Social Media User Geolocation via Hybrid Attention

Cheng Zheng[1*], Jyun-Yu Jiang[1*], Yichao Zhou[1], Sean D. Young[2], and Wei Wang[1]
[1]Department of Computer Science, University of California, Los Angeles
[2]Departments of Emergency Medicine and Informatics, University of California, Irvine
{chengzheng,jyunyu,yz,weiwang}@cs.ucla.edu,syoung5@uci.edu

ABSTRACT

Determining user geolocation is vital to various real-world applications on the internet, such as online marketing and event detection. To identify the geolocations of users, their behaviors on social media like published posts and social interactions can be strong evidence. However, most of the existing social media based approaches individually learn from text contexts and social networks. This separation can not only lead to sub-optimal performance but also ignore the distinct importance of two resources for different users. To address this challenge, we propose a novel end-to-end framework, Hybrid-attentive User Geolocation (HUG), to jointly model post texts and user interactions in social media. The hybrid attention mechanism is introduced to automatically determine the importance of texts and social networks for each user while social media posts and interactions are modeled by a graph attention network and a language attention network. Extensive experiments conducted on three benchmark geolocation datasets using Twitter data demonstrate that HUG significantly outperforms competitive baseline methods. The in-depth analysis also indicates the robustness and interpretability of HUG.

KEYWORDS

Social media user geolocation; Attention mechanism; Graph attention; Hierarchical structure; Interpretability.

ACM Reference Format:

Cheng Zheng, Jyun-Yu Jiang, Yichao Zhou, Sean D. Young, and Wei Wang. 2020. Social Media User Geolocation via Hybrid Attention. In *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401329

1 INTRODUCTION

Nowadays, social media has become one of the most powerful tools for myriad real-world applications, such as online marketing [9] and event detection [17]. To facilitate those applications, the geolocations of social media users are usually required. For example, online marketing needs to decide the target audience based on their locations. A real-world event may be only related to the users within a certain geographical region. However, there are only a limited amount of social media posts annotated with posting geolocations because position sensors and services can be unavailable or prevented. In addition, most of the social media users also do not denote their locations in the user profiles due to the data privacy issue. Hence, it is important to identify user geolocations with only their behaviors on social media.

One of the most intuitive approaches for user geolocation is to analyze the natural languages utilized in social media posts. Users can mention specific entities or events related to geolocations and people living in a certain region may reveal noticeable habits or patterns in their languages. For example, Rahimi et al. [12] extract bag-of-words features from user posts; Wing and Baldridge [18] estimate the word distributions for different regions; Han et al. [6] conduct feature selection to discover location indicative words. However, user language usage sometimes can be too vague and ambiguous to recognize their locations, especially in social media posts with only limited and noisy texts. Several less active users that rarely publish posts may also have insufficient data for geolocation.

In addition to social media posts published by users, social interactions with other users can also be applied to user geolocation. More precisely, a user can be more likely to reach out to the users living in closer areas. For instance, Davis Jr et al. [3] and Jurgens [8] exploit label propagation and rely on the location redundancy through user relationships. Wang et al. [16] derive node embeddings of social networks and location networks as features for user geolocation. Nevertheless, the sparsity of social networks can still lead to unsatisfactory performance. Social connections can be relationships with users living in other locations. Although previous studies utilize text frequency in social networks [13] and train machine learning models with heterogeneous features [4], conventional approaches are significantly affected by the network structures. Moreover, the importance of social networks can be distinct across different users.

In this paper, the framework, Hybrid-attentive User Geolocation (HUG), is proposed to tackle the above issues. Social media posts of each user are first encoded by a hierarchical language attention network. The social network of users is modeled by a graph attention network so that the relations between users can be leveraged in representation learning. Finally, the hybrid attention mechanism is applied to automatically decide the individual importance scores of user posts and the social network for each user, thereby identifying her geolocation. To improve the prediction performance of tail locations, we also propose a novel location regularizer that leverages the knowledge from other locations. In the end, we conduct extensive experiments to show the effectiveness of HUG with in-depth analysis. We also demonstrate the interpretability of HUG with several concrete examples.

*Equal contribution.

SIGIR ’20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401329

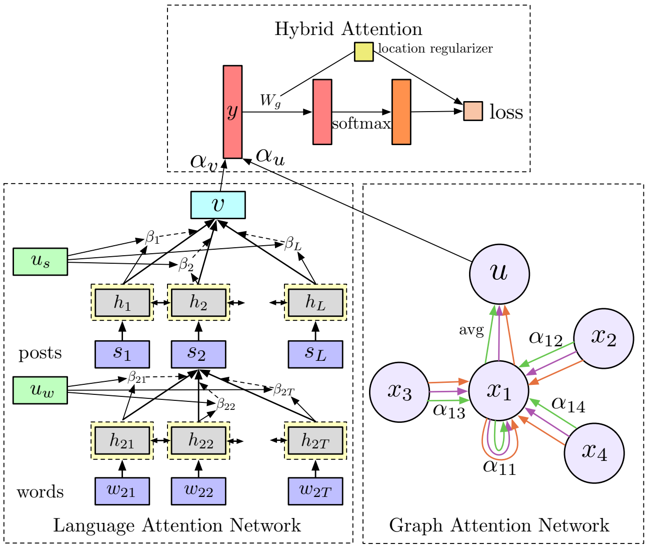


Figure 1: The overview of the proposed framework, HUG.

In the literature, social media user geolocation has attracted increasing attention in recent years. Some of the conventional methods focus on modeling social media posts [9, 12, 18], while several studies rely on social network information [1, 3, 8, 13, 16]. Although some approaches [4, 7, 13] simultaneously consider language and network knowledge, texts and social networks are individually and evenly modeled without considering distinct importance for different users. To the best of our knowledge, this work is the first study that dynamically adjusts the influence of different resources upon the model for social media user geolocation.

2 HYBRID-ATTENTIVE USER GEOLOCATION

In this section, we first formally define the task of social media user geolocation and then introduce our proposed approach, Hybrid-attentive User Geolocation (HUG).

Problem Statement. Suppose we have a set of social media users U and their social network G. For each user $k \in U$, $W^k = \{w^k_{ij}\}$ denotes the social media posts published by the user, where w^k_{ij} represents the j-th word of the i-th post in W^k. The social network $G = (U, A)$ treats each user $k \in U$ as a node and models their relations with an edge set $A \subseteq U \times U$ that indicates the relations between users. Given the social media posts W^k of a user k and the social network G, the goal of this work is to predict the geolocation of the user $L_k \in \mathcal{L}$, where $\mathcal{L}$ is a set of candidate locations.

2.1 Multi-head Graph Attention Network

To learn the structural knowledge from social networks, we employ the graph attention network [15] to derive node representations as user graph vectors.

Node Features. For each user k, the input node features x^{in}_k are the node attributes such as user profiles or bag-of-words features. We use a fully-connected layer to learn the hidden node features as $x^0_k = \mathcal{F}(W^0 x^{in}_k)$, where W^0 is the weight matrix and $\mathcal{F}(\cdot)$ is a nonlinear activation function.

Multi-head Graph Attention Layer. The graph attention network consists of several stacked graph attention layers passing node features x^i_k on different levels. For the i-th layer, the importance score s^i_{jk} of each edge $a_{jk} \in A$ between the users j and k can be estimated by the self-attention mechanism [20, 21] as $s^i_{jk} = \langle Q^i x^{i-1}_j, Q^i x^{i-1}_k \rangle$, where Q^i is the weight matrix applied to every node. The node features in the i-th layer x^i_k can then be obtained as:

$$x^i_k = \sigma\left(\sum_{j \in \mathcal{N}(k)} \alpha^i_{jk} \mathcal{W}^i x^{i-1}_j\right), \alpha^i_{jk} = \frac{\exp(s^i_{jk})}{\sum_{j' \in \mathcal{N}(k)} \exp(s^i_{j'k})}$$

where $\mathcal{N}(k)$ indicates the neighbors of the user k in the social network; $\mathcal{W}^i$ is a weight matrix for feature projection; $\sigma(\cdot)$ is a nonlinear activation function. Specifically, we utilize the multi-head attention mechanism to have a greater capability of modeling structural knowledge by concatenating the features generated by different weight matrices Q^i_z and $\mathcal{W}^i_z$, where $1 \le z \le H$; H is the number of heads. Finally, in the last layer, we average the multi-head features and delay the employment of the nonlinear activation to derive the user graph vector u_k as:

$$u_k = \sigma\left(\frac{1}{H}\sum_{z=1}^{H} \sum_{j \in \mathcal{N}(k)} \alpha^N_{jk,z} \mathcal{W}^i_z x^{N-1}_j\right),$$

where N is the number of graph attention layers.

2.2 Language Attention Network

We use a hierarchical language attention network [19] to encode the textual features for each user. The language attention model is composed of several parts, including a word embedding layer, a post encoder, and a user encoder.

Word Embedding Layer. We convert each word w_{ij} into a one-hot encoding representation $\tilde{w}_{ij}$ and embed the words to vectors e with an embedding matrix E, where $e_{ij} = E \cdot \tilde{w}_{ij}$.

Post Encoder. For each post of a user, we feed the word embeddings to a bidirectional Recurrent Neural Network (BiRNN) to learn a hidden state of each word with sequential information as:

$$\overleftarrow{h}_{ij} = \mathrm{GRU}(\overleftarrow{h}_{i,j+1}, e_{ij}), \overrightarrow{h}_{ij} = \mathrm{GRU}(\overrightarrow{h}_{i,j-1}, e_{ij}), h_{ij} = [\overleftarrow{h}_{ij}, \overrightarrow{h}_{ij}],$$

where $\mathrm{GRU}(\cdot)$ is the recurrent neural unit of the BiRNN. Here we choose GRU instead of LSTM because of its computational efficiency. To derive the post representation, we introduce an attention layer to obtain a weighted sum of the hidden states from the BiRNN layer. To be specific, we initialize a context vector u_w and calculate the attention scores β_{ij} for the words in the post as:

$$u_{ij} = \tanh(W_w \cdot h_{ij} + b_w), \beta_{ij} = \frac{exp(u^T_{ij} \cdot u_w)}{\sum_j exp(u^T_{ij} \cdot u_w)}, s_i = \sum_j \beta_{ij} \cdot h_{ij},$$

where W_w and b_w are the weight matrix and the bias to map each word into a hidden space for estimating importance.

User Encoder. Similarly, we employ a BiRNN model using GRU units to derive the hidden representations h_i for the posts of each user according to their published times as:

$$\overleftarrow{h}_i = \mathrm{GRU}(\overleftarrow{h}_{i+1}, s_i), \overrightarrow{h}_i = \mathrm{GRU}(\overrightarrow{h}_{i-1}, s_i), h_i = [\overleftarrow{h}_i, \overrightarrow{h}_i].$$

The other context vector u_s is then learned to estimate the importance score β_i for each post and aggregate the post representations h_i to form a user language vector v as follows:

$$u_i = \tanh(W_u \cdot h_i + b_u), \beta_i = \frac{exp(u_i^T \cdot u_s)}{\sum_j exp(u_i^T \cdot u_s)}, v = \sum_i \beta_i \cdot h_i,$$

where W_u and b_u are the weight matrix and the bias.

2.3 Hybrid Attention

To dynamically adjust the importance of two resources for a certain user, we propose the hybrid attention to jointly model texts and social networks. Precisely, a context vector c_h is applied to estimate the importance scores of graph and language user vectors as:

$$\alpha_v = \frac{\exp(o_v \cdot c_h)}{\exp(o_v \cdot c_h) + \exp(o_u \cdot c_h)}, \alpha_u = \frac{\exp(o_u \cdot c_h)}{\exp(o_v \cdot c_h) + \exp(o_u \cdot c_h)},$$

where $o_v = \tanh(W_h \cdot v + b_h)$; $o_u = \tanh(W_h \cdot u + b_h)$; W_h and b_h are the weight matrix and the bias for a nonlinear projection. Therefore, the ultimate feature vector produced by the hybrid attention can be obtained as $y = \alpha_v \cdot v + \alpha_u \cdot u$.

2.4 Location-regularized User Geolocation

User Geolocation. Based on the feature vector y, we use a fully-connected hidden layer without a bias to estimate the probability of being the geolocation of the user k for each location i as:

$$P(L_k = i) = \text{Softmax}(W_g y),$$

where W_g is the weight matrix for the hidden layer.

Location-based Regularization. To leverage the knowledge across different locations, we regularize the model weights W_g for inferring location probabilities by the corresponding distances. Specifically, we have the location-based regularization loss Loss_R as:

$$\text{Loss}_R = \sum_{j \in \mathcal{L}} \sum_{k \in \mathcal{L} - j} \frac{|W_g(j) - W_g(k)|^2}{D(j,k)},$$

where $D(j,k)$ denotes the distance between the locations j and k.

Learning and Optimization. Finally, the loss function of *HUG* for optimization can be derived by the cross-entropy loss for classification and the location-based regularization loss as:

$$\text{Loss} = \sum_i \mathbb{1}[L_k = i] \cdot P(L_k = i) + \gamma \cdot \text{Loss}_R,$$

where γ is the weight of regularization loss.

3 EXPERIMENTS

3.1 Experimental Setup

Datasets. We employ three public Twitter user geolocation datasets: (1) GeoText [5], (2) Twitter-US [14] and (3) Twitter-World [6]. The datasets are pre-partitioned into training, development and test sets. In each dataset, user tweets are concatenated into single documents. The social graphs are extracted with the mention relations between users, where two users are connected if one mentions the other, or they co-mention a third user. The node attributes in the social graphs are the bag-of-words and TFIDF features. The labels are the discretized geographical coordinates of the training users using a k-d tree [14]. Dataset statistics are summarized in Table 1.

Table 1: Dataset statistics.

	GeoText	Twitter-US	Twitter-World
Users	9,475	449,200	1,386,766
Classes	129	256	930
Train	5,685	429,200	1,366,766
Dev	1,895	10,000	10,000
Test	1,895	10,000	10,000

Baselines. We compare the proposed HUG against 6 baselines that are trained based on the text and network features to determine user geolocations, including: (1) MLP + k-d tree [12], a text-based multilayer perceptron model; (2) GCN-LP [13], a network-based model using one-hot neighbor encoding as the node attributes; (3) MENET [4], a multiview neural network model that utilize multi-entry data to infer users' locations; (4) MLP-TXT+NET [13], a multilayer perceptron model with the concatenation of text features and adjacent lists as input; (5) GCN [13], a graph convolution network model [10] with the bag-of-words as node attributes; (6) HLPNN [7], a feature fusion model with city and country objectives.

Evaluation. We evaluate the models with three commonly used metrics: (1) **Acc@161**, the accuracy of predicting a user within 161km or 100 miles from the labeled location; (2) **Mean**, the mean error between the predicted and labeled location; (3) **Median**, the median error between the predicted and labeled location.

Implementation Details. We implement the proposed HUG in PyTorch framework for efficient GPU computation. The language attention network has bidirectional GRUs with hidden dimensions in $\{50, 100, 200\}$ and the word embeddings are initialized with the Glove vectors [11] pre-trained on the Twitter corpus. The entity-level aggregation network has two layers with the hidden dimension $D_e \in \{64, 128, 256\}$. The number of heads in multi-head graph attention is searched in $\{1, 2, 4, 8, 16\}$. We apply Adam optimizer for training and the initial learning rate is set as 5.0×10^{-4}. The activation functions are ELU [2].

3.2 Experimental Results

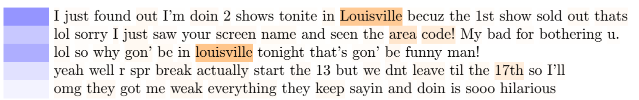
Table 2 summarizes the model performance of Twitter user geolocation prediction on all datasets. Overall, HUG is able to outperform other baselines across the three datasets on all metrics. We make the following observations. (1) Compared with MLP + k-d tree [12] and GCN-LP [13] that only utilizes a single source of data, e.g. text or network features, HUG outperforms by simultaneously learning the important language features and network structure features. (2) As the feature fusion models, MENET [4], MLP-TXT+NET [13] and HLPNN [7] incorporate the fixed network embeddings as features. In contrast, our proposed HUG can adaptively fine-tune both attention models to favor the geolocation prediction by the hybrid attention mechanism. (3) Compared with GCN [13], our attention-based approach can better understand the hierarchical language features and assign different importance to nodes of the same neighborhood. (4) We conduct the ablation study by removing the graph attention, language attention and location-based regularization one by one at a time. As the results on the Twitter-US dataset shown

Table 2: Twitter user geolocation prediction performance.

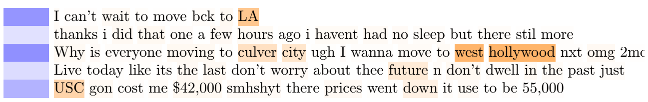
	GeoText			Twitter-US			Twitter-World		
	Acc@161 ↑	Mean ↓	Median ↓	Acc@161 ↑	Mean ↓	Median ↓	Acc@161 ↑	Mean ↓	Median ↓
MLP + k-d tree	38%	844	389	54%	554	120	34%	1456	415
GCN-LP	58%	576	56	53%	653	126	45%	2357	279
MENET	62%	532	32	66%	433	45	53%	1044	118
MLP-TXT+NET	58%	554	58	66%	420	56	58%	1030	53
GCN	60%	546	45	62%	485	71	54%	1130	108
HLPNN	-	-	-	71%	362	32	59%	828	60
HUG	**64%**	**516**	**30**	**72%**	**359**	**31**	**62%**	**818**	**49**

Table 3: Ablation study on Twitter-US.

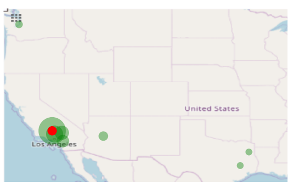
	Acc@161 ↑	Mean ↓	Median ↓
HUG	**72%**	**359**	**31**
w/o graph attention	51%	531	57
w/o language attention	58%	612	63
w/o location regularization	59%	562	51



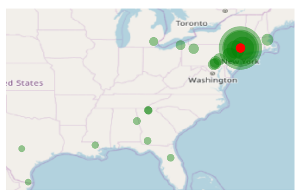
(a) User A



(b) User B



(c) User C



(d) User D

Figure 2: Attention weight analysis. (a)-(b) Documents of users A and B. (c)-(d) Geolocations and attention weights of users C and D with their one-hop neighbors.

in Table 3, each module contributes to the performance improvement and the proposed HUG benefits from the combination and the hybrid attention mechanism.

We further investigate the interpretability of the proposed HUG. Figure 2 shows the text and graph examples from the GeoText dataset. In (a) and (b), we show the social media posts of two users (A, B), whose hybrid attention weights for texts are $\alpha_v = 0.643$ and 0.794, respectively. The blue blocks denote the tweet-level attention weights. The orange denotes the word attention weights and our model can select the words with a strong indication of geolocations like *Louisville*, *LA*, *USC* and *West Hollywood*. Figure 2 (c) and (d) demonstrate two users (C, D) with the geolocations and attention weights of their one-hop neighbors. The hybrid attention weights for graph vector are $\alpha_u = 0.844$ and 0.942 for user C and D, respectively. We plot the geolocations of user C and D in red dots. The green dots are the geolocations of the one-hop neighbors and the dot sizes denote the attention weights. Our proposed HUG also works in terms of the graph attention and location-based regularization, by assigning the higher weights to closer neighbors and lower weights to farther neighbors.

4 CONCLUSION

In this paper, we propose a novel end-to-end framework, Hybrid-attentive User Geolocation (HUG), to jointly model the post texts and user interactions in social media and predict user geolocations. We introduce the hybrid attention mechanism to automatically determine the importance of texts and social networks while social media posts and interactions are modeled by a graph attention network and a language attention network. The experimental study on three benchmark geolocation datasets from Twitter shows that HUG consistently renders superior prediction performance against baseline approaches. We also demonstrate the interpretability of HUG with in-depth analysis of attention weights.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their careful reading and insightful comments on our manuscript. The work was partially supported by NIAID (R56AI125105, 7R01AI132030), NIMH (5R01MH106415), and NSF (DGE-1829071, IIS-2031187).

REFERENCES

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*.
[2] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*.
[3] Clodoveu A Davis Jr, Gisele L Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L. Arcanjo. 2011. Inferring the location of twitter messages based on user relationships. *Transactions in GIS* (2011).
[4] 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* (2017).
[5] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *EMNLP*.
[6] Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *COLING*.
[7] Binxuan Huang and Kathleen M Carley. 2019. A Hierarchical Location Prediction Neural Network for Twitter User Geolocation. In *EMNLP*.
[8] David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *ICWSM*.
[9] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. " I'm eating a sandwich in Glasgow" modeling locations with tweets. In *SMUC*.
[10] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.
[11] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.
[12] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. A Neural Model for User Geolocation and Lexical Dialectology. In *ACL*.
[13] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised User Geolocation via Graph Convolutional Networks. In *ACL*.
[14] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *EMNLP*.
[15] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.
[16] Fengjiao Wang, Chun-Ta Lu, Yongzhi Qu, and S Yu Philip. 2017. Collective geographical embedding for geolocating social network users. In *PAKDD*.
[17] Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In *ICWSM*.
[18] Benjamin P Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *ACL*.
[19] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.
[20] Cheng Zheng, Qin Zhang, Guodong Long, Chengqi Zhang, Sean D Young, and Wei Wang. 2020. Measuring Time-Sensitive and Topic-Specific Influence in Social Networks with LSTM and Self-Attention. *IEEE Access* (2020).
[21] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434* (2018).