Geolocation with Attention-Based Multitask Learning Models

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

Abstract

Geolocation, predicting the location of a post based on text and other information, has a huge potential for several social media applications. Typically, the problem is modeled as either multi-class classification or regression. In the first case, the classes are geographic areas previously identified; in the second, the models directly predict geographic coordinates. The former requires discretization of the coordinates, but yields better performance. The latter is potentially more precise and true to the nature of the problem, but often results in worse performance. We propose to combine the two approaches in an attention-based multitask convolutional neural network that jointly predicts both discrete locations and continuous geographic coordinates. We evaluate the multi-task (MTL) model against single-task models and prior work. We find that MTL significantly improves performance, reporting large gains on one data set, but also note that the correlation between labels and coordinates has a marked impact on the effectiveness of including a regression task.

1 Introduction

Knowing the location of a social media post is useful for a variety of applications: from improving content relevance for the socio-cultural environment of a geographic area (Rakesh et al., 2013), to the understanding of demographic distributions for disaster relief (Lingad et al., 2013).

However, most social media posts do not include location. On Twitter, one of the most studied social media, geotagging is enabled for at most 5% of the posts (Sloan and Morgan, 2015; Cebeillac and Rault, 2016). In order to address this issue, samples of geolocated data have been used to create corpora of geo-tagged texts. Those corpora allow us to train supervised models to predict the geographic location for a post, relying on the post's text and, possibly, users' interaction information and other meta-data provided by the social media. While a lot of work has gone into this problem, it is still far from solved.

The task is usually framed as a multi-class *classification* problem, but actual location information is normally given as a pair of continuous-valued latitude/longitude coordinates (e.g.: $51.5074°$ N, $0.1278°$ W). Using these coordinates in classification requires translation into labels corresponding to a geographic area (e.g., cities, states, countries). This translation is another non-trivial task (Wing and Baldridge, 2014), and necessarily loses information. Much less frequently, geolocation is framed as *regression*, i.e., direct prediction of the coordinates. While potentially more accurate, regression over geographic coordinates presents a host of challenges (values are continuous but bounded, can be negative, and distances are non-Euclidean, due to the Earth's curvature). It is therefore usually less effective than classification.

Ideally, we would like to combine the advantages of both approaches, i.e., let the regression over continuous-valued coordinates guide the discrete location classification. So far, however, no work has tried to combine the two approaches. With recent advances in multi-task learning (MTL), we have the opportunity to combine them. In this paper, we do exactly that.

We combine classification and regression in a multi-task attention-based convolutional neural network (MTL-Att-CNN), which jointly learns to predict the geographic labels *and* the relative coordinates. We evaluate on two data sets widely used in the geolocation literature, TWITTER-US and TWITTER-WORLD (Section 3). In line with prior research on MTL (Alonso and Plank, 2017; Bingel and Søgaard, 2017), we do find that auxiliary regression can indeed help classification performance, but under a somewhat surprising con-

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 217–223
Hong Kong, Nov 4, 2019.

dition: when there are enough classification labels. We show this by evaluating on two different schemes for discretizing coordinates into labels. The first (Rahimi et al., 2017b) identifies irregular areas via k-d trees, and is the most common in the literature. The second (Fornaciari and Hovy, 2019b) directly identifies towns of at least 15K inhabitants and allows the evaluation of the method in a more realistic scenario, but results in 3–6 times more labels.

Contributions 1) We propose a novel multi-task CNN model, which learns geographic label prediction and coordinate regression together. 2) Based on Fornaciari and Hovy (2019b), we propose an alternative coordinate discretization, which correlates more with geocoordinates (Section 3). We find that label granularity impacts the effectiveness of MTL.

2 Related Work

Most successful recent approaches to geolocation use Deep Learning architectures for the task (Liu and Inkpen, 2015; Iso et al., 2017; Han et al., 2016). Many authors (Miura et al., 2016; Bakerman et al., 2018; Rahimi et al., 2018; Ebrahimi et al., 2018; Do et al., 2018; Fornaciari and Hovy, 2019a) follow a hybrid approach, combining the text representation with network information and further meta-data. However, recent works explore the effectiveness of purely textual data for geolocation (Tang et al., 2019).

Other researchers have directly predicted the geographic coordinates associated with the texts. Eisenstein et al. (2010) was the first to formulate the problem as a regression task predicting the coordinate values as numerical values. Lourentzou et al. (2017) use very simple labels, but create a neural model which separately performs both the classification task and the prediction of the geographic coordinates. They evaluate the relative performance of each approach.

Rahimi et al. (2017a) created a dense representation of bi-dimensional points using Mixture Density Networks (Bishop, 1994). They motivate the higher complexity of such multi-dimensional representations with the limits of the loss minimization in uni-modal distributions for multi-target scenarios. In particular, they underline that minimizing the squared loss is equivalent to positioning the predicted point in the middle of the possible outputs, when more flexible representations would be useful for geographic prediction: "a user who mentions content in both NYC and LA is predicted to be in the centre of the U.S.".

We address this point with a model which jointly solves the classification and regression problem, similar to the approach by Subramanian et al. (2018), who combine regression with a classification-like "ordinal regression" in order to predict both the number of votes for a petition as well as the voting threshold it reaches.

There is a rich literature on the use of multi-task learning (Caruana, 1996; Caruana et al., 1996; Caruana, 1997) in NLP, highlighting the importance of choosing the right auxiliary tasks (Alonso and Plank, 2017; Bingel and Søgaard, 2017; Benton et al., 2017; Lamprinidis et al., 2018).

3 Data

Corpora We use two publicly available data sets commonly used for geolocation, known as TWITTER-US and TWITTER-WORLD. They were released by Roller et al. (2012) and Han et al. (2012) respectively. Both data sets consist of geolocated tweets written in English, coming from North America and from everywhere in the World. Each instance consists of a set of tweets from a single user, associated with a pair of geographic coordinates (latitude and longitude). TWITTER-US has 449 694 instances, TWITTER-WORLD 1 386 766. Both corpora have predefined development and test sets of 10 000 records each. These corpora were used in the shared task of W-NUT 2016, providing the basis for comparison with other models in the literature.

Labels Since the location is represented as coordinates, there is no single best solution for translating them into meaningful labels (i.e., geographic areas). We follow two distinct discretizing approaches, resulting in different label sets. First, to allow comparison with prior work, we implement the coordinate clustering method proposed by Rahimi et al. (2017b). It relies on the k-d tree procedure (Maneewongvatana and Mount, 1999) and led to the identification of 256 geographic areas for TWITTER-US and 930 for TWITTER-WORLD. These areas, however, are quite large and do not always correspond to any meaningful territorial division (e.g., city, county, state, etc).

In order to create labels sets corresponding more closely to existing geographic distinctions, we follow the Point2City - P2C, another algorithm

based on k-d tree with additional steps, proposed by Fornaciari and Hovy (2019b). This results in more fine-grained geographic labels.

P2C clusters all points closer than 11 km (which correspond to the first decimal point on the longitude axis), then iteratively merges the centroids until no centroids are closer than 11 km to each other. Finally, these points are labeled with the name of the closest city of at least 15 000 inhabitants, according to the information provided by the free database GeoNames. We refer the reader to Fornaciari and Hovy (2019b) for more details of the method.

The mean distance between P2C labels and the respective actual city centers is less than 3.5 km. P2C results in 1 593 labels for TWITTER-US and 2 975 for TWITTER-WORLD, a factor of respectively 6 and 3 greater than the method used by Rahimi et al. (2017b). We provide our labels and our models on GitHub Bocconi-NLPLab.

Pre-processing and feature selection We preprocess the text by converting it to lowercase, removing URLs and stop-words. We reduce numbers to 0, except for those appearing in mentions (e.g., @abc123). In order to make the vocabulary size computationally tractable, we restrict the allowed words to those with a minimum frequency of 5 for each corpus. Since this removes about 80% of the vocabulary, losing possibly relevant information, we convert a part of the low-frequency words into replacement tokens. In particular, considering the training set only, we selected all those appearing uniquely in the same place according to the P2C labels. We discarded the low frequency terms found in more than one geographic area. In this way, the resulting vocabulary size is 1.470M words for TWITTER-US and 470K for TWITTER-WORLD.

We follow Han et al. (2014) and Forman (2003) in limiting both vocabularies to the same number of tokens, i.e., 470K tokens, by filtering the terms according to their Information Gain Ratio (IGR). This is a measure of the degree of informativeness for each term, according to its distribution among a set of labels – geographic areas in our case.

4 Methods

We train embeddings for both corpora, and use them as input to the multi-task learning model.

Embeddings Since tweets are natively short texts further reduced by removing stop words, we use a small context window size of 5 words. We trained our embeddings on the training sets of each corpus. As we are interested in potentially rare geographically informative words, we use the skip-gram model, which is more sensitive to low-frequency terms than CBOW (Mikolov et al., 2013) and train for 50 epochs. We use an embedding size of 512, choosing a power of 2 for memory efficiency, and the size as a compromise between a rich representation and the computational tractability of the embeddings matrix. For the same reason, we limit the length of each instance to 800 words for TWITTER-US and 400 words for TWITTER-WORLD, which preserves the entire text for 99.5% of the instances in each corpus.

MTL-Att-CNN We implement a CNN with the following structure. The input layer has the word indices of the text, converted via the embedding matrix into a matrix of shape $words \times embeddings$. We convolve two parallel channels with max-pooling layers and convolutional window sizes 4 and 8 over the input. The two window sizes account for both short and relatively long patterns in the texts. In both channels, the initial number of filters is 128 for the first convolution, and 256 in the second one. We join the output of the convolutional channels and pass it through an attention mechanism (Bahdanau et al., 2014; Vaswani et al., 2017) to emphasize the weight of any meaningful pattern recognized by the convolutions. We use the implementation of Yang et al. (2016). The output consists of two independent, fully-connected layers for the predictions, respectively in the form of discrete labels for classification and of continuous latitude and longitude values for regression.

Gradient Normalization Multi-task networks are quite sensitive to the choice of auxiliary tasks and the associated loss (Benton et al., 2017). If the loss function outputs of different tasks differ in scale, backpropagation also involves errors at different scales. This can imbalance the relative contributions of each task on the overall results: the "lighter" task can therefore be disadvantaged up to the point to become untrainable, since the backpropagation becomes dominated by the task with the larger error scale. To prevent this problem, we first normalize the coordinates to the range $0 - 1$. Since coordinates include negative values, we transform them by adding 180 and dividing by

TWITTER-US						
method	model	# labels	Acc	Acc@161	mean	median
Han et al. (2014)	NB + IGR	378	26%	45%	-	260
Rahimi et al. (2017b)	MLP + k-means	256	-	**55%**	**581**	**91**
k-d labels	STL-Att-CNN	256	21.06%	44.51%	845.23†	272.15
	MTL-Att-CNN	256	20.75%	44.35%	856.60	276.99
P2C labels	STL-Att-CNN	1,593	31.22%	44.48%	944.89	304.99
	MTL-Att-CNN	1,593	**31.36%**	44.64%	889.98**	293.26

TWITTER-WORLD						
Han et al. (2014)	NB + IGR	3135	13%	26%	-	913
Rahimi et al. (2017b)	MLP + k-means	930	-	36%	1417	373
k-d labels	STL-Att-CNN	930	30.67%	48.13%	1656.06	202.68
	MTL-Att-CNN	930	30.70%	48.46%	1640.16	**195.18**
P2C labels	STL-Att-CNN	2,975	35.67%	47.95%	1695.85	203.50
	MTL-Att-CNN	2,975	**36.07%**	**48.48%***	1643.29**	195.54

Table 1: Performance of prior work and proposed model. NB= Naive Bayes, MLP=Multi-Layer Perceptron, CNN=Convolutional Neural Net, STL=Single Task, MTL=Multi Task. Significance on MTL vs. STL: $^{*}: p \leq 0.05$, $^{**}: p \leq 0.01$, $^{\dagger}: p \leq 0.005$

360. As loss function, we compute the Euclidean distance between the predicted and the target coordinates.[1] We rescale all distances to within $0-1$ as well, i.e., to the same scale as the softmax output of the classification task.

For the main task (i.e., classification), we use the Adam optimizer (Kingma and Ba, 2014). This gradient descent optimizer is widely used as it uses moving averages of the parameters (i.e., the momentum), in practice adjusting the step size during the training (Bengio, 2012). The Adam optimizer, though, requires a high number of parameter. For the auxiliary task (i.e., regression), we simply used standard gradient descent.

5 Experiments

We carry out 8 experiments, 4 on TWITTER-US and 4 on TWITTER-WORLD. For each data set, we compare the performance of multi-task (MTL) and single-task (i.e., classification) models (STL), both with the labels of Rahimi et al. (2017b) and our own label set. For each of the 8 conditions, we report results averaged over three runs to reduce the impact of the random initializations. For each condition, we compute significance between STL and MTL via bootstrap sampling (Berg-Kirkpatrick et al., 2012; Søgaard et al., 2014).

[1]We also experimented with incorporating radians into the distance measure, but did not find any particular improvement, since it is learned directly during the training process.

TWITTER-US and TWITTER-WORLD are two remarkably different data sets. Not only they address areas of different size, with different geographic density of the entities to locate, they also differ in vocabulary size (larger in TWITTER-US), even considering different pre-processing procedures. Therefore, the performance difference many studies report is not surprising.

The outcomes are shown in Table 1. On both data sets, MTL yields the best results for exact accuracy. On TWITTER-US, we outperform Han et al. (2014) in exact accuracy, but cannot compare to Rahimi et al. (2017b), and do not reach their acc@161 or distance measures. For TWITTER-WORLD, we report the best results for both types of accuracy and median distance. Interestingly, mean distance is higher, suggesting a very long tail of far-away predictions.

The effectiveness of MTL increases with label granularity. This makes sense, since under a more fine-grained label scheme, the correlation between coordinates and labels is higher, which is exactly what we learn in the auxiliary task. Under the broader labeling scheme by Rahimi et al. (2017b), label areas are of irregular size, and so the correlation with the coordinates varies. With the k-d tree labels, the mean distance between the coordinates and the cluster centroids is 50 Km for TWITTER-US and 40 km for TWITTER-WORLD, while with our labels the mean distance is 16 and 7 km, re-

spectively. With highly granular P2C labels, MTL consistently outperforms STL; in contrast, with wider areas, STL mean distance beats MTL in TWITTER-US. The auxiliary regression adds valuable information to the classification task: MTL improves significantly over STL.

6 Ablation study

In order to verify the impact of the network components on the overall performance, we carry out a brief ablation study. In particular, we are interested in the attention mechanism, implemented following Yang et al. (2016). To this end, we train a MTL model *without* attention mechanism. We note that they are not directly comparable to those shown in table 1, since they used different, randomly initialized embeddings, and should be interpreted with caution. The results do suggest, though, that we can expect the attention mechanism to increase performance by about 10 points percent (both for accuracy and for acc@161), and to increase median distance by about 150 km. This effect holds for both multi-task and single-task models.

7 Conclusion

IN this paper, we propose a novel multi-task learning framework with attention for geolocation, combining label classification with regression over geo-coordinates.

We find that the granularity of the labels (and their correlation with the coordinates) has a direct impact on the effectiveness of MTL, with more labels counter-intuitively resulting in higher exact accuracy. Besides the labels commonly adopted in the literature, we also evaluate with a greater number and more specific locations (arguably a more realistic way to evaluate the geolocation for many real life applications). This effect holds independent of whether the model is trained with attention or not.

The auxiliary regression task is helpful for classification when using more fine-grained labels, which address specific rather than broad areas. Our models are optimized for exact accuracy, rather than to Acc@161, and we report some of the best accuracy measures for TWITTER-WORLD, and competitive results for TWITTER-US.

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

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? Semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 44–53.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jordan Bakerman, Karl Pazdernik, Alyson Wilson, Geoffrey Fairchild, and Rian Bahran. 2018. Twitter geolocation: A hybrid approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(3):34.

Yoshua Bengio. 2012. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 152–162.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 164–169.

Christopher M Bishop. 1994. Mixture density networks. Technical report, Citeseer.

Rich Caruana. 1996. Algorithms and applications for multitask learning. In *ICML*, pages 87–95.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Rich Caruana, Shumeet Baluja, and Tom Mitchell. 1996. Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation. In *Advances in neural information processing systems*, pages 959–965.

Alexandre Cebeillac and Yves-Marie Rault. 2016. Contribution of geotagged twitter data in the study of a social group's activity space. the case of the upper middle class in delhi, india. *Netcom. Réseaux, communication et territoires*, 30(3/4):231–248.

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. 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.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305.

Tommaso Fornaciari and Dirk Hovy. 2019a. Dense Node Representation for Geolocation. 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)*.

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.

Hayate Iso, Shoko Wakamiya, and Eiji Aramaki. 2017. Density estimation for geolocation via convolutional mixture density network. *arXiv preprint arXiv:1705.02750*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sotiris Lamprinidis, Daniel Hardt, and Dirk Hovy. 2018. Predicting news headline popularity with syntactic and semantic knowledge using multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 659–664.

John Lingad, Sarvnaz Karimi, and Jie Yin. 2013. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on world wide web*, pages 1017–1020. ACM.

Ji Liu and Diana Inkpen. 2015. Estimating user location in social media with stacked denoising autoencoders. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 201–210.

Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2017. Text-based geolocation prediction of social media users with neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 696–705. IEEE.

Songrit Maneewongvatana and David M Mount. 1999. It's okay to be skinny, if your friends are fat. In *Center for Geometric Computing 4th Annual Workshop on Computational Geometry*, volume 2, pages 1–8.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. 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.

Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017a. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Empirical Methods in Natural Language Processing*. 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. 2017b. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*, pages 209–216.

Vineeth Rakesh, Chandan K Reddy, and Dilpreet Singh. 2013. Location-specific tweet detection and topic summarization in twitter. In *Proceedings of*

the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 1441–1444. ACM.

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.

Luke Sloan and Jeffrey Morgan. 2015. Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PloS one*, 10(11):e0142209.

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.

Shivashankar Subramanian, Timothy Baldwin, and Trevor Cohn. 2018. Content-based Popularity Prediction of Online Petitions Using a Deep Regression Model. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 182–188. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.