Semi-supervised Text Regression with Conditional Generative Adversarial Networks

Tao Li
Purdue University
taoli@purdue.edu

Xudong Liu
ObEN, Inc.
xudong@oben.com

Shihan Su
California Institute of Technology
ssu@caltech.edu

Abstract

Enormous online textual information provides intriguing opportunities for understandings of social and economic semantics. In this paper, we propose a novel text regression model based on a conditional generative adversarial network (GAN), with an attempt to associate textual data and social outcomes in a semi-supervised manner. Besides promising potential of predicting capabilities, our superiorities are twofold: (i) the model works with unbalanced datasets of limited labelled data, which align with real-world scenarios; and (ii) predictions are obtained by an end-to-end framework, without explicitly selecting high-level representations. Finally we point out related datasets for experiments and future research directions.

1 Introduction

With millions of textual information uploaded every day, the Internet embeds tremendous data of social and economic phenomena, and have attracted consistent interests not only from sociologists and economists but also statisticians and computer scientists. For example, Joshi et al. (2010) forecasted movie revenues using online reviews; based on social media data, Lampos and Cristianini (2010) monitored flu pandemic and Lampos et al. (2013) predicted election results.

To our best knowledge, the concept of text regression was first introduced by Kogan et al. (2009) who described it as: given a piece of text, predict a real-world continuous quantity associated with the text's meaning. They applied a linear model to estimate financial risks by using financial reports directly and claimed a significant outperformance compared to previous methods. Subse-

quently, several linear text regression models were proposed; to name a few: Volkova et al. (2014); Lampos et al. (2014); Preoţiuc-Pietro et al. (2015).

Although easy for interpretation and implementation, linear models rely heavily on specific selections of high-level textual representations and fail to properly capture complicated distributions. Recent successese of deep neural networks in the field of computer vision (e.g., Ledig et al. (2017) and Liu et al. (2019)) encourage reseachers to discover their potential in natural language processing. Unlike image synthesis, using deep networks for natural language generation (NLG) is notoriously difficult (Li et al., 2018a), as the feature space of a sentence is discrete and thereby discontinuous and non-differentiable. Kusner and Hernández-Lobato (2016) attacked this issue by using one-hot vectors obtained from softmax function for backpropergation. Lin et al. (2017) used ranking scores instead of real/fake prediction for the objective function of the discriminator.

Our idea of using GANs for text regression was inspired by recent advances in NLG (e.g., Yu et al. (2017) and Li et al. (2017)). We further shift the focus from realistic language synthesis to the generation of adversarial samples from a LSTM (Hochreiter and Schmidhuber, 1997), who competes against a discriminator for regression (see Figure 1). The performance of our model is guaranteed by deep neural networks' power of capturing complicated distributions especially when obtained in an adversarial manner. The capability of training with limited supervision also facilitates promising future applications.

The rest of the paper is organized as follows: in Section 2 we discuss existing text regression techniques and previous works in semi-supervised learning with GANs; the model is detailed in Section 3; we conclude the paper in Section 4 by future works.

2 Related Work

2.1 Text Regression

Previous attempts at text regression mainly focused on linear models. Kogan et al. (2009) adopted a support vector regression (SVR) (Drucker et al., 1997) in financial reports to predict the volatility of stock returns, a widely used measure of financial risk, and reported a significant outperformance compared to state-of-the-arts. To correlate movies' online reviews and corresponding revenues, Joshi et al. (2010) extracted high-level features of textual reviews and incorporated them into a elastic net model (Zou and Hastie, 2005). Lampos et al. (2013) exploited a multi-task learning scheme that leverages textual data with user profiles for voting intention prediction. As mentioned earlier, linear models sometimes are oversimplified and fail to properly capture real-world scenarios. Bitvai and Cohn (2015) proposed the first non-linear model, a deep convolutional neural network, for text regression which surpassed previous state of the art even with limited supervision.

2.2 Semi-supervised Learning

Semi-supervised learning tackles the problem of learning a mapping between data and labels when only a small subset of labels are available. Earlier approaches of generative models with semi-supervised learning consider Gaussian mixture models (Zhu, 2006) and non-parametric density models (Kemp et al., 2004), but suffer from limitations of scalability and inference accuracy. Recently Kingma et al. (2014) addresses this problem by developing stochastic variational inference algorithms for join optimization of model and variational parameters.

Since generative adversarial networks (GANs) has been shown to be promising in generating realistic images (Goodfellow et al., 2014), several approaches have been proposed to use GANs in semi-supervised learning. Springenberg (2015) extends the discriminator (D) to be a K class classifier with objective function to minimize prediction certainty on generated images, while generator aims for maximize the same objective. Odena (2016) augments the K class discriminator to include a $K + 1$ label as fake for the generated images. These work have shown that incorporating adversarial objectives can make the learning of classifier robust and data efficient. While previous works mainly focus on classification setting, in our work, we extend the GAN based semi-supervised learning to regression task.

3 The TR-GAN Model

In this section, we detail the conditional generative adversarial network for text regression in a semi-supervised setting (TR-GAN). We first introduce the word embedding method.

3.1 Word Embedding

Word embedding method learns a high dimension representation for each word, thereby incorporate semantic information that cannot be captured by the single token. In our work, we adopted pre-trained word embedding for each word in the text input. Then each document in data can be represented by a $D \times N$ matrix, where D is the number of words in the document and N is the dimension of word embedding in the pretrained model.

3.2 Model Architecture

As illustrated in Figure 1, the network architecture is a conditional GAN with a generator and a discriminator. A long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) is deployed as the generator for natural languages. As the embedding is fed into LSTM, the generator is a LSTM-based sentence decoder. The discriminator is a convolutional neural network (CNN) (Kalchbrenner et al., 2014), where serval residual blocks (He et al., 2016) are followed by batch normalization with $ReLU$ as the activate function. Subsequently, two fully connected layers are finalized for adversarial learning and the regression task.

The objective function adopt mean absolute error (MAE) for regression tasks and adversarial loss for sequence generation. Not only can this model generate realistic sentences through the optimized generator but the discriminator is also trained as a regression model for multiple prediction tasks (e.g., auto sales prediction, public opinion tracking, and even epidemiological surveillance from social media), which are of great interest to a wide range of stakeholders.

4 Future Work

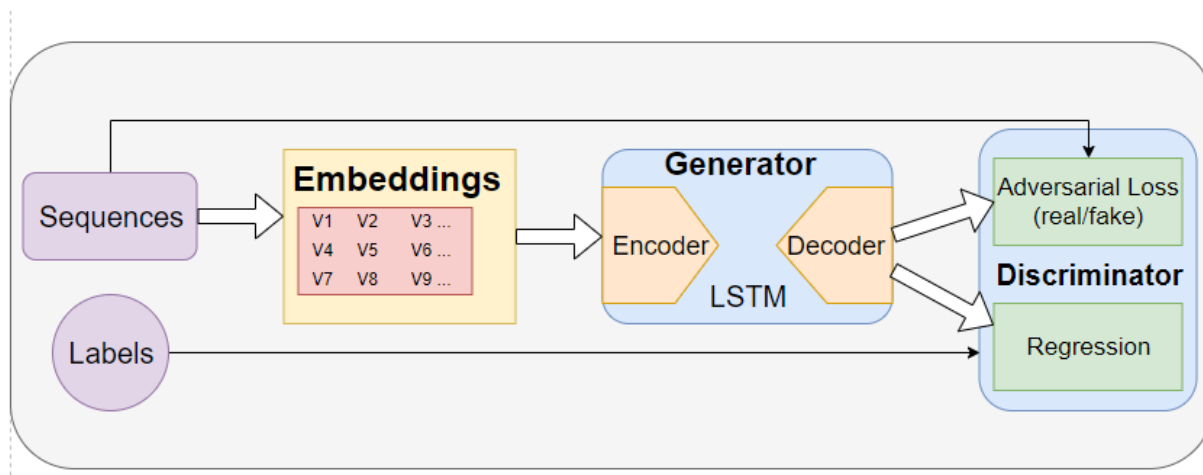We are excited about the idea of using GANs for text regression. Given the nature of the TR-GAN

Figure 1: Architecture of the TR-GAN model.

model, it is not challenging to find an experimental dataset; for example, Li et al. (2018b) collected 50,000 textual comments below YouTube videos, among which 20,000 are labelled by state-of-the-art algorithms and 1,000 are labelled manually. We also are interested to see how the generated languages look like, given that existing literatures of using GANs for NLG merely report original experimental results but instead numerical metrics.

Acknowledgments

We thank the authors of Liu et al. (2019) and Wang et al. (2019) for helpful discussions and Li et al. (2018b) for collecting the dataset.

References

Zsolt Bitvai and Trevor Cohn. 2015. Non-linear text regression with a deep convolutional neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 180–185.

Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A Smith. 2010. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Charles Kemp, Thomas L Griffiths, Sean Stromsten, and Joshua B Tenenbaum. 2004. Semi-supervised learning with trees. In *Advances in neural information processing systems*, pages 257–264.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics.

Matt J Kusner and José Miguel Hernández-Lobato. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 405–413.

Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE.

Vasileios Lampos, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. A user-centric model of voting intention from social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 993–1003.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Tao Li, Kaiming Fu, Minsoo Choi, Xudong Liu, and Ying Chen. 2018a. Toward robust and efficient training of generative adversarial networks with bayesian approximation. In *the Approximation Theory and Machine Learning Conference*.

Tao Li, Lei Lin, Minsoo Choi, Kaiming Fu, Siyuan Gong, and Jian Wang. 2018b. Youtube av 50k: an annotated corpus for comments in autonomous vehicles. In *the 13th International Joint Symposium on Artificial Intelligence and Natural Language Processing*.

Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial ranking for language generation. In *Advances in Neural Information Processing Systems*, pages 3155–3165.

Xudong Liu and Guodong Guo. 2018. Attributes in multiple facial images. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 318–324. IEEE.

Xudong Liu, Tao Li, Hao Peng, Iris Chuoying Ouyang, Taehwan Kim, Ruizhe Wang, and Guodong Guo. 2019. Mining semantic descriptions from data for beauty understanding. In *2019 IEEE Winter Conference on Applications of Computer Vision*.

Augustus Odena. 2016. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1754–1764.

Jost Tobias Springenberg. 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 186–196.

Chaojie Wang, Siyuan Gong, Anye Zhou, Tao Li, and Srinivas Peeta. 2019. Cooperative adaptive cruise control for connected autonomous vehicles by factoring communication-related constraints. In *the 23rd International Symposium on Transportation and Traffic Theory*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858.

Yizhe Zhang, Zhe Gan, and Lawrence Carin. 2016. Generating text via adversarial training. In *NIPS workshop on Adversarial Training*, volume 21.

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. 2017. Adversarial feature matching for text generation. *arXiv preprint arXiv:1706.03850*.

Xiaojin Zhu. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2(3):4.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.