Point-of-Interest Type Prediction using Text and Images

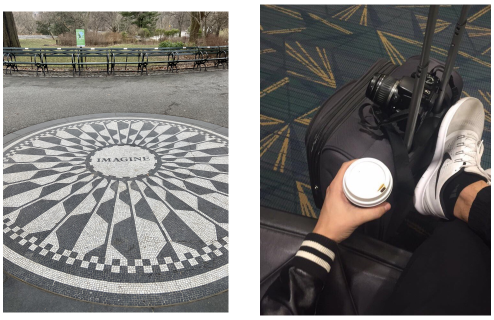
Danae Sánchez Villegas **Nikolaos Aletras**
Computer Science Department, University of Sheffield, UK
{dsanchezvillegas1, n.aletras}@sheffield.ac.uk

Abstract

Point-of-interest (POI) type prediction is the task of inferring the type of a place from where a social media post was shared. Inferring a POI's type is useful for studies in computational social science including sociolinguistics, geosemiotics, and cultural geography, and has applications in geosocial networking technologies such as recommendation and visualization systems. Prior efforts in POI type prediction focus solely on text, without taking visual information into account. However in reality, the variety of modalities, as well as their semiotic relationships with one another, shape communication and interactions in social media. This paper presents a study on POI type prediction using multimodal information from text and images available at posting time. For that purpose, we enrich a currently available data set for POI type prediction with the images that accompany the text messages. Our proposed method extracts relevant information from each modality to effectively capture interactions between text and image achieving a macro F1 of 47.21 across eight categories significantly outperforming the state-of-the-art method for POI type prediction based on text-only methods. Finally, we provide a detailed analysis to shed light on cross-modal interactions and the limitations of our best performing model.[1]

1 Introduction

A place is typically described as a physical space infused with human meaning and experiences that facilitate communication (Tuan, 1977). The multimodal content of social media posts (e.g. text, images, emojis) generated by users from specific places such as restaurants, shops, and parks, contribute to shaping a place's identity, by offering information about feelings elicited by participating in an activity or living an experience in that place (Tanasescu et al., 2013).



imagine all the people sharing all the world ∼

Next stop: NYC ✈️

Figure 1: Example of text and image content of sample tweets. Users share content that is relevant to their experiences and feelings in the location.

Fig. 1 shows examples of Twitter posts consisting of image-text pairs, shared from two different places or Point-of-Interests (POIs). Users share content that is relevant to their experience in the location. For example, the text *imagine all the people sharing all the world* which is accompanied by a photograph of the Imagine Mosaic in Central Park; and the text *Next stop: NYC* along with a picture of descriptive items that people carry at an airport such as luggage, a camera and a takeaway coffee cup.

Developing computational methods to infer the type of a POI from social media posts (Liu et al., 2012; Sánchez Villegas et al., 2020) is useful for complementing studies in computational social science including sociolinguistics, geosemiotics, and cultural geography (Kress et al., 1996; Scollon and Scollon, 2003; Al Zydjaly, 2014), and has applications in geosocial networking technologies such as recommendation and visualization systems (Alazzawi et al., 2012; Zhang and Cheng, 2018; van Weerdenburg et al., 2019; Liu et al., 2020b).

Previous work in natural language processing (NLP) has investigated the language that people

[1] Code and data are available here: https://github.com/danaesavi/poi-type-prediction

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7785–7797
November 7–11, 2021. ©2021 Association for Computational Linguistics

use in social media from different locations, by inferring the type of a POI of a given social media post using only text and posting time, ignoring the visual context (Sánchez Villegas et al., 2020). However, communication and interactions in social media are naturally shaped by the variety of available modalities and their semiotic relationships (i.e. how meaning is created and communicated) with one another (Georgakopoulou and Spilioti, 2015; Kruk et al., 2019; Vempala and Preoţiuc-Pietro, 2019).

In this paper, we propose POI type prediction using multimodal content available at posting time by taking into account textual and visual information. Our contributions are as follows:

- We enrich a publicly available data set of social media posts and POI types with images;
- We propose a multimodal model that combines text and images in two levels using: (i) a modality gate to control the amount of information needed from the text and image; (ii) a cross-attention mechanism to learn cross-modal interactions. Our model significantly outperforms the best state-of-the-art method proposed by Sánchez Villegas et al. (2020);
- We provide an in-depth analysis to uncover the limitations of our model and uncover cross-modal characteristics of POI types.

2 Related Work

2.1 POI Analysis

POIs have been studied to classify functional regions (e.g. residential, business, and transportation areas) and to analyze activity patterns using social media check-in data and geo-referenced images (Zhi et al., 2016; Liu et al., 2020a; Zhou et al., 2020a; Zhang et al., 2020). Zhou et al. (2020a) presents a model for classifying POI function types (e.g. bank, entertainment, culture) using POI names and a list of results produced by searching for the POI name in a web search engine. Zhang et al. (2020) makes use of social media check-ins and street-level images to compare the different activity patterns of visitors and locals, and uncover inconspicuous but interesting places for them in a city. A framework for extracting emotions (e.g. joy, happiness) from photos taken at various locations in social media is described in Kang et al. (2019).

2.2 POI Type Prediction

POI type prediction is related to geolocation prediction of social media posts that has been widely studied in NLP (Eisenstein et al., 2010; Roller et al., 2012; Dredze et al., 2016). However, while geolocation prediction aims to infer the exact geographical location of a post using language variation and geographical cues, POI type prediction is focused on identifying the characteristics associated with each type of place, regardless of its geographic location.

Previous work on POI type prediction from social media content has used Twitter posts (text and posting time), to identify the POI type from where a post was sent from (Liu et al., 2012; Sánchez Villegas et al., 2020). Liu et al. (2012) incorporate text, temporal features (posting hour) and user history information into probabilistic text classification models. Rather than a user-based study, our research aims to uncover the characteristics associated with various types of POIs. Sánchez Villegas et al. (2020) analyze semantic place information of different types of POIs by using text and temporal information (hour, and day of the week) of a Twitter's post. To the best of our knowledge, this is the first study to combine textual and visual features to classify POI types (e.g. arts & entertainment, nightlife spot) from social media messages, regardless of its geographic location.

2.3 Social Media Analysis using Text and Images

The combination of text and images of social media posts has been largely used for different applications such as sentiment analysis, (Nguyen and Shirai, 2015; Chambers et al., 2015), sarcasm detection (Cai et al., 2019) and text-image relation classification (Vempala and Preoţiuc-Pietro, 2019; Kruk et al., 2019). Moon et al. (2018b) propose a model for recognizing named entities from short social media texts using image and text. Cai et al. (2019) use a hierarchical fusion model to integrate image and text context with an attention-based fusion. Chinnappa et al. (2019) examine the possession relationships from text-image pairs in social media posts. Wang et al. (2020) use texts and images for predicting the keyphrases (i.e. representative terms) for a post by aligning and capturing the cross-modal interactions via cross-attention. Previous text-image classification in social media requires that the data is fully paired, i.e. every post

	Train		Dev		Test		
Category	**# Tweets**	**# Images**	**# Tweets**	**# Images**	**# Tweets**	**# Images**	**Tokens**
Arts & Entertainment	40,417	20,711	4,755	2,527	5,284	2,740	14.41
College & University	21,275	9,112	2,418	1,057	2,884	1,252	15.52
Food	6,676	2,969	869	351	724	280	14.34
Great Outdoors	27,763	13,422	4,173	2,102	3,653	1,948	13.49
Nightlife Spot	5,545	2,532	876	385	656	353	15.46
Professional & Other Places	30,640	13,888	3,381	1,499	3,762	1,712	16.46
Shop & Service	8,285	3,455	886	266	812	353	15.31
Travel & Transport	16,428	6,681	2,201	829	1,872	789	14.88
All	157,029	72,679 (46.28%)	19,559	9,006 (46.05%)	19,647	9,410 (47.90%)	14.92

Table 1: POI categories and data set statistics showing the number of tweets for each category, and number (%) of tweets having an accompanying image

contains an image and a text. However, this requirement may not be satisfied since not all posts contain both modalities [2]. This work considers both cases, (1) all modalities (text-image pairs) are available, and content in only one modality (text or image) is available.

Social media analysis research has also looked at the semiotic properties of text-image pairs in posts (Alikhani et al., 2019; Vempala and Preoţiuc-Pietro, 2019; Kruk et al., 2019). Vempala and Preoţiuc-Pietro (2019) investigate the relationship between text and image content by identifying overlapping meaning in both modalities, those where one modality contributes with additional details, and cases where each modality contributes with different information. Kruk et al. (2019) analyze the relationship between the text-image pairs and find that when the image and caption diverge semiotically, the benefit from multimodal modeling is greater.

3 Task & Data

Sánchez Villegas et al. (2020) define POI type prediction as a multi-class classification task where given the text content of a post, the goal is to classify it in one of the M POI categories. In this work, we extend this task definition to include images in order to capture the semiotic relationships between the two modalities. For that purpose, we consider a social media post P (e.g. tweet) to comprise of a text and image pair (x^t, x^v), where $x^t \in \mathbb{R}^{d_t}$ and $x^v \in \mathbb{R}^{d_v}$ are the textual and visual vector representations respectively.

3.1 POI Data

We use the data set introduced by Sánchez Villegas et al. (2020) which contains $196,235$ tweets written in English, labeled with one out of the eight POI broad type categories shown in Table 1, which correspond to the 8 primary top-level POI categories in ‘Places by Foursquare’, a database of over 105 million POIs worldwide managed by Foursquare. To generalize to locations not present in the training set, we use the same location-level data splits (train, dev, test) as in Sánchez Villegas et al. (2020), where each split contains tweets from different locations.

3.2 Image Collection

We use the Twitter API to collect the images that accompany each textual post in the data set. For the tweets that have more than one image, we select the first available only. This results in $91,224$ tweets with at least one image. During the image processing (see Section 5.3) we removed 129 images because we found they were either damaged, absent[3], or no objects were detected, resulting in $91,095$ text-image pairs (see Table 1 for data statistics). In order to deal with the rest of the tweets with no associated image, we pair them with a single ‘average’ image computed over all images in the train set: $x^v = avg(x^v_{tr})$. The intuition behind this approach is to generate a ‘noisy’ image that is not related and does not add to the meaning (Vempala and Preoţiuc-Pietro, 2019).[4]

3.3 Exploratory Analysis of Image Data

To shed light on the characteristics of the collected images, we apply object detection on the images

[2] See: https://buffer.com/resources/twitter-data-1-million-tweets/

[3] Removed by Twitter due to violations to the Twitter Rules and Terms of Service.

[4] Early experimentation with associating tweets with the image of the most similar tweet that contains a real image from the training data yielded similar performance.

Category	Common Objects in Images
Arts & Entertainment	light, pants, shirt, arm, picture, hair, glasses, line, girl, jacket
College & University	pants, shirt, line, hair, arm, picture, light, glasses, girl, trees
Food	cup, picture, spoon, meat, knife, arm, glasses, shirt, pants, handle
Great Outdoors	trees, arm, pants, cloud, hill, line, shirt, grass, picture, glasses
Nightlife Spot	arm, picture, shirt, light, hair, pants, glasses, mouth, girl, cup
Professional & Other Places	pants, shirt, picture, light, hair, screen, line, arm, glasses, girl
Shop & Service	picture, pants, arm, shirt, glasses, light, hair, line, girl, letters
Travel & Transport	pants, shirt, light, screen, arm, hair, glasses, picture, chair, line

Table 2: Most common objects for each POI category.

collected using Faster-RCNN (Ren et al., 2016) pretrained on Visual Genome (Krishna et al., 2017; Anderson et al., 2018). Table 2 shows the most common objects for each specific category. We observe that most objects are related to items one would find in each place category (e.g. 'spoon', 'meat', 'knife' in *Food*). Clothing items are common across category types (e.g. 'shirt', 'jacket', 'pants') suggesting the presence of people in the images. A common object tag of the *Shop & Service* category is 'letters', which concerns images that contain embedded text. Finally, the category *Great Outdoors* includes object tags such as 'cloud', 'hill', and 'grass', words that describe the landscape of this type of place.

4 Multimodal POI Type Prediction

4.1 Text and Image Representation

Given a text-image post $P = (x^t, x^v)$, $x^t \in \mathbb{R}^{d_t}$, $x^v \in \mathbb{R}^{d_v}$, we first compute text and image encoding vectors f^t, f^v respectively.

Text We use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) to obtain the text feature representations f^t by extracting the 'classification' [CLS] token.

Image For encoding the images, we use Xception (Chollet, 2017) pre-trained on ImageNet (Deng et al., 2009).[5] We extract convolutional feature maps for each image and we apply average pooling to obtain the image representation f^v.

4.2 MM-Gate

Given the complex semiotic relationship between text and image, we need a weighting strategy that assigns more importance to the most relevant modality while suppressing irrelevant information. Thus, a first approach is to use gated multimodal fusion (MM-Gate), similar to the approach proposed by Arevalo et al. (2020) to control the contribution of text and image to the POI type prediction. Given f^t, f^v the text and visual vectors, we obtain the multimodal representation h of a post P as follows:

$$h^t = tanh(W^t f^t + b^t) \quad (1)$$

$$h^v = tanh(W^v f^v + b^v) \quad (2)$$

$$z = \sigma(W^z[f^t; f^v] + b^z) \quad (3)$$

$$h = z * h^t + (1 - z) * h^v \quad (4)$$

where $W^t \in \mathbb{R}^{d_t}$, $W^v \in \mathbb{R}^{d_v}$ and $W^z \in R^{d_t + d_v}$ are learnable parameters, *tanh* is the activation function and $h^t, h^v \in \mathbb{R}$ are projections of f^t and f^v. $[;]$ denotes concatenation and σ is the sigmoid activation function. h is a weighted combination of the textual and visual information h^t and h^v respectively. We fine-tune the entire model by adding a classification layer with a softmax activation function for POI type prediction

4.3 MM-XAtt

The MM-Gate model does not capture interactions between text and image that might be beneficial for learning semiotic relationships. To model cross-modal interactions, we adapt the cross-attention mechanism (Tsai et al., 2019; Tan and Bansal, 2019) to combine text and image information for multimodal POI type prediction (MM-XAtt). Cross-attention consists of two attention layers, one from textual f^t to visual features f^v and one from visual to textual features. We first linearly project the text and visual representations to obtain the same dimensionality (d_{proj}). Then, we compute the scaled dot attention ($a = softmax\frac{(Q(K)^T)}{\sqrt{d_{proj}}}V$) with the projected textual vector as query (Q), and the projected image vector as the key (K) and values (V), and vice versa. The multimodal representation h is the sum of the resulting attention layers. The entire model is fine-tuned by adding a classification layer with a softmax activation function.

4.4 MM-Gated-XAtt

Vempala and Preoţiuc-Pietro (2019) have demonstrated that the relationship between the text and

[5] Early experimentation with ResNet101 (He et al., 2016) and EfficientNet (Tan and Le, 2019) yielded similar results.

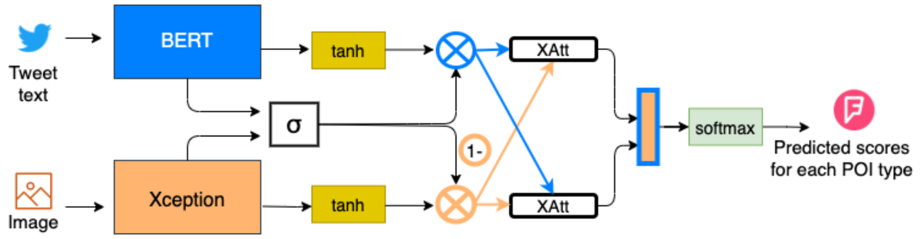


Figure 2: Overview of our MM-Gated-XAtt model which combines features from text and image modalities for POI type prediction.

image in a social media post is complex. Images may or may not add meaning to the post and the text content (or meaning) may or may not correspond to the image. We hypothesize that this might actually happen in posts made from particular locations, i.e. language and visual information may or may not be related. To address this, we propose (1) using gated multimodal fusion to manage the flow of information from each modality, and (2) also learn cross-modal interactions by using cross-attention on top of the gated multimodal mechanism. Fig. 2 shows an overview of our model architecture (MM-Gated-XAtt). Given the text and image representations f^t, f^v respectively, we compute h^t, h^v, and z as in Equation 1, 2 and 3. Next, we apply cross-attention using two attention layers where the query and context vectors are the weighted representations of the text and visual modalities, $z * h^t$ and $(1 - z) * h^v$, and vice versa. The multimodal context vector h is the sum of the resulting attention layers. Finally, we fine-tune the model by passing h through a classification layer for POI type prediction with a softmax activation function.

5 Experimental Setup

5.1 Baselines

We compare our models against (1) text-only; (2) image-only; and (3) other state-of-the-art multimodal approaches.[6]

Text-only We fine-tune BERT for POI type classification by adding a classification layer with softmax activation function on top of the [CLS] token which is the best performing model in Sánchez Villegas et al. (2020).

Image-only We fine-tune three pre-trained models that are popular in various computer vision classification tasks: (1) ResNet101 (He et al., 2016); (2) EfficientNet (Tan and Le, 2019); and (3) Xception (Chollet, 2017). Each model is fine-tuned on POI type classification by adding an output softmax layer.

Text and Image For combining text and image information, we experiment with different standard fusion strategies: (1) we project the image representation f^v, to the same dimensionality as $f^t \in \mathbb{R}^{d_t}$ using a linear layer and then we concatenate the vectors (**Concat**); (2) we project the textual and visual features to the same space and then we apply self-attention to learn weights for each modality (**Attention**); (3) we also adapt the guided attention introduced by Anderson et al. (2018) for learning attention weights at the object-level (and other salient regions) rather than equally sized grid-regions (**Guided Attention**); (4) we compare against **LXMERT**, a transformer-based model that has been pre-trained on text and image pairs for learning cross-modality interactions (Tan and Bansal, 2019). All models are fine-tuned by adding a classification layer with a softmax activation function for POI type prediction. Finally, we evaluate a simple ensemble strategy by using LXMERT for classifying tweets that are originally accompanied by an image and BERT for classifying text-only tweets (**Ensemble**).

5.2 Text Processing

We use the same tokenization settings as in Sánchez Villegas et al. (2020). For each tweet, we lowercase text and replace URLs and @-mentions of users with placeholder tokens.

5.3 Image Processing

Each image is resized to (224 × 224) pixels representing a value for the red, green and blue color in the range of $[0, 255]$. The pixel values of all images are normalized. For LXMERT and Guided Attention fusion, we extract *object-level* features using Faster-RCNN (Ren et al., 2016) pretrained

[6] We include a majority class baseline (i.e. assigning all instances in the test set the most frequent label in the train set).

Model	F1	P	R
Majority	5.30	3.36	12.50
BERT (Sánchez Villegas et al., 2020)	43.67 (0.01)	**48.44** (0.02)	41.33 (0.01)
ResNet	21.11 (1.81)	23.23 (2.09)	29.90 (3.31)
EfficientNet	24.72 (0.76)	28.05 (0.28)	35.48 (0.23)
Xception	23.64 (0.44)	25.62 (0.50)	34.12 (0.49)
Concat-BERT+ResNet	43.28 (0.37)	42.72 (0.51)	47.59 (0.45)
Concat-BERT+EfficientNet	41.56 (0.71)	41.54 (0.88)	43.97 (0.79)
Concat-BERT+Xception	44.00 (0.52)	43.34 (0.70)	48.35 (0.75)
Attention-BERT+Xception	42.89 (0.44)	42.74 (0.19)	46.78 (1.28)
Guided Attention-BERT+Xception	41.53 (0.57)	41.10 (0.55)	45.36 (0.48)
LXMERT	40.17 (0.62)	40.26 (0.24)	42.25 (2.38)
Ensemble-BERT+LXMERT	43.82 (0.47)	43.50 (0.20)	44.67 (0.66)
MM-Gate	44.64 (0.65)	43.67 (0.49)	48.50 (0.18)
MM-XAtt	27.31 (1.58)	37.06 (2.66)	29.71 (0.60)
MM-Gated-XAtt (Ours)	**47.21**† (1.70)	46.83 (1.45)	**50.69** (2.21)

Table 3: Macro F1-Score, precision (P) and recall (R) for POI type prediction (± std. dev.) Best results are in bold. † indicates statistically significant improvement (t-test, $p < 0.05$) over BERT (Sánchez Villegas et al., 2020).

on Visual Genome (Krishna et al., 2017) following Anderson et al. (2018). We keep 36 objects for each image as in Tan and Bansal (2019).

5.4 Implementation Details

We select the hyperparameters for all models using early stopping by monitoring the validation loss using the Adam optimizer (Kingma and Ba, 2014). Because the data is imbalanced, we estimate the class weights using the 'balanced' heuristic (King and Zeng, 2001). All experiments are performed using a Nvidia V100 GPU.

Text-only We fine-tune BERT for 20 epochs and choose the epoch with the lowest validation loss. We use the pre-trained base-uncased model for BERT (Vaswani et al., 2017; Devlin et al., 2019) from HuggingFace library (12-layer, 768-dimensional) with a maximal sequence length of 50 tokens. We fine-tune BERT for 2 epochs and learning rate $\eta = 2e^{-5}$ with $\eta \in \{2e^{-5}, 3e^{-5}, 5e^{-5}\}$.

Image-only For ResNet101, we fine-tune for 5 epochs with learning rate $\eta = 1e^{-4}$ and dropout $\delta = 0.2$ (δ in $[0, 0.5]$ using random search) before passing the image representation through the classification layer. EfficientNet is fine-tuned for 7 epochs with $\eta = 1e^{-5}$ and $\delta = 0.5$. Xception is fine-tuned for 6 epochs with $\eta = 1e^{-5}$ and $\delta = 0.5$.

Text and Image Concat-BERT+Xception, Concat-BERT+ResNet and Guided Attention-BERT+Xception are fine-tuned for 2 epochs with $\eta = 1e^{-5}$ and $\delta = 0.25$; Concat-BERT+EfficientNet for 4 epochs with $\eta = 1e^{-5}$ and $\delta = 0.25$; Attention-BERT+Xception for 3 epochs with $\eta = 1e^{-5}$ and $\delta = 0.25$; MM-XAtt for 3 epochs with $\eta = 1e^{-5}$ and $\delta = 0.15$; MM-Gate and MM-Gated-XAtt for 2 epochs with $\eta = 1e^{-5}$ and $\delta = 0.05$; $\eta \in \{2e^{-5}, 3e^{-5}, 5e^{-5}\}$, δ from $[0, 0.5]$ (random search) before passing through the classification layer. The dimensionality of the multimodal representation h (Eq. 4) is set to 200. We fine-tune LXMERT for 4 epochs with $\eta = 1e^{-5}$ where $\eta \in \{1e^{-3}, 1e^{-4}, 1e^{-5}\}$ and dropout $\delta = 0.25$ (δ in $[0, 0.5]$, random search) before passing through the classification layer.

5.5 Evaluation

We evaluate the performance of all models using macro F1, precision, and recall. Results are obtained over three runs using different random seeds reporting the average and the standard deviation.

6 Results

The results of POI type prediction are presented in Table 3. We first examine the impact of each modality by analyzing the performance of the unimodal models, then we investigate the effect of multimodal methods for POI type prediction, and finally we examine the performance of our proposed model MM-Gated-XAtt by analyzing each component independently.

Text-Image Only	
Model	**F1**
LXMERT	47.72 (0.98)
MM-Gate	45.87 (1.48)
MM-XAtt	48.93 (2.08)
MM-Gated-XAtt (Ours)	**57.64** (3.64)

Table 4: Macro F1-Score for POI type prediction on tweets that are originally accompanied by an image. Best results are in bold.

We observe that the text-only model (BERT) achieves 43.67 F1 which is substantially higher than the performance of image-only models (e.g. the best performing EfficientNet model obtains 24.72 F1). This suggests that text encapsulates more relevant information for this task than images on their own, similar to other studies in multimodal computational social science (Wang et al., 2020; Ma et al., 2021).

Models that simply concatenate text and image vectors have close performance to BERT (44.0 for Concat-BERT+Xception) or lower (41.56 for Concat-BERT+EfficientNet). This suggests that assigning equal importance to text and image information can deteriorate performance. It also shows that modeling cross-modal interactions is necessary to boost performance of POI type classification models.

Surprisingly, we observe that the pre-trained multimodal LXMERT fails to improve over BERT (40.17 F1) while its performance is lower than simpler concatenative fusion models. We speculate that this is because LXMERT is pretrained on data where both, text and image modalities share common semantic relationships which is the case in standard vision-language tasks including image captioning and visual question answering (Zhou et al., 2020b; Lu et al., 2019). On the other hand, text-image relationships in social media data for inferring the type of location from which a message was sent are more diverse, highlighting the particular challenges for modeling text and images together (Hessel and Lee, 2020).

Our proposed MM-Gated-XAtt model achieves 47.21 F1 which significantly (t-test, $p < 0.05$) improves over BERT, the best performing model in Sánchez Villegas et al. (2020) and consistently outperforms all other image-only and multimodal approaches. This confirms our main hypothesis that modeling text with image jointly to learn the interactions between modalities benefit performance in POI type prediction. We also observe that using only the gating mechanism (MM-Gate) outperforms (44.64 F1) all other models except for MM-Gated-XAtt. This highlights the importance of controlling the information flow for the two modalities. Using cross-attention on its own (MM-XAtt), on the other hand, fails to improve over other multimodal approaches, implying that learning cross-modal interactions is not sufficient on its own. This supports our hypothesis that language and visual information in posts sent from specific locations may be or may not be related, and that managing the flow of information from each modality improves the classifier's performance.

Finally, we investigate using less noisy text-image pairs in alignment with related computational social science studies involving text and images (Moon et al., 2018b; Cai et al., 2019; Chinnappa et al., 2019). We train and test LXMERT, MM-Gate, MM-XAtt, and MM-Gated-XAtt on tweets that are originally accompanied by an image (see Section 3), excluding all text-only tweets. The results are shown in Table 4. In general, performance is higher for all models using less noisy data. Our proposed model MM-Gated-XAtt consistently achieves the best performance (57.64 F1). In addition, we observe that LXMERT and MM-XAtt produce similar results (47.72 and 48.93 F1 respectively) suggesting that cross-attention can be applied directly to text-image pairs in low-noise settings without hurting the model performance. The benefit of controlling the flow of information through a gating mechanism, on the other hand, strongly improves model robustness.

6.1 Training on Text-Image Pairs Only

To compare the effect of the 'average' image (see Section 3) on the performance of the models, we train MM-Gate, MM-XAtt, and MM-Gated-XAtt on tweets that are originally accompanied by an image excluding all text-only tweets; and we test on all tweets as in our original setting (text-only tweets are paired with the 'average' image). The results are shown in Table 5. MM-Gated-XAtt is consistently the best performing model, followed by MM-Gate. However, their performance is inferior than when models are trained on all tweets using the 'average' image as in the original setting. This suggests that the gate operation not only regulates the flow of information for each modality but also learns how to use the noisy modality to

Text-Image Only -> All	
Model	**F1**
MM-Gate	40.67 (0.45)
MM-XAtt	31.00 (0.89)
MM-Gated-XAtt (Ours)	**42.45** (2.94)

Table 5: Macro F1-Score for POI type prediction. Models are trained on tweets that are originally accompanied by an image. Results are on all tweets. Best results are in bold.

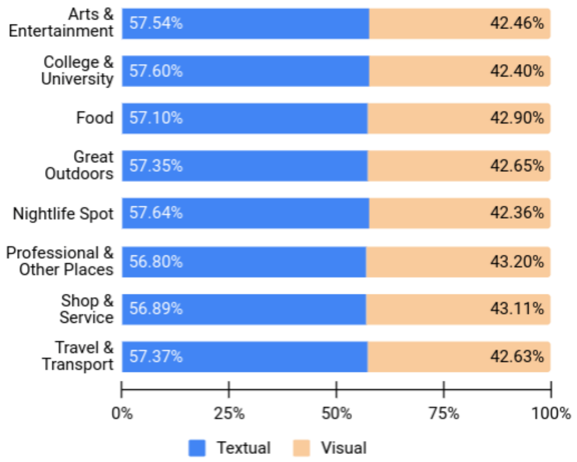


Figure 3: Average percentage of MM-Gated-XAtt activations for the textual and visual modalities for each POI category on the test set.

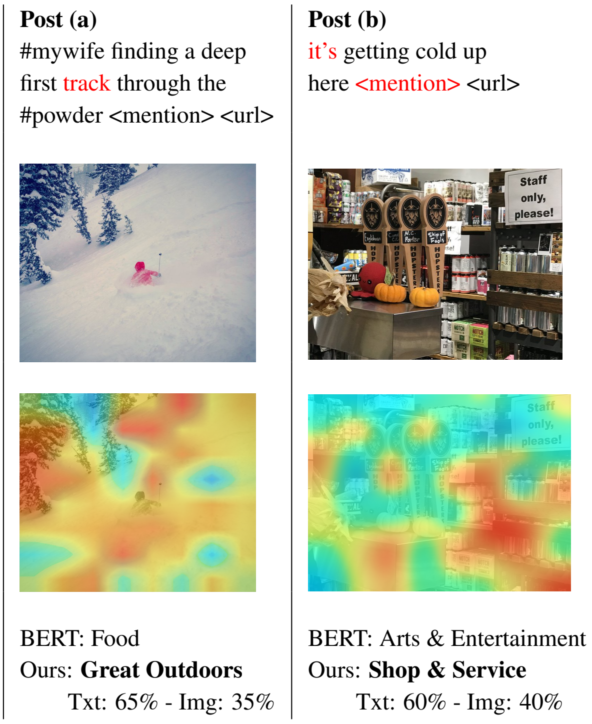
improve classification prediction. This result is similar to findings by (Arevalo et al., 2020).

7 Analysis

7.1 Modality Contribution

To determine the influence of each modality in MM-Gated-XAtt when assigning a particular label to a tweet, we compute the average percentage of activations for the textual and visual modalities for each POI category on the test set. The outcome of this analysis is depicted in Fig. 3. As anticipated, the textual modality has a greater influence on the model prediction, which is consistent with our findings in Section 6. The category where the visual modality has greater impact on the predicted label is *Professional & Other Places* (43.20%) followed by *Shop & Service* (43.11%).

To examine how the visual information impacts the POI type prediction task, Fig. 4 shows examples of posts where the contribution of the image is large while the text-only model (BERT) misclassified the POI category. We observe that the text content of Post (a) misled BERT towards *Food*, probably due to the term 'powder'. On the other hand, MM-Gated-XAtt can filter irrelevant information from the text, and prioritize relevant content from the image in order to assign the correct POI category for Post (a) (*Great Outdoors*). Likewise, Post (b) was correctly classified by MM-Gated-XAtt as *Shop & Service* and misclassified by BERT as *Arts & Entertainment*. For this post 40% of the contribution corresponds to the image and 60% to text. This shows how image information can help to address the ambiguity in short texts (Moon et al., 2018a), improving POI type prediction.


Post (a)
#mywife finding a deep first track through the #powder <mention> <url>

Post (b)
it's getting cold up here <mention> <url>

BERT: Food
Ours: **Great Outdoors**
Txt: 65% - Img: 35%

BERT: Arts & Entertainment
Ours: **Shop & Service**
Txt: 60% - Img: 40%

Figure 4: POI type predictions of MM-Gated-XAtt (Ours) and BERT Sánchez Villegas et al. (2020) showing the contribution of each modality (%) and the XAtt visualization. Correct predictions are in bold.

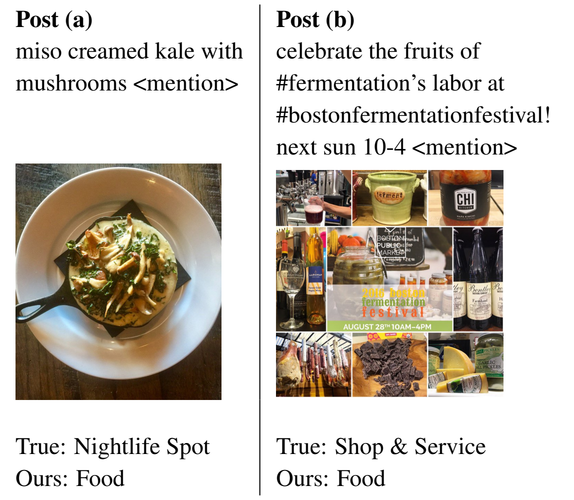
7.2 Cross-attention (XAtt)

Fig. 4 shows examples of the XAtt visualization. We note that the model focuses on relevant nouns and pronouns (e.g. 'track', 'it'), which are common informative words in vision-and-language tasks Tan et al. (2019). Moreover, our model focuses on relevant words such as 'track' for classifying Post (a) as *Great Outdoors*. Lastly, we observe that the XAtt often captures a general image information, with emphasis on specific sections for the predicted POI category such as the pine trees for *Great Outdoors* and the display racks for *Shop & Service*.

7.3 Error Analysis

To shed light on the limitations of our multimodal MM-Gated-XAtt model for predicting POI types, we performed an analysis of misclassifications. In general, we observe that the model struggles with identifying POI categories where people might perform similar activities in each of them such as *Food*, *Nightlife Spot*, and *Shop & Service* similar to findings by Ye et al. (2011).

Fig. 5 (a) and (b) show examples of tweets misclassified as *Food* by the MM-Gated-XAtt model. Post (a) belongs to the category *Nightlife Spot* and Post (b) belongs to the *Shop & Service* category. In both cases, the text and image content is related to the *Food* category, misleading the classifier towards this POI type. Posting about food is a common practice in hospitality establishments such as restaurants and bars (Zhu et al., 2019), where customers are more likely to share content such as photos of dishes and beverages, intentionally designed to show that are associated with the particular context and lifestyle that a specific place represents (Homburg et al., 2015; Brunner et al., 2016; Apaolaza et al., 2021). Similarly, Post (b) shows an example of a tweet that promotes a POI by communicating specific characteristics of the place (Kruk et al., 2019; Aydin, 2020). To correctly classify the category of POIs, the model might need access to deeper contextual information about the locations (e.g. finer subcategories of a type of place and how POI types are related to one another).


Post (a)
miso creamed kale with mushrooms <mention>

True: Nightlife Spot
Ours: Food

Post (b)
celebrate the fruits of #fermentation's labor at #bostonfermentationfestival! next sun 10-4 <mention>

True: Shop & Service
Ours: Food

Figure 5: Example of misclassifications made by our MM-Gated-XAtt model.

8 Conclusion and Future Work

This paper presents the first study on multimodal POI type classification using text and images from social media posts motivated by studies in geosemiotics, visual semiotics and cultural geography. We enrich a publicly available data set with images and we propose a multimodal model that uses: (1) a gate mechanism to control the information flow from each modality; (2) a cross-attention mechanism to align and capture the interactions between modalities. Our model achieves state-of-the-art performance for POI type prediction significantly outperforming the previous text-only model and competitive pretrained multimodal models.

In future work, we plan to perform more granular prediction of POI types and user information to provide additional context to the models. Our models could also be used for modeling other tasks where text and images naturally occur in social media such as analyzing political ads (Sánchez Villegas et al., 2021), parody (Maronikolakis et al., 2020) and complaints (Preoţiuc-Pietro et al., 2019; Jin and Aletras, 2020, 2021).

Acknowledgments

We would like to thank Mali Jin, Panayiotis Karachristou and all reviewers for their valuable feedback. DSV is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1. NA is supported by a Leverhulme Trust Research Project Grant.

Ethical Statement

Our work complies with Twitter data policy for research,[7] and has received approval from the Ethics Committee of our institution (Ref. No 039665).

References

Najma Al Zydjaly. 2014. 8. geosemiotics: Discourses in place. In *Interactions, Images and Texts*, pages 63–76. De Gruyter Mouton.

Ahmed N Alazzawi, Alia I Abdelmoty, and Christopher B Jones. 2012. What can I do there? Towards the automatic discovery of place-related services and activities. *International Journal of Geographical Information Science*, 26(2):345–364.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A cor-

[7] See: https://developer.twitter.com/en/developer-terms/agreement-and-policy

pus of image-text discourse relations. 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)*, pages 570–575, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Vanessa Apaolaza, Mario R. Paredes, Patrick Hartmann, and Clare D'Souza. 2021. How does restaurant's symbolic design affect photo-posting on instagram? the moderating role of community commitment and coolness. *Journal of Hospitality Marketing & Management*, 30(1):21–37.

John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A González. 2020. Gated multimodal networks. *Neural Computing and Applications*, pages 1–20.

Gökhan Aydin. 2020. Social media engagement and organic post effectiveness: A roadmap for increasing the effectiveness of social media use in hospitality industry. *Journal of Hospitality Marketing & Management*, 29(1):1–21.

Christian Boris Brunner, Sebastian Ullrich, Patrik Jungen, and Franz-Rudolf Esch. 2016. Impact of symbolic product design on brand evaluations. *Journal of product & brand management*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. 2015. Identifying political sentiment between nation states with social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 65–75, Lisbon, Portugal. Association for Computational Linguistics.

Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco. 2019. Extracting possessions from social media: Images complement language. 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)*, pages 663–672, Hong Kong, China. Association for Computational Linguistics.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. 2016. Geolocation for Twitter: Timing matters. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1064–1069, San Diego, California. Association for Computational Linguistics.

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, Cambridge, MA. Association for Computational Linguistics.

Alexandra Georgakopoulou and Tereza Spilioti. 2015. *The Routledge handbook of language and digital communication*. Routledge.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Christian Homburg, Martin Schwemmle, and Christina Kuehnl. 2015. New product design: Concept, measurement, and consequences. *Journal of marketing*, 79(3):41–56.

Mali Jin and Nikolaos Aletras. 2020. Complaint identification in social media with transformer networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1765–1771, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mali Jin and Nikolaos Aletras. 2021. Modeling the severity of complaints in social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2264–2274, Online. Association for Computational Linguistics.

Yuhao Kang, Qingyuan Jia, Song Gao, Xiaohuan Zeng, Yueyao Wang, Stephan Angsuesser, Yu Liu, Xinyue Ye, and Teng Fei. 2019. Extracting human emotions at different places based on facial expressions and spatial clustering analysis. *Transactions in GIS*, 23(3):450–480.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

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

Gunther R Kress, Theo Van Leeuwen, et al. 1996. *Reading images: The grammar of visual design*. Psychology Press.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. 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)*, pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Haibin Liu, Bo Luo, and Dongwon Lee. 2012. Location type classification using tweet content. In *2012 11th International Conference on Machine Learning and Applications*, volume 1, pages 232–237. IEEE.

Kang Liu, Ling Yin, Feng Lu, and Naixia Mou. 2020a. Visualizing and exploring poi configurations of urban regions on poi-type semantic space. *Cities*, 99:102610.

Tongcun Liu, Jianxin Liao, Zhigen Wu, Yulong Wang, and Jingyu Wang. 2020b. Exploiting geographical-temporal awareness attention for next point-of-interest recommendation. *Neurocomputing*, 400:227–237.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chunpeng Ma, Aili Shen, Hiyori Yoshikawa, Tomoya Iwakura, Daniel Beck, and Timothy Baldwin. 2021. On the (in)effectiveness of images for text classification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 42–48, Online. Association for Computational Linguistics.

Antonios Maronikolakis, Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. Analyzing political parody in social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4373–4384, Online. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018a. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018b. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364, Beijing, China. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

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, Jeju Island, Korea. Association for Computational Linguistics.

Danae Sánchez Villegas, Saeid Mokaram, and Nikolaos Aletras. 2021. Analyzing online political advertisements. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3669–3680, Online. Association for Computational Linguistics.

Danae Sánchez Villegas, Daniel Preotiuc-Pietro, and Nikolaos Aletras. 2020. Point-of-interest type inference from social media text. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 804–810, Suzhou, China. Association for Computational Linguistics.

Ron Scollon and Suzie Wong Scollon. 2003. *Discourses in place: Language in the material world*. Routledge.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. 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)*, pages 2610–2621, Minneapolis, Minnesota. Association for Computational Linguistics.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Vlad Tanasescu, Christopher B Jones, Gualtiero Colombo, Martin J Chorley, Stuart M Allen, and Roger M Whitaker. 2013. The personality of venues: Places and the five-factors ('big five') model of personality. In *Fourth IEEE International Conference on Computing for Geospatial Research and Application*, pages 76–81.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Yi-Fu Tuan. 1977. *Space and place: The perspective of experience*. U of Minnesota Press.

Demi van Weerdenburg, Simon Scheider, Benjamin Adams, Bas Spierings, and Egbert van der Zee. 2019. Where to go and what to do: Extracting leisure activity potentials from web data on urban space. *Computers, Environment and Urban Systems*, 73:143–156.

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.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of Twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, Florence, Italy. Association for Computational Linguistics.

Yue Wang, Jing Li, Michael Lyu, and Irwin King. 2020. Cross-media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3311–3324, Online. Association for Computational Linguistics.

Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. 2011. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–528.

Fan Zhang, Jinyan Zu, Mingyuan Hu, Di Zhu, Yuhao Kang, Song Gao, Yi Zhang, and Zhou Huang. 2020. Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems*, 81:101478.

Siyuan Zhang and Hong Cheng. 2018. Exploiting context graph attention for poi recommendation in location-based social networks. In *International Conference on Database Systems for Advanced Applications*, pages 83–99. Springer.

Ye Zhi, Haifeng Li, Dashan Wang, Min Deng, Shaowen Wang, Jing Gao, Zhengyu Duan, and Yu Liu. 2016. Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data. *Geospatial Information Science*, 19(2):94–105.

Chaoran Zhou, Hang Yang, Jianping Zhao, and Xin Zhang. 2020a. Poi classification method based on feature extension and deep learning. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 24(7):944–952.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020b. Unified Vision-Language Pre-Training for Image Captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049.

Jiang Zhu, Lan Jiang, Wenyu Dou, and Liang Liang. 2019. Post, eat, change: the effects of posting food photos on consumers' dining experiences and brand evaluation. *Journal of Interactive Marketing*, 46:101–112.