Are People Located in the Places They Mention in Their Tweets? A Multimodal Approach

Zhaomin Xiao, Eduardo Blanco
University of North Texas, University of Arizona

Motivations

- Spatial information plays an important role in many applications, such as transportation planning and emergency management systems.
- Most previous work on spatial information with social media are on named entity recognition/disambiguation and location prediction.
- Location prediction, whose goal is to assign a location to a user, targets home location or real-time location.
- Taking into account images has been proven useful in many NLP tasks, such as sentiment analysis, although it is still underutilized in location prediction with tweets.
- We are the first to tackle the problem of real-time location prediction using text and images.

Background

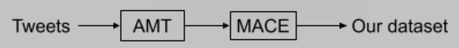
- AMT (Amazon Mechanical Turk) is a platform for annotation collection that allows people to publish and complete annotation tasks.
- MACE [1] is a tool that is designed to identify which annotators in AMT are trustworthy and predict the correct underlying annotations.
- BERT [2] is a pretrained language encoder, designed to generate a vector to capture the information contained in a sequence of text.
- VGG16 [3] is a pretrained image encoder, designed to generate a vector to capture the image information.

Objectives

- Create a dataset of tweets, with different modalities for each tweet annotated with spatial annotations about Twitter users.
- Analyze if the annotations change depending on the modality shown to annotators.
- Build multimodal neural networks to automatically predict the real-time locations of Twitter users.
- Analyze when and why different modalities complement each other.

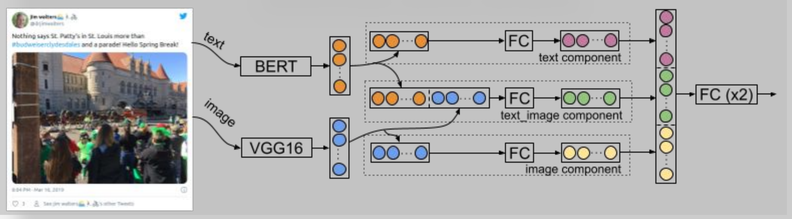
Methods

Dataset Creation



- Annotation question: Was the Twitter user at the mentioned location when the tweet was posted?
  - yes: The Twitter user was at the mentioned location.
  - no: I cannot tell if the Twitter user was at the mentioned location.
- We collect annotations for three variants of each tweet: 1) tweet not showing text, 2) tweet not showing image, and 3) full tweet.

Multimodal Neural Network



Results/Discussion

- Our dataset consists of 6,540 tweets, with each tweet having three annotations, corresponding whether it contains only text, only images, or both. Most annotations change if we show different modalities to annotators.
- Many annotations change if we show different modalities to annotators.

	yes	no
Text and image	51.09%	48.91%
Only text	80.93%	19.07%
Only image	69.74%	30.26%

Annotation Examples

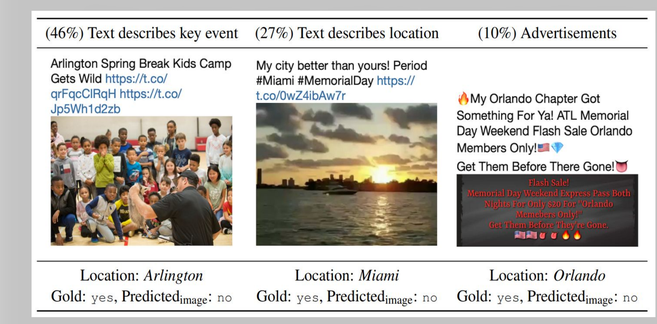
- Our experimental results show that taking into account both modalities is beneficial.

	P	R	F1
text	0.65	0.66	0.65
image	0.64	0.65	0.64
text_image	0.62	0.68	0.65
text + image + text_image	**0.64**	**0.74**	**0.68**

Results/Discussion (Cont.)

Most common errors made by text but fixed by images

Most common errors made by images but fixed by text



Most common remaining errors

References

1. Hovy et al., Learning Whom to Trust with MACE (2013).
2. Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019).
3. Karen Simonyan and Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition (2015).