A Transformer-based Framework for POI-level Social Post Geolocation

Menglin Li[1][0000−0002−7890−7636], Kwan Hui Lim[1][0000−0002−4569−0901], Teng Guo[2][0000−0001−6604−475X], and Junhua Liu[1,3][0000−0003−4477−7439]

[1] Singapore University of Technology and Design, Singapore
[2] Dalian University of Technology, China
[3] Forth AI, Singapore
menglin_li@mymail.sutd.edu.sg, kwanhui_lim@sutd.edu.sg,
teng.guo@outlook.com, j@forth.ai

Abstract. POI-level geo-information of social posts is critical to many location-based applications and services. However, the multi-modality, complexity, and diverse nature of social media data and their platforms limit the performance of inferring such fine-grained locations and their subsequent applications. To address this issue, we present a transformer-based general framework, which builds upon pre-trained language models and considers non-textual data, for social post geolocation at the POI level. To this end, inputs are categorized to handle different social data, and an optimal combination strategy is provided for feature representations. Moreover, a uniform representation of hierarchy is proposed to learn temporal information, and a concatenated version of encodings is employed to capture feature-wise positions better. Experimental results on various social media datasets demonstrate that the three variants of our proposed framework outperform multiple state-of-art baselines by a large margin in terms of accuracy and distance error metrics.

Keywords: Location Prediction · Geolocation · Social Media · Twitter · Transformer

1 Introduction

Knowing the posting location of social media data is important for many useful applications, including local event/place recommendations [8,24], location-based advertisements [6, 11], emergency location identification and disaster response [23,44]. However, geotagged social posts are very limited as less than 1% of tweets are labeled with geo-coordinates [1]. This constraint motivates our research on geolocation, which is a topic that has received significant attention in the past decade. However, most prior studies concentrate on user geolocation, which is estimating the home location of users [34,39,45,46]. This type of geo-information is insufficient for applications like emergency location identification and natural disaster response [21], which require the location of individual posts. Hence, in this paper, we focus on the problem of social post geolocation to infer the locations of individual posts.

For social post geolocation, previous efforts typically aim at inferring locations at the city level [2, 21, 43]. Although there is good performance at the city level, location information at such a coarse-grained level is still insufficient for the various applications mentioned earlier. While some researchers studied the task of geo-coordinates estimation, it is challenging to achieve high accuracy [28, 31]. In real-life scenarios, semantic toponyms are more practical and understandable compared to numerical latitude and longitude [42]. Therefore, we study the problem of social post geolocation at the Point-Of-Interest (POI) level, a fine-grained semantic level.

However, Social Post Geolocation at the POI level is a challenging problem due to the complexity, multi-modality, and diverse nature of social media data and their platforms. Firstly, the user-generated textual content is short, free-form, and often noisy, containing acronyms, misspellings, and special tokens. It is non-trivial to understand such complex text precisely for location estimation. Secondly, there are other non-textual contents such as time, social networks, images, and videos, which can be used for this task but also lead to the multi-modality issue. The ability to represent and fuse different data types is vital for geolocation. Lastly, it is increasingly important to develop a geolocation framework with a generalization ability to deal with the emergence of diverse social platforms, like photo-sharing and micro-blogging platforms. Many works focus on a single social platform with specific inputs, thus limiting their performance on other social platforms due to the difference in data fields. For better generalizability across platforms, some approaches utilize text content solely for geolocation but at the expense of missing out on other non-textual content and limiting performance.

To address these limitations, we present a transformer-based model, named transTagger, for POI-level social post geolocation, which is a general framework that builds upon the Bidirectional Encoder Representations from Transformers (BERT) model with good generalization ability across different social platforms for accurate fine-grained location inference. The main contributions of this work can be summarized as follows:

- We design a general categorization to tackle the multi-modality and diverse nature of social media data and their platforms and provide four datasets with ground truth covering two cities and two platforms.
- We fuse features and learn their correlations using transformer encoders with a concatenated version of positional encodings, along with a novel temporal representation to provide an optimal combination strategy of representations for multi-modality fusion. We denote this model, transTagger.
- We construct two additional variants, hierTagger and mtlTagger, by incorporating the hierarchy of locations into transTagger, and experimental results demonstrate that our models outperform state-of-the-art baselines by a considerate margin in terms of accuracy and distance error metrics.[1]

[1] Our code and dataset are made publicly available at https://github.com/lazylml/transTagger.

The rest of the paper is organized as follows. In Section 2, we review the critical related work in the geolocation field and briefly introduce hierarchical classification techniques. In Section 3, we first present the problem formulation and then describe our proposed model transTagger and two variants in detail. Then Section 4 introduces the experimental setting, while Section 5 presents and discusses our experimental results. Following that, we summarize and conclude this paper in Section 6.

2 Related Work

In this section, we review two main categories of work that are related to our research, namely social post geolocation and hierarchical geolocation works.

2.1 Post Geolocation

Post geolocation focuses on estimating the originating locations of social posts. Unlike user geolocation, which leverages a user's entire posting history, post geolocation considers only an individual post or tweet and uses that as input. For example, the work [13] uses the convolutional mixture density network for location estimation with single tweet content. Term co-occurrences in tweets, which exhibit spatial clustering or dispersion tendency, are detected and used to extend feature space in probabilistic language models [32]. For location prediction during disaster events, Ouaret et al. [31] present an iterative Random Forest fitting-prediction framework to learn semi-supervised models. A name entity recognizer [28] is developed for geolocating tweets with the help of GeoNames gazetteer. Kulkarni et al. [19] present a multi-level geocoding model that learns to associate texts with geographical locations and represent locations using S2 hierarchy. Others propose to locate tweets based on BERT architecture with different tokenization settings, like vocabulary sizes [36]. In special cases, historical locations of users are involved to boost location inference performance, like using the Markov model to formalize tweet geolocation in a flood-related disaster based on history tweets [38].

Many researchers consider metadata to infer tweet locations [2, 17, 20]. Pliakos and Kotropoulos construct a hypergraph based on images, users, geotags and tags of Flickr, which is further used for simultaneous image tagging and geolocation prediction [33]. A refined language model that is learned from massive corpora of social content, including tags, titles, descriptions, user ids, and image ids, is proposed to estimate the location of a post [16]. Miura et al. [29] propose a simple neural network structure with fully-connected layers and an average pooling process based on message text and user metadata for geolocation prediction. To classify the microblogs of WeiBo into 8 semantic categories, the work [42] explores the effect of user attributes and designs a neural network-based architecture with 4 feature fusion strategies.

2.2 Hierarchical Geolocation

Although the class hierarchy has been shown to be effective in closely relevant fields, like text classification [12,18,27,48], this problem has not thus far received the attention it deserves. Only a handful of existing works estimate the locations of tweets and explore geolocation performance using hierarchical locations. Previous efforts [26, 43] represent locations as a tree and construct a local classifier for each parent node to infer locations, which corresponds to a typical hierarchical classification technique, Local Classifier per Parent Node (LCPN) [37]. Multi-Task Learning (MTL) is incorporated to combine losses across multiple levels and predict locations at each level simultaneously [9, 19]. Most of these works aim at user location inference, whereas we study post geolocation.

Similar to our work, some research has attempted to infer fine-grained locations of tweets [3,4,30]. By investigating two properties, spatial focus and spatial homophily, a learning-to-rank framework [3, 4] is designed by ranking candidate venues. The work [30] extracts semantic similarities between tweets and POI reviews locally and globally to provide a Spatially-aware Geotext Matching model building upon MLP. Both methods need to compute similarity features explicitly with additional datasets, like check-in data or POI reviews from Foursquare, which is non-trivial and time-consuming to collect. While these works advance the task of tweet geolocation, our work differs from these earlier works in various ways, which we discuss next. Our method takes in tweet content and metadata of the Twitter dataset directly as inputs, building upon BERT and using transformer encoders to learn correlations among features. Additionally, we employ a uniform representation of decomposed hierarchical time elements to further boost performance as the importance of temporal features is highlighted by many studies [21, 25, 30, 38]. Moreover, we explore the effect of location hierarchy on the post geolocation performance by leveraging LCPN and MTL in our proposed models.

3 Method

3.1 Problem Formulation

The **Social Post Geolocation** problem is defined as estimating the originating location of tweets. In the same spirit as prior studies [21, 42, 43], the task is formulated as a classification problem where the predicted target is a location. Unlike these earlier works, which classify posts into countries or cities, we aim at inferring locations at a finer-granularity level, that is at the landmark or POI level. More specifically, the social post geolocation problem is represented as inferring POIs, given text and metadata of social media as input.

3.2 Method Overview

The overall structure of our proposed model, transTagger, is shown in Figure 1. To tackle the inconsistency of different social platforms, we classify the inputs

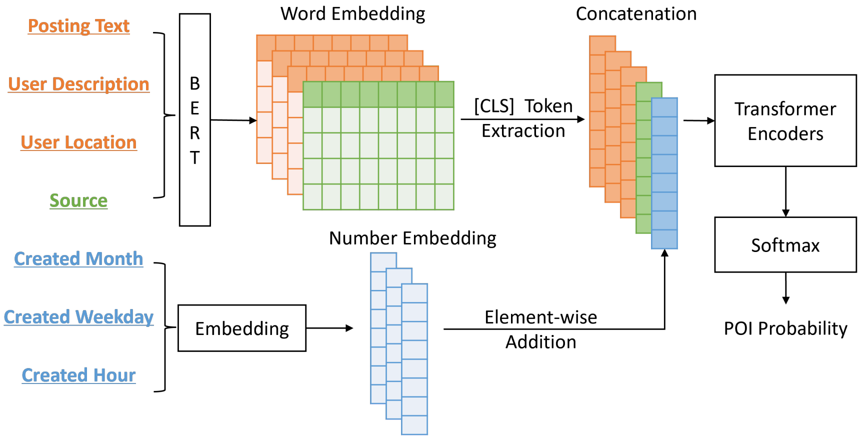

Fig. 1. The architecture of our proposed model transTagger

of social media data into three categories. Information contained in social media data can be divided into user-generated and system-generated according to sources. The user-generated content is of free form and could be very noisy. Besides posting text, the user-generated content also includes user locations, user descriptions and so on. They form the first category of inputs and we denote it **Text**. System-generated content comprises textual fields and numerical fields. The former is mostly categorical text, like source (indicating whether the tweet is posted from the phone or web platform), which falls into the second category of inputs: **Categorical Text (CT)**. For the latter, numerical fields, a typical one is the time (when the post is created), and others are less explored and employed in post geolocation and we leave them for future research. The third category is **Time** and we discuss the various representation techniques used in later sections. **Text**, **CT**, and **Time** are depicted in orange, green, and blue, respectively in Figure 1.

Our model applies BERT to learn semantic information and contextual information of **Text** and **CT** and maps features into a word embedding space. Following that, the representations of [CLS] tokens are extracted from all textual features and combined with embeddings of **Time**. Then we use several layers of transformer encoders to learn the correlation of all features. The POI probability of each post is calculated using a fully-connected layer with softmax as the activation function.

3.3 Feature Representation

We apply the pre-trained model by plugging in the post geolocation task-specific inputs and outputs into BERT. At the architecture level, BERT is an L-layer bi-directional transformer encoder [5]. The hidden size and the number of self-attention heads for each component are denoted as H and A, respectively.

Text, including posting texts, user locations, user descriptions, and **CT**, like sources, are all used as inputs. Here a degenerate text-∅ pair corresponds to sentence A and sentence B since we formulate the post geolocation task as a classification problem and there is no ”sentence” pair. An input sample is regarded as a sentence in this paper although it may actually contain multiple sentences. During tokenization, each sentence is converted into a sequence of tokens and a special classification token, [CLS], is injected in front of every input sample [5]. Then the first token becomes [CLS]. Apart from the above token embedding, other embeddings are utilized to take the position information inside sentences or between sentence pairs into consideration. Position embedding represents the position of each token in a sentence. In contrast, segment embedding is used to distinguish sentences A and B and thus is set to all zero in our case. The element-wise addition of token, position and segment embeddings forms the input representation [5].

We denote the learned embedding in the final hidden layer of each input sample as $E \in \mathbb{R}^{N\times H}$ where N is the sentence length. The corresponding embedding of the [CLS] token is represented as $C \in \mathbb{R}^{H}$. This token embedding can be seen as the aggregation of sentence representation, which is used for subsequent applications. Note that all the parameters are fine-tuned in an end-to-end manner based on our task, post geolocation.

Time is a vital factor in relation to human mobility and thus, of great importance for location inference. However, most works simply represent it as one-hot encoding based on the timestamp, which does not capture the full extent of temporal information and ignores the hierarchy of time elements, like hours and months. Inspired by this work [47], we propose a uniform representation of hierarchical time elements, UniHier, to learn temporal information. Hierarchical time elements are extracted from **Time**, including hours, weekdays, and months. Then each element is represented as a learnable embedding vector with dimension H and limited vocab size. A uniform representation of time is constructed by the element-wise addition of all embedding vectors.

3.4 Feature Fusion

Assuming that **Text** contains m fields, **CT** contains n fields, we extract [CLS] token vectors of **Text** and **CT**, and concatenate them with the UniHier representation of **Time**, then a feature matrix $F \in \mathbb{R}^{(m+n+1)\times H}$ is generated.

To learn the correlation of all features, we employ a multi-layer transformer encoder as described in the work [41]. Positional encodings are represented using sine and cosine functions of different frequencies as below and *pos* is the position, i is the dimension:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/H}) \tag{1}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/H}) \tag{2}$$

These positional encodings are fixed during training and with dimension H. In contrast to the now ubiquitous transformer encoder that sums feature representations and the corresponding positional encodings, we concatenate them and term it the concatenated version of positional encodings. Experiments demonstrate that this approach improves performance.

After concatenating with positional encodings, this feature matrix is utilized to calculate POI probabilities with a softmax layer. This model is then trained using the Adam update rule as the optimizer.

3.5 Hierarchical Prediction

The hierarchy of locations enables the application of hierarchical prediction and thus improves the performance of post geolocation. We incorporate LCPN, a typical hierarchical classification approach, with transTagger, and construct a variant, hierTagger. By combining the class hierarchy with MTL, we build upon our earlier described transTagger and propose another variant, mtlTagger. Due to space constraints, we briefly describe how to build these two variants and refer interested readers to our released source code for the implementation details.

hierTagger The LCPN approach aims to train a multi-class classifier for each parent node in the class hierarchy, to distinguish between its child nodes [37]. The class hierarchy is typically a tree or a Direct Acyclic Graph (DAG), which is represented as a tree in our case. We build the tree of toponyms at different scales, from coarse to fine, starting from a root node that covers the whole research area. For every parent node in this tree, we employ transTagger to construct a local classifier, which is trained independently. Then a top-down class prediction approach is applied during the testing phase.

mtlTagger MTL provides models with better generalization ability by sharing representations between related tasks [35]. The predictions of post location at coarser levels are designed as auxiliary tasks. We incorporate transTagger with hard parameter sharing, a commonly used approach with MTL in neural networks, to predict post location at different scales, from coarse to fine. The prediction result for the coarser level, denoted as q, is further utilized to constrain the finer level prediction by adding q to the loss function of the finer level. A correlation matrix between the two levels is employed to help the loss function of the finer level better understand the coarser level's prediction result.

4 Experimental Setting

4.1 Datasets

We perform our experiments using datasets from two different social media platforms, Flickr and Twitter, for two cities of Melbourne and Singapore.

Twitter We collected 266,614 geotagged tweets that were posted in Melbourne from 2010 to 2018, and 482,765 geotagged tweets that were posted in Singapore from 2018 to 2022. We also combined tweets from Melbourne and Singapore for experiments to test the robustness of our models. The Twitter datasets of Melbourne, and Singapore, and their combination are denoted as Twitter-Mel, Twitter-SG, and Twitter-SM, respectively.

Flickr The Flickr dataset comprises 78,131 geotagged images that were posted in Melbourne from 2004 to 2020, extracted using the Flickr API or from the Yahoo! Flickr Creative Commons 100M (YFCC-100M) [40]. We further augmented this dataset by collecting the metadata of Flickr users. This dataset is denoted as Flickr-Mel.[2]

A list of POIs and their categories are obtained using the Google Place API.[3] For Singapore, our research area is the whole country/city and there are 9,666 POIs. For Melbourne, we concentrate on the central city area and there are 242 POIs. To implement hierarchical prediction, POI themes and POI sub-themes are involved as labels to construct the class hierarchy. Specifically, there are 16 POI themes (eg., Leisure/Recreation), 49 POI sub-themes (e.g., Park/Garden), and 242 POIs (e.g., Batman Park).

Our work aims to predict the specific POI where a post is sent from, in contrast to existing efforts that focus on coarse-level predictions at the city, country, or even continent level. To this end, we label a tweet tw in the Twitter dataset (or image im in the case of the Flickr dataset) as one and only one POI. Following the proximity principle [22], we compare the distance between tw (or im) and the POI location using their latitude and longitude coordinates, and label it with the POI if their distance differs by less than 100 meters. Any tw (or im) that is not assigned a POI label is then filtered out. Note that the above statistics of the Twitter and Flickr datasets are computed after POI-labelling preprocessing.

Our two variants involve the use of class hierarchy of POIs. For example, hierTagger utilizes POI-theme level and POI-level labels, while mtlTagger contains three loss functions that are designed for POI theme, POI sub-theme, and POI predictions, respectively.

4.2 Evaluation Metrics

We use two evaluation metrics that are frequently used in geolocation tasks, namely accuracy and distance error. Accuracy, denoted as $\boldsymbol{acc@k}$, reflects the proportion of correct predictions based on the top-k results and we evaluate with k as 1, 5, 10, and 20. Mean distance error, represented as $\boldsymbol{mean}$, measures the mean distance between the predicted location and actual POI location. We also experimented using median distance error and observe that our models

[2] We also collected a Flickr dataset for Singapore but excluded it for further experimentation due to a low number of data points.

[3] https://developers.google.com/maps/documentation/places/web-service/overview

achieve 0 error, thus we do not report the results for concision. Unless otherwise specified, all results reported in this paper are at the POI level to make our models comparable.

4.3 Parameter Setting

In our experiments, the max sequence length for text and other textual features is 100. To represent **Time** inputs using UniHier, they are randomly initialized from a uniform distribution $U(-1.0, 1.0)$ with dimension 128 (this value corresponds to the dimension of word embeddings) and vocab size is limited to 60 since the finest granularity is a minute. These embeddings are then learned during training.

The hyperparameter tuning is conducted using Bayesian optimization on the learning rate, the number of encoder layers, the number of heads, hidden size, and batch size. The number of layers, the number of attention heads, and the hidden size of the transformer encoder before the softmax layer are set as 3, 48, and 1300, respectively. The training of our model is performed using Adam with an initial learning rate of 3e-4 and a batch size of 128. We train the model with 4 epochs. Additionally, the block threshold for hierTagger is set as 0.01 and the loss weights for mtlTagger are 0.1, 0.1, and 1.

4.4 Baselines

We compare our proposed model and two variants with various popular geolocation models, including **MNB-Ngrams** (Multinomial Naive Bayes with Uni/Bi/Tri-grams) [2, 4, 7, 26, 32], **CNN-TT** (Convolutional Neural Network with Text and Time) [22], and **HLPNN** (Hierarchical Location Prediction Neural Network) [9]. The CNN text classification model [15] is widely used for geolocation [10,13,25], which we include as a baseline **CNN** in our experiments, along with its variant that uses one-hot encoding **CNN-1Hot** [14]. Besides HLPNN, another hierarchical classification model, **HDLTex** (Hierarchical Deep Learning for Text Classification) [18] is utilized as one of baselines. Our two proposed variants, hierTagger and mtlTagger, are also involved in comparisons.

Table 1. Baseline comparison on Flickr-Mel and Twitter-Mel

	Flickr-Mel					Twitter-Mel				
	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(m)↓	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(m)↓
HLPNN	68.68	83.62	88.95	93.87	247.6	61.45	76.7	81.85	87.05	433.2
HDLTex	56.89	64.71	66.49	70.14	604	56.2	64.67	66.33	67.69	512.5
CNN-TT	75.49	87.63	90.83	94.14	241	67.85	80.69	84.93	89.19	351.9
CNN	59.4	74.19	81.16	88.43	528	60.45	77.27	83.45	88.54	408.5
CNN-1Hot	59.91	76.69	83.25	90.14	697.7	63.08	76.89	80.92	85.43	362
MNB-Ngrams	54.35	71.71	79.93	88.61	1071	49.82	73.6	79.05	84.62	500.7
transTagger	**77.88**	89.85	**93.05**	93.05	**175.8**	**71.96**	84.64	**88.2**	88.2	**303.3**
hierTagger	77.59	**90.13**	92.91	**95.87**	183.5	71.42	84.34	88.12	**91.49**	319.5
mtlTagger	77.22	89.44	92.86	95.73	182.9	71.84	**84.67**	88.03	91.44	317.9

Table 2. Baseline comparison on Twitter-SG and Twitter-SM

	Twitter-SG					Twitter-SM				
	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(km)↓	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(km)↓
CNN-TT	53.76	67.27	70.29	72.81	2.617	54.8	68.64	72.41	75.7	154.3
CNN	49.77	62.04	64.72	67.66	2.949	48.98	62.11	65.5	69.09	542.7
CNN-1Hot	38.34	50.11	52.96	55.86	3.536	38.85	52.05	55.54	59.21	882.9
transTagger	**61.94**	**73.75**	**76.75**	**76.75**	**2.215**	**64.88**	**76.8**	**80.08**	**80.08**	**3.69**

5 Experimental Results

5.1 Baseline Comparison

To verify the effectiveness of our proposed models, experiments are designed to compare the performance of three variants and various baselines on the Flickr-Mel and Twitter-Mel datasets, as shown in Table 1. Similar experiments are conducted on the Twitter-SG and Twitter-SM datasets as well as to further examine the robustness of geolocation performance, as presented in Table 2. We only report results for transTagger and three baselines as the hierarchical labels are not available for the latter two datasets.

Overall, transTagger, hierTagger, and mtlTagger outperform all baselines, including the hierarchical ones, across all four datasets. Compared with a strong baseline like CNN-TT, transTagger outperforms by a substantial margin, obtaining an improvement of 2.39%, 4.11%, 8.18% and 10.08% in accuracy (acc@1) on Flickr-Mel, Twitter-Mel, Twitter-SG, and Twitter-SM, respectively. The latter two datasets contain many more POIs and the improvement of transTagger over the baselines is even larger. This indicates that our model is versatile enough to handle a large number of classes (POIs) well. In addition to accuracy, the mean distance error is also greatly reduced. To be specific, transTagger reduces the mean distance error by 65.2, 48.6, 402, and 1174 meters, compared with CNN-TT. In contrast to Table 1, we use kilometers (km) to denote distance in Table 2 because Twitter-SG and Twitter-SM cover much larger areas and thus values of mean distance error are relatively higher. In addition, the distance calculation of Twitter-SM involves two cities and thus is quite sensitive to prediction accuracy as this dataset is a mixture of Twitter-SG and Twitter-Mel. Therefore, the distance errors would increase greatly in comparison to the corresponding accuracy that decreases slightly, as shown in Table 2 and Table 4.

The overall results show that our proposed models provide superior performance for POI-level post geolocation across all cities and platforms, compared to the various baselines.

5.2 Representation Combination Selection

Taking generalization into consideration, we categorize inputs into three types: **Text**, **CT**, and **Time**, and performed a representation for each type, as previously described in Section 3. However, there are multiple ways to represent each input type. For **CT**, one way is to treat categorical texts as normal texts

Table 3. Representation combination selection on Flickr-Mel and Twitter-Mel

	Flickr-Mel					Twitter-Mel				
	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(m)↓	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(m)↓
transTagger										
Text-Text	77.88	89.85	93.05	93.05	175.8	**71.96**	**84.64**	88.2	88.2	**303.3**
1Hot-Text	77.88	89.85	93.05	93.05	175.8	71.69	84.6	**88.29**	**88.29**	322.1
Text-UniHier	**78.04**	**90.16**	**93.28**	**93.28**	**171.6**	70.03	84.32	88.09	88.09	313.8
1Hot-UniHier	**78.04**	**90.16**	**93.28**	**93.28**	**171.6**	69.69	84.25	87.87	87.87	308.3
Text-1Hot	77.49	89.66	92.99	92.99	184.1	69.91	84.13	87.71	87.71	316.9
1Hot-1Hot	77.49	89.66	92.99	92.99	184.1	69.5	84.26	88.04	88.04	321.6
hierTagger										
Text-Text	77.59	**90.13**	92.91	**95.87**	183.5	71.42	84.34	88.12	91.49	319.5
1Hot-Text	77.59	**90.13**	92.91	**95.87**	183.5	**71.49**	**84.45**	**88.15**	**91.56**	324.5
Text-UniHier	**78.18**	89.94	**93.15**	95.83	**169.5**	70.03	84.29	88.03	91.52	314.4
1Hot-UniHier	**78.18**	89.94	**93.15**	95.83	**169.5**	69.6	84.18	87.78	91.07	**308.8**
Text-1Hot	77.23	89.43	92.82	95.56	190.2	69.82	84.04	87.57	91.16	316.7
1Hot-1Hot	77.23	89.43	92.82	95.56	190.2	69.35	84.19	87.96	91.46	321.9
mtlTagger										
Text-Text	77.22	89.44	92.86	95.73	182.9	**71.84**	**84.67**	88.03	91.44	317.9
1Hot-Text	77.22	89.44	92.86	95.73	182.9	71.48	84.45	**88.04**	**91.64**	315.1
Text-UniHier	**78.93**	**90.18**	93.31	95.97	**168.3**	69.91	84.3	87.94	91.52	**312.9**
1Hot-UniHier	**78.93**	**90.18**	93.31	95.97	**168.3**	69.16	84.06	88.01	91.39	314.2
Text-1Hot	77.84	89.9	**93.36**	**96.26**	179.1	69.62	84.18	87.83	91.35	317.3
1Hot-1Hot	77.84	89.9	**93.36**	**96.26**	179.1	69.39	84	87.72	91.33	314.9

and use BERT or other language models to generate representations, and we call this Text embedding. Another commonly used approach is one-hot encoding. For **Time**, one way is to treat date/time as a standard text and generate temporal embedding using language models. Hence, there are two ways to represent **CT**: text and one-hot, and three ways for **Time**: text, one-hot, and UniHier. This results in six combinations of these representation methods, which we further experiment to find an optimal representation combination strategy. The results are illustrated in Table 3 and Table 4, where Text denotes Text embedding, and Text-UniHier refers to using Text embedding for **CT** and UniHier representation for **Time**, and so forth. Note that the results of Text-Text and 1Hot-Text are duplicated for Flickr since there are no **CT** fields. Similarly for Text-UniHier and 1Hot-UniHier, Text-1Hot and 1Hot-1Hot.

The results show that Text-Text delivers the overall best performance across all Twitter datasets. However, Text-UniHier (or 1Hot-UniHier) outperforms others for the Flickr dataset. One possible reason is that Flickr contains more time

Table 4. Representation combination selection of transTagger on Twitter-SG and Twitter-SM

	Twitter-SG					Twitter-SM				
	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(km)↓	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(km)↓
Text-Text	**61.94**	**73.75**	**76.75**	**76.75**	**2.215**	**64.88**	76.8	**80.08**	**80.08**	**3.69**
1Hot-Text	61.37	73.26	76.36	76.36	2.292	64.84	**76.88**	80.06	80.06	3.263
Text-UniHier	58.1	72.71	75.92	75.92	2.318	61.9	76.1	79.55	79.55	56.63
1Hot-UniHier	57.82	72.63	75.94	75.94	2.332	61.53	76.13	79.48	79.48	69.64
Text-1Hot	58.13	72.71	75.92	75.92	2.334	61.74	76	79.33	79.33	67.52
1Hot-1Hot	57.3	72.43	75.65	75.65	2.349	61.21	75.8	79.24	79.24	58.33

fields, including photo taken time and photo posted time, compared to Twitter that only contains tweet created time. Therefore, the best representation combination is Text-Text. In the event where multiple time inputs are involved, it is recommended to represent temporal inputs using UniHier.

We further compare the performance of three variants. Contrary to our expectations, hierTagger and mtlTagger show no distinct advantage, except for acc@20. Hence, these two variants are recommended when this specific metric is important. The intuition of utilizing hierarchical locations is that the prediction results at coarser level can help guide the geolocation at target level. However, this process might involve error propagation and thus impair the expressive power of the whole architecture. An effective mechanism for correcting these prediction errors is a promising direction to boost geolocation performance, and we leave this for future work.

Table 5. Ablation study on Twitter-SG and Twitter-SM

	Twitter-SG					Twitter-SM				
	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(km)↓	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Mean(km)↓
transTagger	**61.94**	**73.75**	**76.75**	**76.75**	**2.215**	**64.88**	**76.8**	80.08	80.08	**3.69**
w/o transformer	60.3	72.44	75.64	75.64	2.338	63.92	76.7	**80.1**	**80.1**	6.108
w/o position	61.28	73.01	75.96	75.96	2.292	64.39	76.43	79.75	79.75	4.917

5.3 Ablation Study

We compare transTagger with two ablations to examine the effectiveness of two model components, namely transformer encoders and position encodings. Table 5 shows the performance breakdown on Twitter-SG and Twitter-SM. For w/o position, we replace the concatenation version of positional encodings with the commonly used add-on version. The w/o transformer ablation removes the transformer encoders which are used to learn the correlation of features. The results demonstrate that all components contribute to improving the post geolocation performance of transTagger. Among all components, encoders have the greatest effect as shown by how it increases accuracy (including acc@1, acc@5, acc@10, and acc@20) and reduces the mean distance error by the largest margin.

Table 6. Coarse-Level Geolocation

	Flickr-Mel(POI-Theme)				Flickr-Mel(POI)			
	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑	Acc@1↑	Acc@5↑	Acc@10↑	Acc@20↑
HLPNN	79.92	97.16	**100**	**100**	68.68	83.62	88.95	93.87
hierTagger	**83.22**	**97.93**	**100**	**100**	**77.59**	**90.13**	**92.91**	**95.87**
mtlTagger	81.57	97.49	99.97	**100**	77.22	89.44	92.86	95.73

5.4 Coarse-Level Geolocation

We now study the prediction results of coarse-level geolocation since our two hierarchical variants both incorporate the toponym hierarchy. Although mtlTagger is capable of inferring locations at three levels, only the results of POI theme and POI are listed in Table 6 to make mtlTagger consistent and comparable with hierTagger. We observed that our models not only outperform at the target level (POI) by a large margin but also present outstanding coarse-level (POI theme) performance, even when compared with the competitive hierarchical geolocation algorithm HLPNN [9]. Furthermore, hierTagger obtained an absolute improvement of almost 2 points compared to mtlTagger (acc@1) for POI-theme geolocation even though the two have a similar capability of estimating POI-level locations. To force the model to focus more on our target task, POI geolocation, we set the weights of mtlTagger as 0.1, 0.1, and 1, for the loss functions of POI theme, POI sub-theme, and POI, respectively. In turn, this might be the cause of a negative impact on coarse-level prediction.

6 Conclusion

In this paper, we propose a transformer-based general framework, transTagger, for POI-level post geolocation. The inputs are categorized into three types: **Text**, **CT**, and **Time** to handle different social data, and the optimal representation combination, Text-Text, is provided by experimenting with all combinations. A novel representation of time, UniHier, is presented and verified to be useful in the case of multiple temporal inputs. Transformer encoders are employed to enhance geolocation performance and a concatenated version of encodings is incorporated to capture feature-wise positions. The effectiveness and robustness of our model are demonstrated on four datasets, covering two cities and two social platforms. Two variants, hierTagger and mtlTagger, by incorporating respective LCPN and MTL with transTagger, are shown to lift acc@20 effectively.

While these results are encouraging, we believe our approach can be further improved via two future directions. Firstly, we can explore more representation methods for different inputs, like numeral embeddings to extract time entities accurately. Secondly, we can also incorporate other modalities in addition to text and numbers, such as images and videos to provide more comprehensive knowledge.

Acknowledgments. This research / project is supported by the National Research Foundation, Singapore, and Ministry of National Development, Singapore under its Cities of Tomorrow R&D Programme (CoT Award COT-V2-2020-1). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Ministry of National Development, Singapore. The computational work was partially performed on resources of the National Supercomputing Centre, Singapore.

References

1. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 759–768 (2010)
2. Chi, L., Lim, K.H., Alam, N., Butler, C.J.: Geolocation prediction in twitter using location indicative words and textual features. In: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). pp. 227–234 (2016)
3. Chong, W.H., Lim, E.P.: Exploiting contextual information for fine-grained tweet geolocation. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 11 (2017)
4. Chong, W.H., Lim, E.P.: Exploiting user and venue characteristics for fine-grained tweet geolocation. ACM Transactions on Information Systems (TOIS) **36**(3), 1–34 (2018)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
6. Evans, C., Moore, P., Thomas, A.: An intelligent mobile advertising system (imas): Location-based advertising to individuals and business. In: 2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems. pp. 959–964. IEEE (2012)
7. Han, B., Cook, P., Baldwin, T.: Text-based twitter user geolocation prediction. Journal of Artificial Intelligence Research **49**, 451–500 (2014)
8. Ho, N.L., Lim, K.H.: Poibert: A transformer-based model for the tour recommendation problem. In: Proceedings of the 2022 IEEE International Conference on Big Data (2022)
9. Huang, B., Carley, K.: A hierarchical location prediction neural network for Twitter user geolocation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4732–4742. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1480, https://aclanthology.org/D19-1480
10. Huang, B., Carley, K.M.: On predicting geolocation of tweets using convolutional neural networks. In: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. pp. 281–291. Springer (2017)
11. Huang, H., Gartner, G., Krisp, J.M., Raubal, M., Van de Weghe, N.: Location based services: ongoing evolution and research agenda. Journal of Location Based Services **12**(2), 63–93 (2018)
12. Huang, W., Chen, E., Liu, Q., Chen, Y., Huang, Z., Liu, Y., Zhao, Z., Zhang, D., Wang, S.: Hierarchical multi-label text classification: An attention-based recurrent network approach. In: Proceedings of the 28th ACM international conference on information and knowledge management. pp. 1051–1060 (2019)
13. Iso, H., Wakamiya, S., Aramaki, E.: Density estimation for geolocation via convolutional mixture density network. CoRR **abs/1705.02750** (2017), http://arxiv.org/abs/1705.02750

14. Johnson, R., Zhang, T.: Effective use of word order for text categorization with convolutional neural networks. In: NAACL (2015)
15. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1181, https://www.aclweb.org/anthology/D14-1181
16. Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, Y.: Geotagging social media content with a refined language modelling approach. In: Pacific-Asia Workshop on Intelligence and Security Informatics. pp. 21–40. Springer (2015)
17. Kordopatis-Zilos, G., Popescu, A., Papadopoulos, S., Kompatsiaris, Y.: Placing images with refined language models and similarity search with pca-reduced vgg features. In: MediaEval (2016)
18. Kowsari, K., Brown, D.E., Heidarysafa, M., Meimandi, K.J., Gerber, M.S., Barnes, L.E.: Hdltex: Hierarchical deep learning for text classification. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). pp. 364–371. IEEE (2017)
19. Kulkarni, S., Jain, S., Hosseini, M.J., Baldridge, J., Ie, E., Zhang, L.: Spatial language representation with multi-level geocoding. arXiv preprint arXiv:2008.09236 (2020)
20. Li, M., Lim, K.H.: Geotagging social media posts to landmarks using hierarchical bert (student abstract). In: Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI'22) (2022)
21. Li, P., Lu, H., Kanhabua, N., Zhao, S., Pan, G.: Location inference for non-geotagged tweets in user timelines. IEEE Transactions on Knowledge and Data Engineering **31**(6), 1150–1165 (2018)
22. Lim, K.H., Karunasekera, S., Harwood, A., George, Y.: Geotagging tweets to landmarks using convolutional neural networks with text and posting time. In: Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion. pp. 61–62 (2019)
23. Liu, J., Singhal, T., Blessing, L.T., Wood, K.L., Lim, K.H.: Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In: Proceedings of the 32nd ACM Conference on Hypertext and Social Media. pp. 133–141 (2021)
24. Liu, J., Wood, K.L., Lim, K.H.: Strategic and crowd-aware itinerary recommendation. In: Proceedings of the 2020 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'20) (2020)
25. Liu, R., Cong, G., Zheng, B., Zheng, K., Su, H.: Location prediction in social networks. In: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. pp. 151–165. Springer (2018)
26. Mahmud, J., Nichols, J., Drews, C.: Where is this tweet from? inferring home locations of twitter users. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 6 (2012)
27. Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised hierarchical text classification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 6826–6833 (2019)
28. Mircea, A.: Real-time classification, geolocation and interactive visualization of covid-19 information shared on social media to better understand global developments. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020 (2020)

29. Miura, Y., Taniguchi, M., Taniguchi, T., Ohkuma, T.: A simple scalable neural networks based model for geolocation prediction in twitter. In: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). pp. 235–239 (2016)
30. Mousset, P., Pitarch, Y., Tamine, L.: End-to-end neural matching for semantic location prediction of tweets. ACM Transactions on Information Systems (TOIS) **39**(1), 1–35 (2020)
31. Ouaret, R., Birregah, B., Soulier, E., Auclair, S., Boulahya, F.: Random forest location prediction from social networks during disaster events. In: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 535–540. IEEE (2019)
32. Ozdikis, O., Ramampiaro, H., Nørvåg, K.: Spatial statistics of term co-occurrences for location prediction of tweets. In: European Conference on Information Retrieval. pp. 494–506. Springer (2018)
33. Pliakos, K., Kotropoulos, C.: Simultaneous image tagging and geo-location prediction within hypergraph ranking framework. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6894–6898. IEEE (2014)
34. Qian, Y., Tang, J., Yang, Z., Huang, B., Wei, W., Carley, K.M.: A probabilistic framework for location inference from social media. arXiv preprint arXiv:1702.07281 (2017)
35. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
36. Scherrer, Y., Ljubešić, N.: Social media variety geolocation with geobert. In: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects. The Association for Computational Linguistics (2021)
37. Silla, C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery **22**(1), 31–72 (2011)
38. Singh, J.P., Dwivedi, Y.K., Rana, N.P., Kumar, A., Kapoor, K.K.: Event classification and location prediction from tweets during disasters. Annals of Operations Research **283**(1), 737–757 (2019)
39. Tao, H., Gao, Y., Wang, Z., Khan, L., Thuraisingham, B.: An episodic learning based geolocation detection framework for imbalanced data. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)
40. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
42. Wang, N., Liu, B., Niu, M., Meng, K., Li, H., Liu, B., Wang, Z.: Semantic place prediction with user attribute in social media. IEEE MultiMedia **28**(4), 29–37 (2021)
43. Wing, B., Baldridge, J.: Hierarchical discriminative classification for text-based geolocation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 336–348 (2014)
44. Zheng, X., Han, J., Sun, A.: A survey of location prediction on twitter. IEEE Transactions on Knowledge and Data Engineering **30**(9), 1652–1671 (2018)
45. Zhong, T., Wang, T., Zhou, F., Trajcevski, G., Zhang, K., Yang, Y.: Interpreting twitter user geolocation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 853–859 (2020)

46. Zhou, F., Qi, X., Zhang, K., Trajcevski, G., Zhong, T.: Metageo: A general framework for social user geolocation identification with few-shot learning. IEEE Transactions on Neural Networks and Learning Systems (2022)
47. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11106–11115 (2021)
48. Zhou, J., Ma, C., Long, D., Xu, G., Ding, N., Zhang, H., Xie, P., Liu, G.: Hierarchy-aware global model for hierarchical text classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1106–1117 (2020)