Dense Node Representation for Geolocation

Tommaso Fornaciari, Dirk Hovy
Bocconi University, Milan, Italy
{fornaciari|dirk.hovy}@unibocconi.it

Abstract

Prior research has shown that geolocation can be substantially improved by including user network information. While effective, it suffers from the curse of dimensionality, since networks are usually represented as sparse adjacency matrices of connections, which grow exponentially with the number of users. In order to incorporate this information, we therefore need to limit the network size, in turn limiting performance and risking sample bias. In this paper, we address these limitations by instead using dense network representations. We explore two methods to learn continuous node representations from either 1) the network structure with `node2vec` (Grover and Leskovec, 2016), or 2) *textual* user mentions via `doc2vec` (Le and Mikolov, 2014). We combine both methods with input from social media posts in an attention-based convolutional neural network and evaluate the contribution of each component on geolocation performance. Our method enables us to incorporate arbitrarily large networks in a fixed-length vector, without limiting the network size. Our models achieve competitive results with similar state-of-the-art methods, but with much fewer model parameters, while being applicable to networks of virtually any size.

1 Introduction

Current state-of-the-art methods for user geolocation in social media rely on a number of data sources. Text is the main source, since people use location-specific terms (Salehi et al., 2017). However, research has shown that text should be augmented with network information, since people interact with other from their local social circles. Even though social media allows for worldwide connections, most people have a larger number of connections with people who live close-by (from their school, workplace, or friend network). The most successful predictive models are therefore architectures that combine these different kinds of inputs (Rahimi et al., 2018; Ebrahimi et al., 2018).

However, incorporating network information is the computational bottleneck of these hybrid approaches: we want to represent the whole network, but we have to do so efficiently. We show that both are possible with dense representations, and indeed improve performance over previous sparse graph network representations. Following graph theory (Bondy et al., 1976), networks are typically represented as connections between entities in a square adjacency matrix, whose size corresponds to the number of users in the network. This means, though, that the matrix grows quadratically with the number of nodes/users. For large-scale social media analysis, where the number of users is often in the millions, this property creates a computational bottleneck: Incorporating such a matrix in a neural architecture, for example through graph-convolution (Kipf and Welling, 2017), easily increases the parameters by orders of magnitude, making training more expensive and increasing the risk of overfitting.

Previous work has therefore usually resorted to sampling methods. While sampling addresses the space issue, it necessarily loses a large amount of information, especially in complex networks, and introduces the risk of sampling biases.

Compounding the problem is the fact that adjacency matrices, despite their size, are very sparse, and do not represent information efficiently. This problem is analogous to sparse word and text representations, which were successfully replaced by dense embeddings (Mikolov et al., 2013a).

We show how to incorporate dense network representations in two ways: 1) with an existing word2vec-based method based on network structure, node2vec (Grover and Leskovec, 2016), and 2) with a new, doc2vec-based method of document

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 224–230

representations (Le and Mikolov, 2014) over the set of user mentions in the text of posts (M2V). Both allow us to represent mentions as dense vectors that encode the network interactions so that similar users will have similar representations. However, they capture different aspects of interactions: people we are connected with vs. people we mention.

We compare the geolocation performance of models that combine a text view with the network views of both node2vec *and* the doc2vec-based method. We measure the contribution of each component to performance. Our results show that dense network representations significantly improve over sparse network representations, but that mention representations (M2V) are more important than structure representations (node2vec).

Contributions The contributions of the study are the following:

- We propose a document embeddings application that builds effective network representations through dense vectors, with no need of sampling procedures even in large networks;
- We show that the node representations can be tuned via two parameters which model the width and strength of their interactions.

2 Related work

Different kinds of data sources and methods can be used for the geolocation of users in Social Media. The the most straightforward approach is to exploit the geographic information conveyed by the linguistic behavior of the user. The first studies relied on the idea of exploiting *Location-Indicative Words* (LIW) (Han et al., 2012, 2014). More recently, neural models have been applied to the same strategy (Rahimi et al., 2017; Tang et al., 2019), improving performance.

The problem, however, can be modeled in different ways, including the different designs of the geographic areas to predict, such as grids (Wing and Baldridge, 2011), hierarchical grids (Wing and Baldridge, 2014), or different kinds of clusters (Han et al., 2012, 2014). In this paper, we test our models both on the set of geographic areas - i.e., labels - used in the shared task of the Workshop on Noisy User-generated Text - W-NUT (Han et al., 2016), and the more fine-grained clusters obtained through the method of Fornaciari and Hovy (2019b). Geographic coordinates themselves can also be exploited, as Fornaciari and Hovy (2019a) showed in a multi-task model that jointly predicts continuous geocoordinates and discrete labels.

In general, geolocation with multi-source models is becoming more popular, as indicated by their increased use in state-of-the-art performances. Miura et al. (2016, 2017) considered text, metadata and network information, modeling the last as a combination user and city embeddings. Similarly to our study, Rahimi et al. (2015) exploited the mentions, even though they used them to build undirected graphs. Ebrahimi et al. (2017, 2018) also used mentions to create an undirected graph, that they pruned and fed into an embedding layer followed by an attention mechanism, in order to create a network representation.

The study of Rahimi et al. (2018) is an example of network segmentation for use in a neural model. They propose a Graph Convolutional Neural Network (GCN), where network and text data are vertically concatenated in a single channel, rather than employed as parallel channels into the same model. Do et al. (2017, 2018) present the Multi-Entry Neural Network (MENET), a model which, similarly to our study, employs node2vec and, separately, includes doc2vec as methods for extraction of document features.

These works represent the state-of-the-art benchmark with respect to the implementation of network views in the models. Other models (Ebrahimi et al., 2017, 2018; Do et al., 2018) also include metadata or other source of information.

3 Methods

3.1 The data sets

We test our methods on three data sets: GEOTEXT (Eisenstein et al., 2010), TWITTER-US (Roller et al., 2012) and TWITTER-WORLD (Han et al., 2012). They contain English tweets, concatenated by author, with geographic coordinates associated with each author. GEOTEXT contains 10K texts, TWITTER-US 450K and TWITTER-WORLD 1.390M. The corpora are each split into training, development and test sets.

3.2 Learning network representations

3.2.1 *node2vec*

Grover and Leskovec (2016) presented node2vec, a method to obtain dense node representations through a skip-gram model. Those representations, however, are obtained through a tiered sam-

pling procedure. While that allows node2vec to explore large networks, by balancing the breadth and depth of the search for the neighbours' identification, it does introduce a random factor. In addition, since node2vec uses the word2vec skip-gram model (Mikolov et al., 2013c), the sequence of the nodes does not carry any meaning, essentially functioning as a further random neighbors selection. In the geolocation scenario, though, network breadth is more important than depth, as similarity between entities grows with their proximity: we would like to preserve this is information entirely, even and especially in large networks. For this reason, we follow the authors settings for the detection of nodes' homophily, rather than their structural similarity in the network, and set the *node2vec* parameters $p = 1$ and $q = 0.5$ (Grover and Leskovec, 2016, p. 11).

3.2.2 *mentions2vec* - M2V

We introduce a novel network representation method which does not depend on graph theory. We bypass the adjacency matrices and directly learn the social interactions from the *content* of social media messages. In many social media this is straightforward, as the users' mentions are introduced by the at sign '@', but in general other forms of Named Entity Recognition (NER) might be considered for the same purpose.

Concretely, we filter from the text everything but the user mentions and apply *doc2vec* to the resulting "texts" (Mikolov et al., 2013b). Basically, we are representing the users according to their communicative behavior directed at other users, in the temporal order these mentions appear in. Therefore, similarly to node2vec, M2V creates a dense representation of the user interactions.

As pointed out earlier, node2vec is applied to a sequence of nodes sampled form the whole network that does not account for temporal ordering. In contrast, M2V does not address nodes, but mentions, which are themselves an evidence of personal connection. The consequence of this choice is two-fold. First, there is no need for a sampling procedure: the whole set of interactions can be considered, even for wide networks. Second, the order of the mentions in the texts reflects the time sequence of the interactions, possibly encoding patterns of social behaviors.

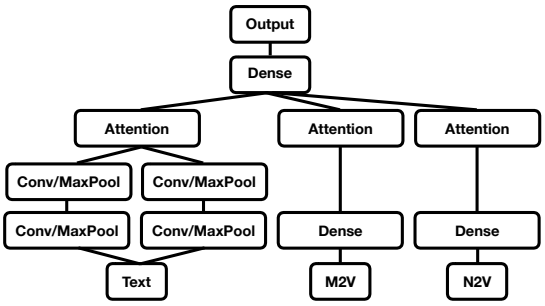


Figure 1: The Multiview Attention-based Convolutional Model. The inputs are the texts, mentions2vec (M2V) and *node2vec* (N2V)

3.3 Labels

For our experiments we use two different sets of labels: those used in the W-NUT 2016 task (Han et al., 2016), and our own labels (Fornaciari and Hovy, 2019b). Our label identification method, called point2city (P2C), clusters all points closer than 25 km and associates each cluster with the closest town of at least 15K people. For further details, see Fornaciari and Hovy (2019b). The resulting labels are highly granular and precise in the identification of meaningful administrative regions.

3.4 Feature selection

The label sets were involved in the preprocessing as follows. Using only the training data, we first select the terms with frequency greater or equal to 10 and 5 for TWITTER-US and TWITTER-WORLD, respectively. This choice is motivated by the different vocabulary size of the two data sets. Any term with frequency greater than 2, but below these thresholds, which is associated with only one label, we replace with label-representative tokens. Low-frequency terms found in more than one place are considered geographically ambiguous and discarded. This allows us to reduce remarkably the vocabulary size, maintaining the useful geographic information of the huge amount of low frequency terms. Considering the terms' Zipf distribution (Powers, 1998), this procedure allows us to replace a small number of types, but a great number of tokens.

Following Han et al. (2012), we further filter the vocabulary by applying Information Gain Ratio (IGR), selecting the terms with the highest values until we reach a manageable vocabulary size: 750K and 470K for TWITTER-US and TWITTER-WORLD.

Authors, method + labels	nr. labels	TWITTER-US Acc	Acc@161	mean	median
Han et al. (2014), NB + IGR	378	26%	45%	-	260
Wing and Baldridge (2014), HierLR k-d	"fewer classes"	-	48%	687	191
Rahimi et al. (2017), MLP + k-tree	256	-	55%	581	91
AttCNN + k-d tree	256	27.86%	57.85%	565.64	64.25
AttCNN-N2V + k-d tree	256	28.2%	56.99%	550.77	68.41
AttCNN-M2V + k-d tree	256	29.84%*	56.77%	546.97	67.91
AttCNN-M2V-N2V + k-d tree	256	29.12%	56.16%	563.52	77.63
AttCNN + P2C	914	51.22%	61.97%	523.42	0
AttCNN-N2V + P2C	914	51.9%	62.36%	518.34	0
AttCNN-M2V + P2C	914	53.04%**	63.64%**	483.09*	0
AttCNN-M2V-N2V + P2C	914	52.93%**	62.91%**	510.98**	0

Table 1: Model performance and significance levels with respect to text-only models: $^{*}: p \leq 0.05$, $^{**}: p \leq 0.01$

Authors, method + labels	nr. labels	TWITTER-WORLD Acc	Acc@161	mean	median
Han et al. (2014), NB + IGR	3135	13%	26%	-	913
Wing and Baldridge (2014), HierLR k-d	"fewer classes"	-	31%	1670	509
Rahimi et al. (2017), MLP + k-tree	930	-	36%	1417	373
AttCNN + k-d tree	930	20.0%	36.39%	1458.63	414.29
AttCNN-N2V + k-d tree	930	22.3%**	40.02%**	1363.11**	330.69**
AttCNN-M2V + k-d tree	930	29.26%**	46.05%**	1155.5**	230.17**
AttCNN-M2V-N2V + k-d tree	930	28.76%**	46.31%**	1191.19**	223.96**
AttCNN + P2C	2818	28.39%	42.5%	1195.92	274.06
AttCNN-N2V + P2C	2818	28.48%	42.18%	1220.02	280.66
AttCNN-M2V + P2C	2818	34.58%**	47.91%**	1134.08**	194.03**
AttCNN-M2V-N2V + P2C	2818	33.98%**	47.19%**	1180.21**	208.03**

Table 2: Model performance and significance levels with respect to text-only model AttCNN: $^{*}: p \leq 0.05$, $^{**}: p \leq 0.01$

3.5 Multiview Attention-based Convolutional Models

Our models are multi-view neural networks with three input channels: the text view, *node2vec*, and *mentions2vec*. The text view, in turn, contains two channels of convolutional/max pooling layers (with window size 4 and 8) followed by an attention mechanism. Both *node2vec* and *mentions2vec* are fed into a dense layer, followed by an attention mechanism. All the outputs are then concatenated and fed into a fully connected output layer. For a graphical representation, see Figure 1.

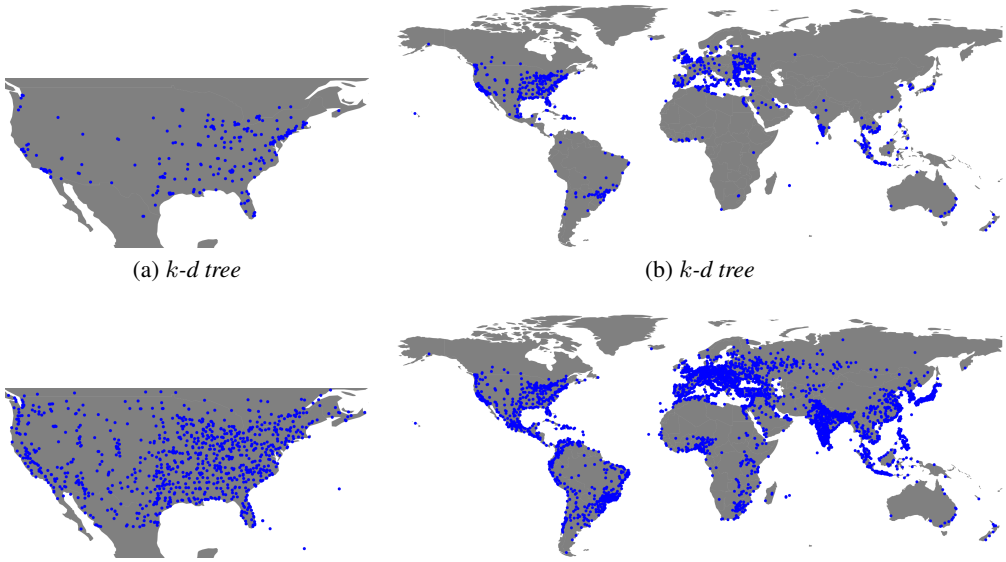
We report the performance metrics commonly considered in the literature: accuracy, acc@161 - i.e., the accuracy within 161 km, or 100 miles, from the target point. This allows us to measure the accuracy of predictions within a reasonable distance from the target point. We also report mean and median distance between the predicted and the target points. We evaluate significance via bootstrap sampling, following Søgaard et al. (2014). The code for the methods described in this paper are available at github.com/Bocconi-NLPLab.

4 Results

Tables 1 and 2 show the performance of our models with and without N2V/M2V, in TWITTER-US and TWITTER-WORLD. Compared to the previous studies using only textual features, our basic model AttCNN shows comparable (TWITTER-WORLD) or better performance (TWITTER-US).

Therefore we consider our base AttCNN model as baseline comparison for the hybrid models AttCNN-N2V, AttCNN-M2V and AttCNN-M2V-N2V. We test two label sets (k-d tree and P2C), and the significance level remarkably changes in these two conditions.

In TWITTER-US, with coarse granularity labels, there is no performance improvement with dense node representations. In contrast, the models with M2V show a significant effect with fine



(a) *k-d tree* (b) *k-d tree*

(c) *P2C* (d) *P2C*

Figure 2: Labels' coordinates in TWITTER-USand TWITTER-WORLD

granularity labels. In TWITTER-WORLD, the dense node representations significantly improve the models' performance, with both kind of labels, even though AttCNN-N2V does not show improvements with P2C labels.

5 Discussion

Mentions2vec is a computationally affordable method for dense network representations, designed to capture social interactions. It proves very effective under most experimental conditions. The results suggest that dense users' network representation enhance geolocation performance, in particular when fine-grained labels identify specific geographic areas, rather than when a small number of labels refers to larger areas, where more different social communities can be found. Figure 2 shows the different density of labels identified by k-d tree and P2C. These settings are particularly useful for M2V, which considers the users' linguistic behavior. In contrast, *Node2vec* does not lead to significant improvement in TWITTER-US, presumably because the sampling procedure of *node2vec* does not allow to detect homophily with sufficient clarity. *Mentions2vec*, which does not suffer from this limitation, appears to be more effective in that context. However, in general, the labels' granularity affects the usefulness of the methods. In TWITTER-US, using labels which cover large areas is detrimental for techniques which address geographical homophily, that is, relatively small cultural/linguistic areas. Even so, it makes sense to use these techniques, as in favourable conditions (for example in TWITTER-WORLD), they lead to remarkable performance improvements.

Acknowledgments

The authors would like to thank the reviewers of the various drafts for their comments. Both authors are members of the Bocconi Institute for Data Science and Analytics (BIDSA) and the Data and Marketing Insights (DMI) unit. This research was supported by a GPU donation from Nvidia, as well as a research grant from CERMES to set up a GPU server, which enabled us to run these experiments.

References

John Adrian Bondy, Uppaluri Siva Ramachandra Murty, et al. 1976. *Graph theory with applications*, volume 290. Citeseer.

Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. 2017. Multiview deep learning for predicting twitter users' location. *arXiv preprint arXiv:1712.08091*.

Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. 2018. Twitter user geolocation using deep multiview learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6304–6308. IEEE.

Mohammad Ebrahimi, Elaheh ShafieiBavani, Raymond Wong, and Fang Chen. 2017. Exploring celebrities on inferring user geolocation in twitter. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 395–406. Springer.

Mohammad Ebrahimi, Elaheh ShafieiBavani, Raymond Wong, and Fang Chen. 2018. A unified neural network model for geolocating twitter users. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 42–53.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287. Association for Computational Linguistics.

Tommaso Fornaciari and Dirk Hovy. 2019a. Geolocation with Attention-Based Multitask Learning Models. In *Proceedings of the 5th Workshop on Noisy User-generated Text (WNUT)*.

Tommaso Fornaciari and Dirk Hovy. 2019b. Identifying Linguistic Areas for Geolocation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (WNUT)*.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012*, pages 1045–1062.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter Geolocation Prediction Shared Task of the 2016 Workshop on Noisy User-generated Text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2016. A simple scalable neural networks based model for geolocation prediction in twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 235–239.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1260–1272.

David MW Powers. 1998. Applications and explanations of zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*, pages 151–160. Association for Computational Linguistics.

Afshin Rahimi, Trevor Cohn, and Tim Baldwin. 2018. Semi-supervised user geolocation via graph convolutional networks. *arXiv preprint arXiv:1804.08049*, pages 2009–2019.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015. Twitter user geolocation using a unified text and network prediction model. *arXiv preprint arXiv:1506.08259*.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*, pages 209–216.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics.

Bahar Salehi, Dirk Hovy, Eduard Hovy, and Anders Søgaard. 2017. Huntsville, hospitals, and hockey teams: Names can reveal your location. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 116–121.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Héctor Martínez Alonso. 2014. What's in a p-value in nlp? In *Proceedings of the eighteenth conference on computational natural language learning*, pages 1–10.

Haina Tang, Xiangpeng Zhao, and Yongmao Ren. 2019. A multilayer recognition model for twitter user geolocation. *Wireless Networks*, pages 1–6.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 336–348.

Benjamin P Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 955–964. Association for Computational Linguistics.