Long-Tail Predictions with Continuous-Output Language Models

Shiran Dudy **Steven Bedrick**
Center for Spoken Language Understanding
Oregon Health & Science University
3181 S.W. Sam Jackson Park Rd.
Portland, Oregon, USA
{dudy,bedricks}@ohsu.edu

Abstract

Neural language models typically employ a categorical approach to prediction and training, leading to well-known computational and numerical limitations. An under-explored alternative approach is to perform prediction directly against a continuous word embedding space, which according to recent research is more akin to how categories are represented in the brain. Choosing this method opens the door for large-vocabulary language models and enables substantially smaller and simpler computational complexities. In this research we explore a different important trait - the continuous output prediction models reach low-frequency vocabulary words which we show are often ignored by the categorical model. Such words are essential, as they can contribute to personalization and user vocabulary adaptation. In this work, we explore continuous-space language modeling in the context of a word prediction task over two different textual domains (newswire text and biomedical journal articles). We investigate both traditional and adversarial training approaches, and report results using several different embedding spaces and decoding mechanisms. We find that our continuous-prediction approach outperforms the standard categorical approach in terms of term diversity, in particular with rare words.

1 Introduction

In recent years neural approaches to language modeling have demonstrated substantial improvements in performance (Melis et al., 2018; Merity et al., 2018), and the latest techniques produce high-quality predictions across many benchmarks (Peters et al., 2018; Devlin et al., 2018). According to (Jozefowicz et al., 2016) "the best (language) models are the largest we were able to fit into a GPU memory" suggesting that good model performance is conditioned on the access to heavy computational resources. This performance comes at a price: most current SotA models employ deep architectures that are computationally complex and require a significant number of parameters to be learned. One major reason for this is that traditional approaches to language modeling (both neural and otherwise) model the task as *categorical* prediction: given a history of discrete symbols (words, subword units, etc.), from a finite vocabulary V, predict the next symbol in a sequence. Recent studies (Huth et al., 2012; Huth et al., 2016), however, reveal that categories in the semantic space of our brain are organized in a distributed fashion [1]

Inspired by meaning representation of categories in the brain, in this work we investigate the retrieval process of continuous representation. We assume that such representations exist, and propose a mechanism to *retrieve* them. This mechanism will be operating in the form of a language model, in a word prediction task. We will compare the "traditional" approach of retrieval of terms by classification, to retrieval through generation of the location of a desired term in a dense and continuous space. The generated location will indicate the vicinity of the model's predicted category, and will be mapped to a specific category representation of a term that is found in the continuous space. To accomplish this, we follow the architecture proposed in (Kumar and Tsvetkov, 2019) and develop a GAN on top of their proposed model. GANs (Goodfellow, 2016) are known to be an effective generative approach in images, as well as other domains (Pascual et al., 2017) that are represented in a continuous fashion, which was

[1] (Huth et al., 2012) hypothesize that the distributed representation has evolved due to efficiency considerations of storage noting that the brain "represents diversity of categories in a compact space"

the driving reason for developing this architecture for our problem. In the next sections we describe the methods and metrics by which we evaluate the models, the models' architectures, and our experimental evaluation, and discuss ways to further develop this continuous approach to word prediction.

2 Methods

We conducted our word prediction experiments across two distributionally-different domains: the annotated English Gigaword corpus (Ferraro et al., 2018, LDC2018T20) (NYT), and full-text biomedical journal articles from the open-access subset of PubMed Central (Beck, 2010) (PMC) containing a longer-tailed distribution of words. For experimental purposes, we assume a fixed [2] set of continuous representations; for each domain, we trained a corpus-specific set of word embeddings using word2vec (Mikolov et al., 2013). We use these embeddings throughout the experiments, and their associated vocabulary entries serve as locations to be predicted by the continuous models. We explored embedding dimensionalities of both 50 and 200. The model's prediction of a point in the embedding space is decoded to a specific lexical item via a nearest-neighbor technique, as well as with an experimental feature-based augmentation (see Sec. 3). The locations learned by the model are the target terms themselves, so a successful model guess occurs when the target embedding was found closest to the location predicted by the model.

Models We developed and evaluated three different families of model architecture. First, a simple categorical-prediction baseline (`ctg`), consisting of an LSTM encoder topped with a softmax classification layer and trained with a cross-entropy loss function. Second, a continuous model (`c`) with a similar architecture aside from its final layer, which instead is a dense and fully-connected layer with the same dimensionality as the input embedding space; we experimented with several loss functions for `c`. Third, we used a GAN-based approach (`G`) that employed the `c` model as its generator, and was trained together with a discriminator `D`. `D` internally imitated `G`, but was also provided with either a genuine ("real") or predicted ("fake") embedding from `G`, following the approach of (Mirza and Osindero, 2014). Inside `D`, the generated embedding $\hat{e}_D$ is compared to $e_{real,fake}$ in order to discern the authenticity of the embedding as described in Eq. 1.

$$D_t = \sigma\left((\hat{e}_D - e_{real,fake})^T\theta + b\right) \tag{1}$$

The dynamic of the proposed GAN is described in Eq. 2.

$$\min_G \max_D L(D,G) = \mathbb{E}_{w\sim p_{data}(w)}[\log D(w_t|w_{history})] + \mathbb{E}_{\hat{w}\sim p_{\hat{w}}(\hat{w})}[\log(1 - D(G(\hat{w}_t|w_{history})))] \tag{2}$$

Evaluation Our experimental focus was on the performance of our models at a word prediction task, and in this work we propose a new metric to look into an often overlooked behaviour of neural language models: their tendency to ignore infrequently-observed vocabulary entries (Holtzman et al., 2020). Our proposed metric (*types*) describes how many correctly-predicted unique vocabulary types were retrieved, though we do also measure *tokens*, describing whether the model guessed a correct/incorrect token. This maps more closely to traditional accuracy metrics of word prediction models. Our interest in this work is primarily on our models' performance at predicting infrequent *types* to measure long tail performance. We also report results using our simple nearest-neighbor decoding approach, as well as our augmented approach. In our evaluation, all models were evaluated for their top-1 and top-10 performance, analogous to text entry prediction in our smartphones (in which top-1 accuracy may not be essential).

3 Results

Table 1 describes overall results of token- and type-level performance. We employed several baselines, including `freq`, which predicting top 10 most common words in train set, and `ugrm`, which sampled from a unigram distribution learned from the training set. `freq` shows that by blindly predicting the 10 most common types, the token-level prediction accuracy reaches 23.39. The `ugrm` baseline performs lower on token accuracy but correctly predicts more types. These serve as lower-bound baselines.

Table 1 indicates that with dimensionality 50, the token prediction is higher for `ctg`, whereas the type prediction is greater in the continuous approaches. All models improve token prediction in dim 200, and type difference is smaller, yet exist (T_{10}). Note that `G` is more diverse than `c`.

[2] In Sec. 4 we address a possible relaxation of that assumption.

Following the high-level analysis in Table 1, Figure 1 describes types stratified by frequency into high-, mid-, and low-frequency bins, and the models' correctly predicted types in each (stratifying T_1). The `ctg` presence is very limited in the mid and low bins, whereas the continuous ones get up to $3,216$ types in the mid-bin, and 700 types in the low-bin. One possible explanation for this dynamic is through the bias-variance perspective, while the `ctg` are more biased towards frequency, the continuous approaches lean towards increasing variance. A similar dynamic may explain the difference in `c` and `G`. The experiments show that the continuous approaches catch the long tail more optimally. A similar, yet stronger pattern was shown in the Pubmed experiment results.

We next explored ways to further enhance the performance of the continuous approaches. One fundamental property of the described embedding space is that words sharing similar context are likely to be found closer together; hence, to some meaningful degree, the space is organized by semantic relations according to the principles of distributional semantics (Harris, 1954). We were curious whether adding *syntactic* information could enhance the underlying results. For the purposes of this evaluation, we employed Part of Speech tags as oracles, such that given a prediction, we augment the search by providing a Part-of-Speech oracle to filter out candidates. We justify our use of an oracle in this paper's experiments as follows: PoS prediction in English is a much more highly-constrained problem than full word prediction, and can be done with simple and straightforward models to a very high level of accuracy.

model	top_1 (top_{10})	T_1 (T_{10})
`freq`	00.89 (23.39)	1 (10)
`ugrm`	00.71 (08.46)	2,190 (5,592)
$\texttt{ctg}_{50}$	19.19 (46.02)	3,982 (7,559)
$\texttt{c}_{50}$	17.31 (28.70)	8,917 (22,509)
$\texttt{G}_{50}$	16.46 (27.45)	11,534 (27,921)
$\texttt{ctg}_{200}$	21.21 (47.87)	4,163 (7,683)
$\texttt{c}_{200}$	18.94 (30.90)	4,335 (13,087)
$\texttt{G}_{200}$	18.15 (29.15)	6,194 (17,905)

Table 1: NYT tokens top_x , types T_x

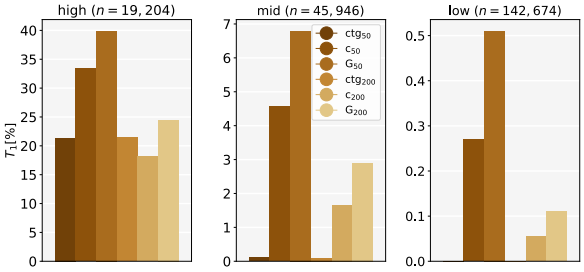

Figure 1: NYT *type* coverage by training frequency bin. n: number of items in each bin; y-axes are percentages over n (note different scales).

model	top_1 (top_{10})	T_1 (T_{10})
$\texttt{cg}_{50_p}$	29.30 (51.38)	4,809 (8,607)
$\texttt{c}_{50_p}$	20.09 (31.83)	10,326 (25,616)
$\texttt{G}_{50_p}$	19.56 (30.30)	13,540 (32,140)
$\texttt{cg}_{200_p}$	30.00 (52.16)	4,752 (8,420)
$\texttt{c}_{200_p}$	21.73 (32.93)	5,348 (15,611)
$\texttt{G}_{200_p}$	20.91 (31.06)	7,659 (21,221)

Table 2: NYT PoS decoding

Table 2 shows that this decoding technique behaved differently depending on the model. For the categorical models it advances the hit rate of tokens, while for the continuous approaches it enhances the diversity of types predicted, reinforcing our explanation of an inherently different mechanism of prediction of the two different approaches.

4 Discussion

In this work we presented an architecture for retrieval of terms, that was successfully retrieving more types (including in the long-tail region) than its categorical counterpart. Predicting the long-tail is key in personalization that targets the terms of mid to low frequency expressed by the user. In addition, we proposed PoS decoding which is a way to improve decoding of the continuous approaches. Beyond the goal of performance improvement, the proposed decoding served as a useful diagnostic tool revealing more about the different prediction nature of the continuous approaches and the categorical one. Learning locations in the embedding space could be further improved, one way is by developing ways to add more embeddings (of different domains) to an existing space. We plan to explore adaptation with the continuous approaches as these models are not bounded to a finite set of terms. We also would look into combining both approaches as they were found to be complementary in behaviour.

References

Jeff Beck. 2010. Report From the Field: PubMed Central, an XML-Based Archive of Life Sciences Journal Articles. In *International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML, Montréal, Canada*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Francis Ferraro, Max Thomas, Wolfe Travis Gormley, Matthew R., Craig Harman, and Benjamin Van Durme. 2018. Concretely Annotated English Gigaword. *Linguistic Data Consortium, Philadelphia*, 4:1.

Ian Goodfellow. 2016. Nips 2016 tutorial: Generative Adversarial Networks. *arXiv preprint arXiv:1701.00160*.

Zellig S Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *ICLR*.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. 2012. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories Across the Human Brain. *Neuron*, 76(6):1210–1224.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural Speech Reveals the Semantic Maps that Tile Human Cerebral Cortex. *Nature*, 532(7600):453–458.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the Limits of Language Modeling. *arXiv preprint arXiv:1602.02410*.

Sachin Kumar and Yulia Tsvetkov. 2019. Von Mises-Fisher Loss for Training Sequence to Sequence Models with Continuous Outputs. In *International Conference on Learning Representations*.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*.

Santiago Pascual, Antonio Bonafonte, and Joan Serrà. 2017. Segan: Speech Enhancement Generative Adversarial Network. *Proc. Interspeech 2017*, pages 3642–3646.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.