CS245 Project Presentation

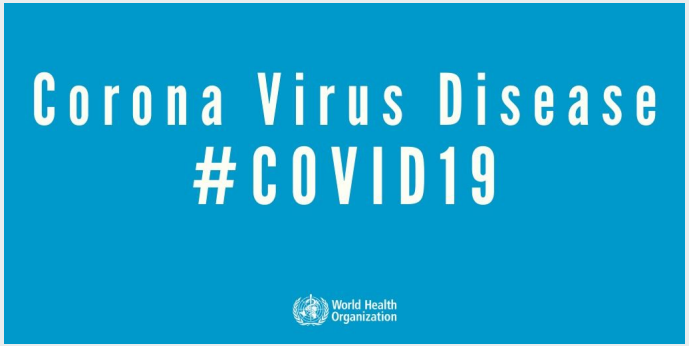
Geotag prediction of COVID-19 related tweets

Ziqi Li, Bowen Zhang, Yuhong Jiang,
Jiangtao Chen, Tianyu Zhang, Jiapeng Wan

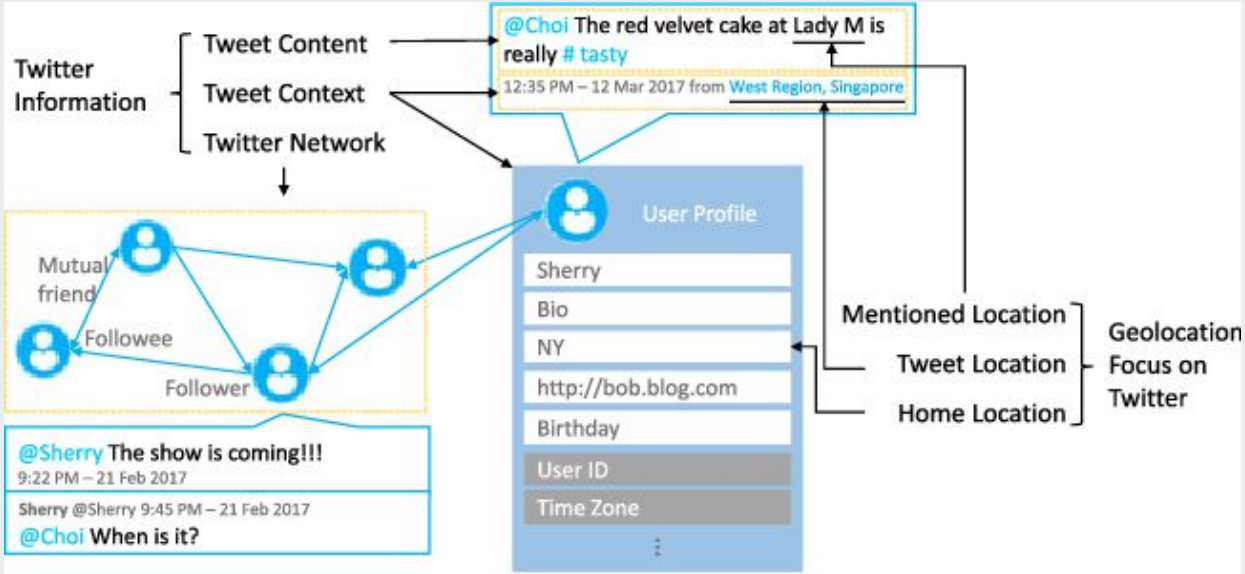
Outline:

- Introduction & Problem statement
- Data retrieval and pre-processing
- Model architecture & Problems of Deepgeo
- Word2vec Deepgeo
- Evaluation and comparison
- Discussion and future work

Intro & problem statement



Location prediction trends



Data Retrieval

- COVID-19 Twitter Streams is deprecated
- Twitter API - Filtered stream with COVID-19 related tags(“covid”, “corona”)

- 2 Accounts, 500,000 tweets per account
- Took more than 3 days to retrieve all the tweets

Fig. Twitter Developer Platform API Counter

Data Retrieval

- Approximately 800,000 tweets in total
- Only 1% with geotag
- 9,693 tweets with geo-location information
- Roughly covers 8 AM to 12 AM

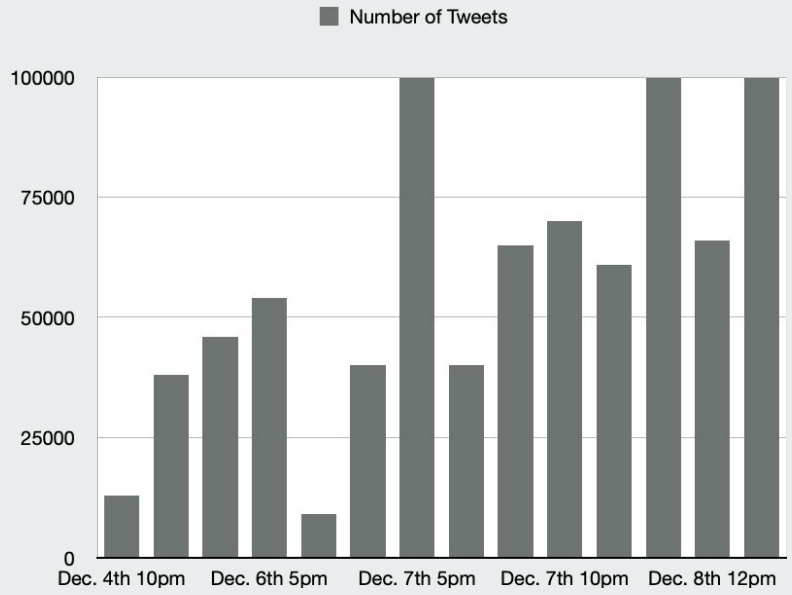


Fig. Data Retrieval Time Distribution

Data Pre-processing

Extract input fields

Preprocessed geo-location

Fig. Data in Raw JSON

Fig. Processed Data in JSON(Input X)

Data Pre-processing

Extract input fields
Preprocessed geo-location

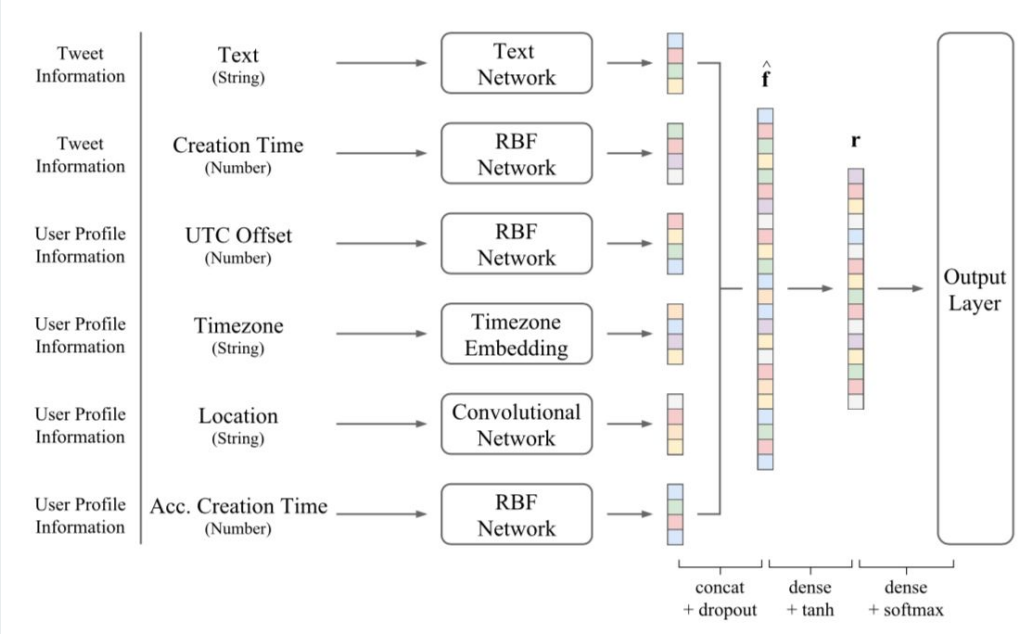
Fig. Data in Raw JSON

Fig. Processed Data in JSON(Label Y)

Model Architecture

- Features as input:
  - Tweet message
  - Tweet creation time
  - Location
- Network used:
  - Text Network
  - RBF Network
  - Convolutional Network

- Deepgeo



Model Architecture

RBF Network

(3 time features)

$$r_i = \exp\left(\frac{-(u-\mu_i)^2}{2\sigma_i^2}\right)$$

$$\mathbf{f}_{\text{rbf}} = [r_0, r_1, \ldots, r_{B-1}]$$

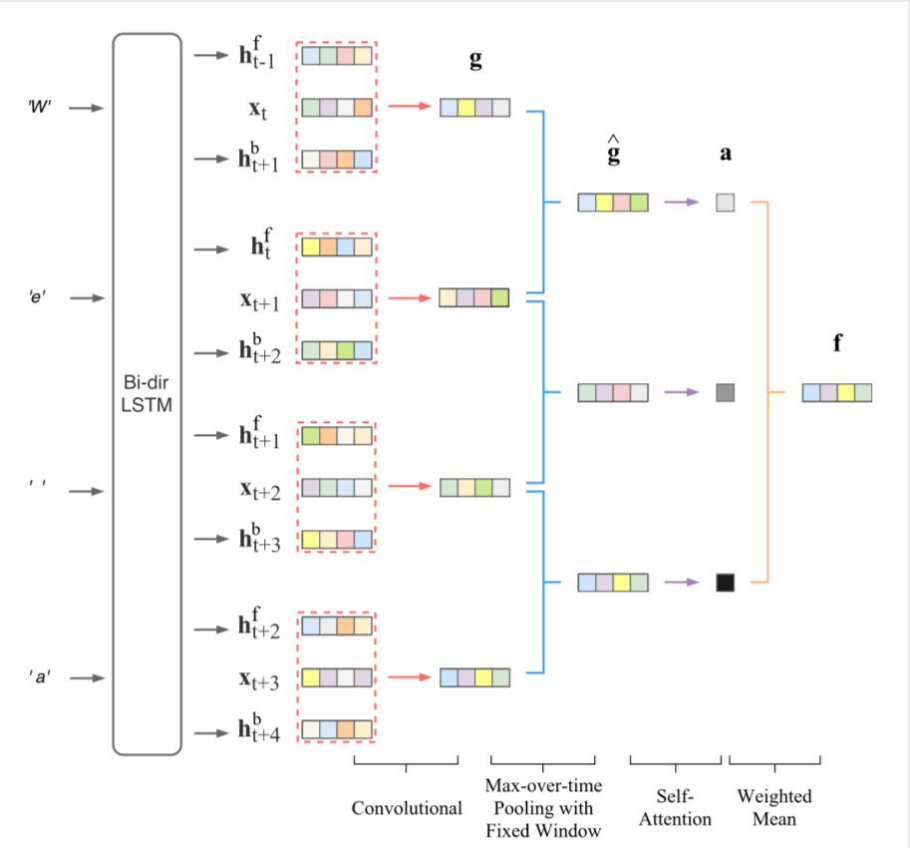
Convolutional Network

(location information)

$$\mathbf{x}_{0:T-1} = \mathbf{x}_0 \oplus \mathbf{x}_1 \oplus \ldots \oplus \mathbf{x}_{T-1}$$

$$\mathbf{g}_t = \text{ReLU}(\mathbf{W}_g \mathbf{x}_{t:t+Q-1})$$
$$\mathbf{f}_{\text{conv}} = \max(\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_{T-Q})$$

Text Network



$$\hat{\mathbf{x}}_t = \mathbf{h}^f_{t-1} \oplus \mathbf{x}_t \oplus \mathbf{h}^b_{t+1}$$
$$\mathbf{g}_t = \mathrm{ReLU}(\mathbf{W}_g \hat{\mathbf{x}}_t)$$

$$\hat{\mathbf{g}}_t = \max(\mathbf{g}_t, \mathbf{g}_{t+1}, \cdots, \mathbf{g}_{t+P-1})$$

$$\alpha_t = \mathbf{v}^\mathsf{T} \tanh(\mathbf{W}_v \hat{\mathbf{g}}_t)$$
$$\mathbf{a} = \mathrm{softmax}(\alpha_0, \alpha_1, \ldots, \alpha_{T-P})$$

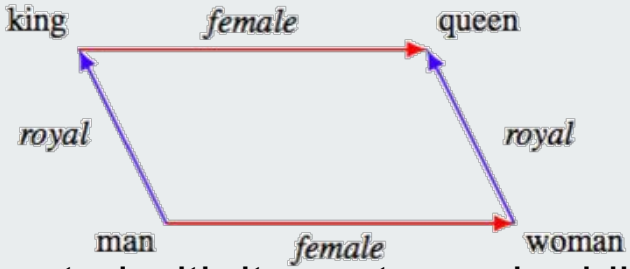
$$\mathbf{f}_{\text{text}} = \sum_{t=0}^{T-P} \mathbf{a}_t \hat{\mathbf{g}}_t$$

Problems of DeepGeo

- 1. Features unavailable
- *User timezone*, *user UTC offset* are no longer provided.
- 2. Features redundant
- Only *Text* and *Location* are important.
- 3. Randomly initialized text embedding
- 4. Character-level embedding may be too fine-grained

Word2Vec

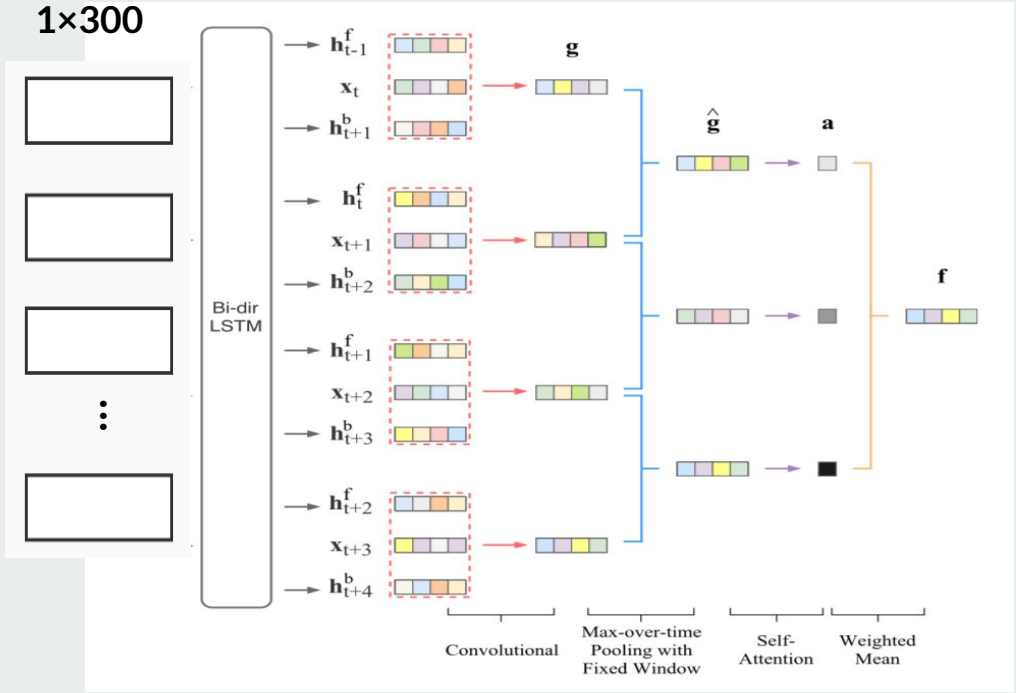
Word2Vec is a pre-trained (i.e., initially learned) model that provides embedding weights at the word level.



Each token (word) is sequentially represented with its vector embedding (Word2Vec) and concatenated with the forward and backward hidden states from a bi-directional LSTM network before applying max-overtime pooling, self-attention, and weighted mean

Model Architecture

- Word2Vec with original Model



Evaluation and comparison

Network	Hyper-Parameter	Message-Only	Tweet+User
Overall	Batch Size	512	
	Epoch No.	10	
	Dropout	0.2	
	Learning Rate	0.001	
	R	400	
Text	Max Length	300	300
	E	200	200
	P	10	10
	O	600	400
Time	B	–	50
UTC Offset	B	–	50
Timezone	Embedding Size	–	50
Location	Max Length	–	20
	E	–	300
	Q	–	3
	O	–	300
Account Time	B	–	10

Table : deepgeo hyper-parameters and values.

Accuracy	System	Features
0.146	Chi et al. (2016)	Message Only
0.212	deepgeo	Message Only
0.409	Miura et al. (2016)	Tweet + User Metadata
0.428	deepgeo	Tweet + User Metadata

Table : Geolocation prediction test accuracy.

Evaluation and comparison

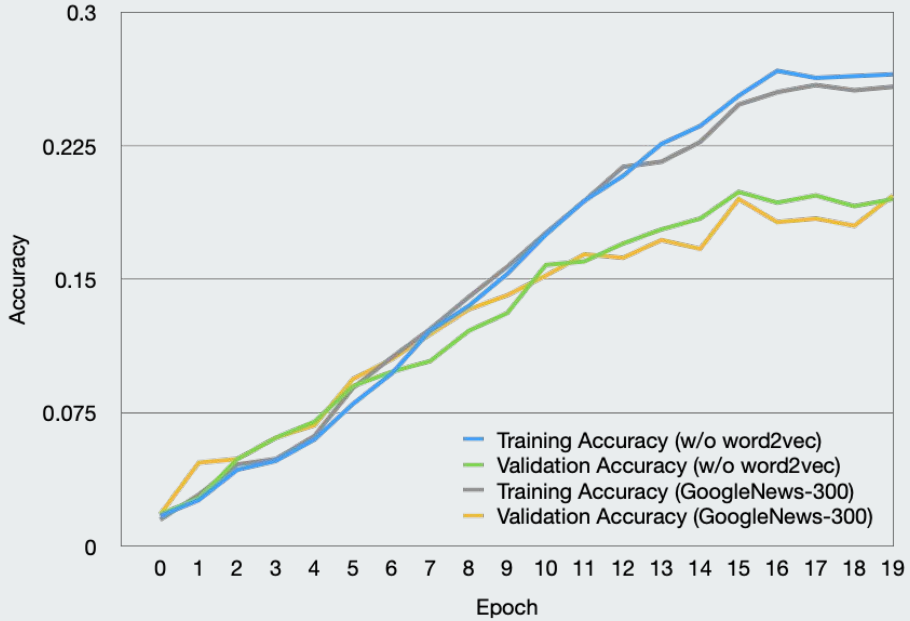


Figure. Training Accuracy and validation accuracy of model trained with original character embeddings and Word2Vec embeddings (GoogleNews-300)

Evaluation and comparison

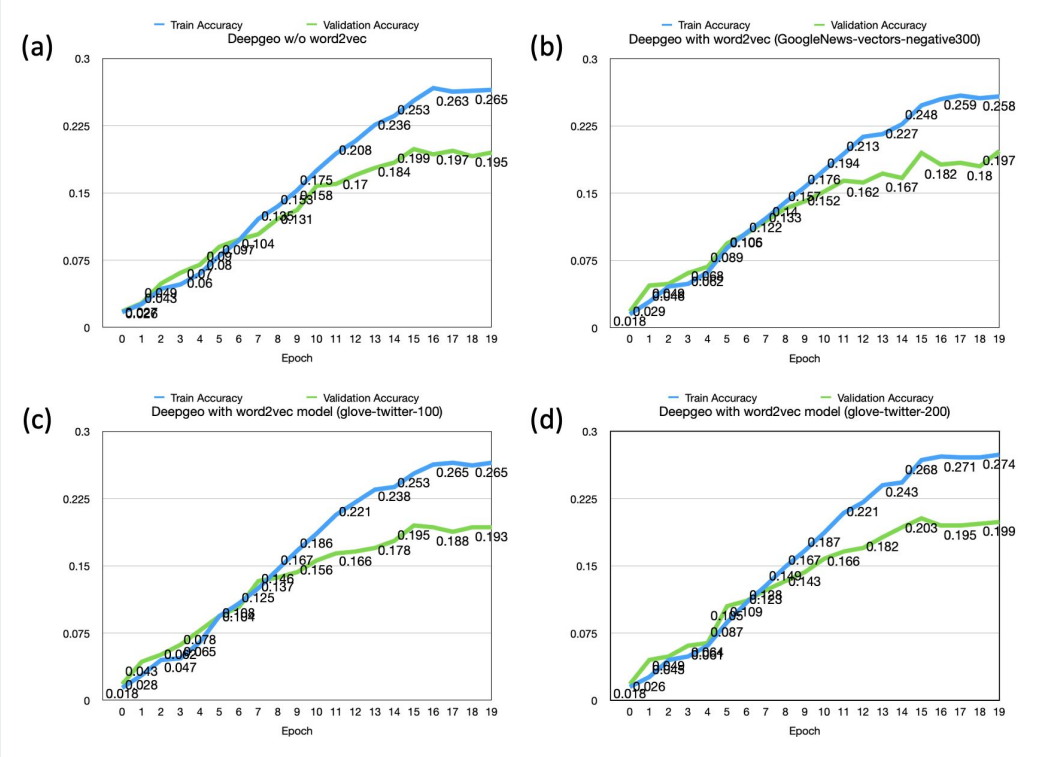


Figure. Training Accuracy and validation accuracy of different models

Evaluation and comparison

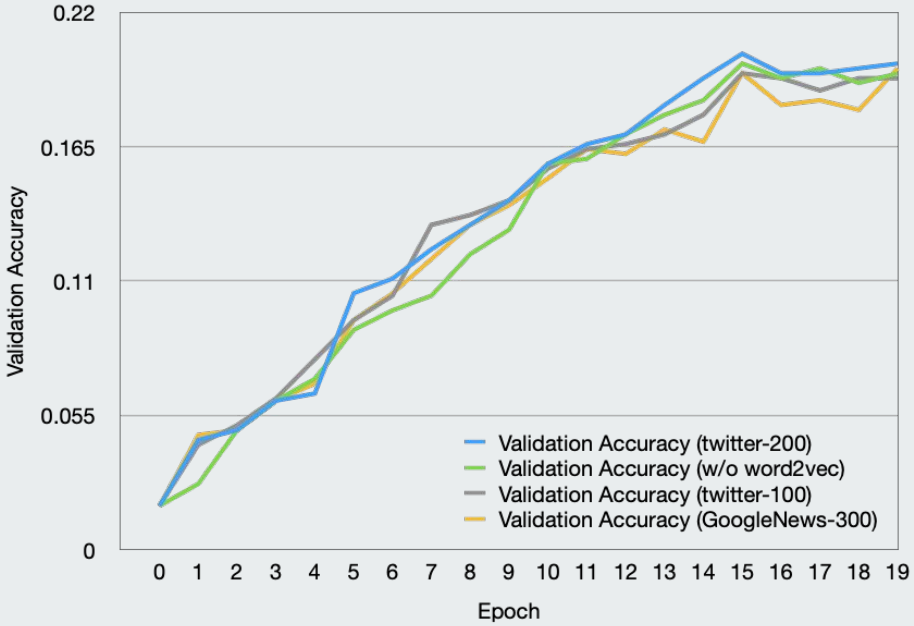


Figure. Validation accuracy of different models

Discussion & future improvement

- Data bias
  - Time period
  - Independence
  - Size
- Feature bias
- Embedding improvement

Questions?