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data](#). In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*, page 9133–9142, Red Hook, NY, USA. Curran Associates Inc.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. [Densely connected convolutional networks](#). In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- Yoon Kim. 2014. [Convolutional neural networks for sentence classification](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
- Sosuke Kobayashi. 2018. [Contextual augmentation: Data augmentation by words with paradigmatic relations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. [Imagenet classification with deep convolutional neural networks](#). In *Advances in neural information processing systems*, pages 1097–1105.
- Xin Li and Dan Roth. 2002. [Learning question classifiers](#). In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. [Robust training under linguistic adversity](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 21–27, Valencia, Spain. Association for Computational Linguistics.
- Elliot Meyerson and Risto Miikkulainen. 2018. [Pseudo-task augmentation: From deep multitask learning to intratask sharing-and back](#). In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3508–3517. PMLR.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Bo Pang and Lillian Lee. 2004. [A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts](#). In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bernardino Romera Paredes, Andreas Argyriou, Nadia Berthouze, and Massimiliano Pontil. 2012. [Exploiting unrelated tasks in multi-task learning](#). In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 951–959, La Palma, Canary Islands. PMLR.
- Piyush Rai and Hal Daumé. 2010. [Infinite predictor subspace models for multitask learning](#). In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 613–620, Chia Laguna Resort, Sardinia, Italy. PMLR.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. [Edinburgh neural machine translation systems for WMT 16](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. [Parsing with compositional vector grammars](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria. Association for Computational Linguistics.
- William Yang Wang and Diyi Yang. 2015. [That’s so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.
- Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. [SwitchOut: an efficient data augmentation algorithm for neural machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.
- Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.
- Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. 2017. [Data noising as smoothing in neural network language models](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. [Qanet: Combining local convolution with global self-attention for reading comprehension](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Sergey Zagoruyko and Nikos Komodakis. 2016. [Wide residual networks](#). In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification](#). In *Proceedings of the 28th International*

*Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.