Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-14296-8

1227: CONTENT-BASED IMAGE RETRIEVAL

Detecting and locating trending places using multimodal social network data

Luis Lucas[1] · **David Tomás**[1] · **Jose Garcia-Rodriguez**[1]

Received: 1 August 2022 / Revised: 10 November 2022 / Accepted: 3 December 2022

Abstract
This paper presents a machine learning-based classifier for detecting points of interest through the combined use of images and text from social networks. This model exploits the transfer learning capabilities of the neural network architecture CLIP (Contrastive Language-Image Pre-Training) in multimodal environments using image and text. Different methodologies based on multimodal information are explored for the geolocation of the places detected. To this end, pre-trained neural network models are used for the classification of images and their associated texts. The result is a system that allows creating new synergies between images and texts in order to detect and geolocate trending places that has not been previously tagged by any other means, providing potentially relevant information for tasks such as cataloging specific types of places in a city for the tourism industry. The experiments carried out reveal that, in general, textual information is more accurate and relevant than visual cues in this multimodal setting.

Keywords Multimodal classification · Location-based retrieval · Transformers · Social networks

1 Introduction

Nowadays, machine learning-based technologies and the analysis of large amounts of data are used in multiple domains. Such is the case of recommender systems that are used on a

Luis Lucas, David Tomás and Jose Garcia-Rodriguez contributed equally to this work.

✉ Luis Lucas
luis.lucas@ua.es

David Tomás
dtomas@dlsi.ua.es

Jose Garcia-Rodriguez
jgr@ua.es

[1] Institute of Informatics Research, University of Alicante, Ctra San Vicente del Raspeig, San Vicente del Raspeig, 03690, Alicante, Spain

daily basis in many personal and professional situations. Even though Internet users generate large amounts of data, gathering all this information poses a challenge since it is mostly generated through proprietary applications belonging to private companies that keep its exploitation rights. Of these data sources, it is social networks that provide the most information or potential knowledge in almost any domain. Previous studies have proposed the use of data from social networks for many different tasks. For example, a topic of great interest at present is the detection of fake news in social media [30].

This work presents an approach to geolocate place mentions in social networks through the combined use of images and text. A set of baselines from pre-trained state-of-the-art models were defined to geolocate user posts using only text, only images, and the combination of the two. Different conditional ensembles were proposed to detect in which situations the system should pay more attention to texts or images in order to improve the baselines. The InstaCities1M corpus [12] was used in the experiments. This dataset contains one million pairs of associated images and texts from Instagram that were extracted by using queries related to the ten most populated English speaking cities over the world, comprising 100,000 image-text pairs for each one.

At a second stage, the Transformer model CLIP (Contrastive Language-Image Pre-Training) [27] was used to identify places on a multimodal setting, using again image-text pairs from Instagram. Following the procedure described in [23], the problem was addressed as a classification task where the possible labels were the 205 categories used in the Places dataset [42]. These labels correspond to locations or scenes, both indoor and outdoor.

CLIP is used in this work as a zero-shot classifier. Zero-shot learning [3] is a machine learning technique in which a pre-trained model is made to generalise on previously unseen categories. This implies that categories at test time are different from categories at training time. Zero-shot approaches generally work by associating the observed (training) and unobserved (testing) categories through some kind of auxiliary information, which encodes the observable properties that characterise the objects. These models are usually pre-trained with large amounts of data to enable transfer learning, that is, to use the knowledge already learned by the model during its training state.

In the present work, zero-shot learning is used to alleviate the problem of data labelling, which is usually a labour-intensive process where it is common to end up with an imbalanced dataset that lacks enough training data for each class. The zero-shot approach makes it possible to correctly categorise places from previously unseen classes, which is a key requirement in any truly autonomous place identification system, taking into account the unlimited number of different locations that may exist. In this learning framework there is no need to worry about the class imbalance of datasets. Furthermore, it avoids the need for building new models or retraining existing ones.

Finally, different applications of the experiments performed on the InstaCities1M dataset are shown, detecting the most popular points of interest or type of places in a given city. These trends can be also compared between cities or time periods. Thus, the approach presented in this work has many applications, especially in areas related to economy, sociology, and tourism.

The remainder of this paper is structured as follows: Section 2 shows related work; Section 3 describes the city classifier and the place identifier models; Section 4 describes the evaluation of the system; finally, Section 5 provides conclusions and suggests paths for future work.

2 Related work

There is a large body of works in the area of machine learning-based multimodal classification. This section summarises some of the most relevant studies in a wide range of domains.

In [5], the authors propose two improvements of existing models, VL-T5 and VL-BART, to label texts based on visual and textual inputs. A multimodal system that uses both text and visual content on Twitter to classify information during emergencies is proposed in [18]. To this end, the authors use LSTM and VGG-16 on texts and images from tweets, respectively. Along the same line, the proposal in [17] defined different models to detect disasters in cultural heritage from social media information focusing on images. Moving to a different domain, the work in [33] used the relationship between images and their associated texts in social media for sarcasm detection. The system relies on an early fusion of pre-trained text and image Transformer models. Finally, in [35] the authors proposed a common image-text embedding space for training a bidirectional network. The retrieval results provided on Flickr30K and MSCOCO datasets improved the state-of-the-art.

In the area of cybersecurity, the paper presented in [2] shows a 4-fold security system using fusion of facial recognition, retina pattern and fingerprint pattern along with the personal password, which is recognised using keystroke dynamics. Their goal was to reduce the probability of false acceptance rates and false rejection rates.

In the field of machine translation, there are multiple proposals based on multimodal inputs. In [37] the authors present a multi-modal self-attention model to solve the issues of noise when text and images are equally treated. The proposed method learns the representations of images based on the text, which avoids encoding irrelevant information in images. Experiments and visualisation analysis demonstrate that the model benefits from visual information and substantially outperforms previous works and competitive baselines using different metrics.

In the task of image captioning, the authors of [40] described a multimodal Transformer model used in machine translation for image captioning. Their approach consisted of simultaneously capturing intra- and inter-modal interactions in a unified attention block. The experimental results show that this method significantly outperforms the previous state-of-the-art approaches in this task. Building an ensemble of seven models, their solution ranked first place on the real-time leaderboard of the MSCOCO[1] image captioning challenge at the time of the writing of this paper.

There are also important contributions In the field of sentiment analysis [4, 15, 16], merging audio, video and image inputs, and also using Transformer models in the case of [15]. More natural language processing applications include the generation of dialogues combining video and text inputs [19], summary generation from audio and video [41], video retrieval [10], and fake news detection [31].

Finally, the work in [36] focuses on multimodal Transformer models. The authors provide a review of the applications and pre-training approaches of these models, together with a summary of common challenges and design decisions shared by these models and their applications.

With respect to multimodal classification techniques, in [25] the authors proposed a method for the integration of natural language understanding in image classification to improve the accuracy by making use of associated metadata. In [6], the authors proposed

[1] https://cocodataset.org/.

a novel deep learning-based multimodal fusion architecture for classification tasks that guarantees compatibility with any kind of learning models, dealing with cross-modal information and preventing performance degradation due to the partial absence of data. The work in [9] presented simple models that combine information from image and text to classify social media content. In this work, a pooling layer and an auxiliary learning task is used to learn a common feature space.

To the best of our knowledge, there is no multimodal approaches based on image and text to the task of geolocating social network content. Nevertheless, there are different works where this task has been addressed with other types of inputs, most of them based on text and/or metadata [1, 20]. In [1] the prediction is at coordinate level. In [20] the geolocation is at city or place level, which is also the goal in this work. A very interesting approach is presented in [43], providing a framework for geolocation identification in social networks based on different sources of information, including text, social relation, and contextual data. The proposal also includes a detailed review of contributions to this topic.

3 Description of the system

The proposed system is divided into two stages. The first one involves a city-level geolocation model, whereas the second one integrates the previous model with a place identifier to extend the evaluation and show potentially useful applications. The following paragraphs describes each of these stages in more detail.

3.1 City classifier

The first stage comprises a system that is able to locate which city a social media post belongs to based on the text and image it contains. The input of the system is the multimodal combination of textual and visual information from the post, and the output is the prediction of which city the post content refers to. The premise is that combining a visual and a textual model can improve their individual performance. Although there are situations in which the text may be more relevant than the image in the final decision (or vice versa), there are also situations where only the combination of both will give a correct result.

In the experiments carried out the problem was reduced to a multiclass classification task over 10 classes, corresponding to the 10 cities in the InstaCities1M dataset. As mentioned before, this dataset contains 100,000 pairs of associated images and texts from Instagram for each of the 10 most populated English speaking cities: London, New York, Sydney, Los Angeles, Chicago, Melbourne, Miami, Toronto, Singapore, and San Francisco.

This dataset was chosen for two main reasons. First, it contains recent social media data combining texts (description and hashtags) and their corresponding images. Secondly, each image is associated to one of the 10 cities mentioned above, which allows to evaluate the geolocation capabilities of the proposal presented in this section.

The textual information of the posts was processed using BERT [7], a contextual language model based on the Transformer architecture [34] that has become the state-of-the-art in many natural language processing tasks. The visual information was processed by RESNET18 [13], a pre-trained Convolutional Neural Network (CNN) that is able to identify 1,000 object categories in images. The output vectors of both models were processed using an ensemble to provide the final prediction of the system. Five different strategies for the ensemble were tested in the experiments carried out:

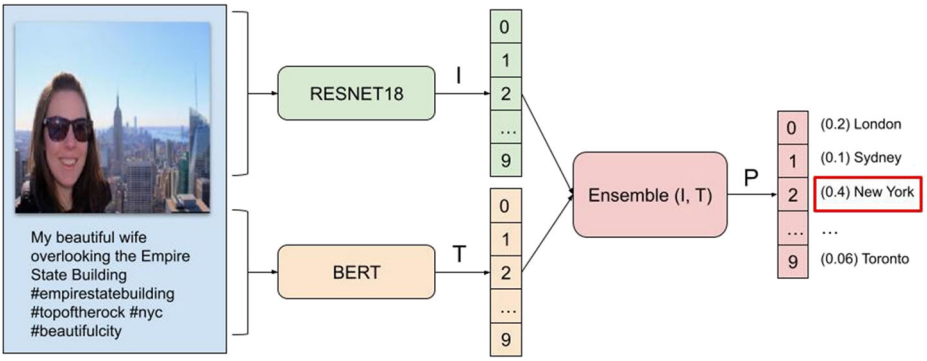

Fig. 1 Components of the city classification system proposed

- $P = T + I$: the final probability given by the system for each class (P) is the sum of the probability given by the textual model BERT (T) and the image model RESNET18 (I)
- $P = 2*T + I$: considering that textual information provided much better performance than the image information in the baseline experiments (see Section 4.1), in this case the prediction of BERT is given twice the importance of the image prediction
- $P = 0.8*T + 0.2*I$: the output of the text model is weighted 80% and the output of the image model 20%
- *L1 normalisation*: in this case, L1 normalisation is applied to the sum of the outputs of the text and image models
- *Top predicted*: the model (text or image) that provides the highest prediction value is chosen as the final result

Figure 1 shows the data flow in the system.

3.2 Place identifier

The second stage of the proposed system adds a place identifier to the city classifier described above. This new component is based on zero-shot deep learning techniques. The

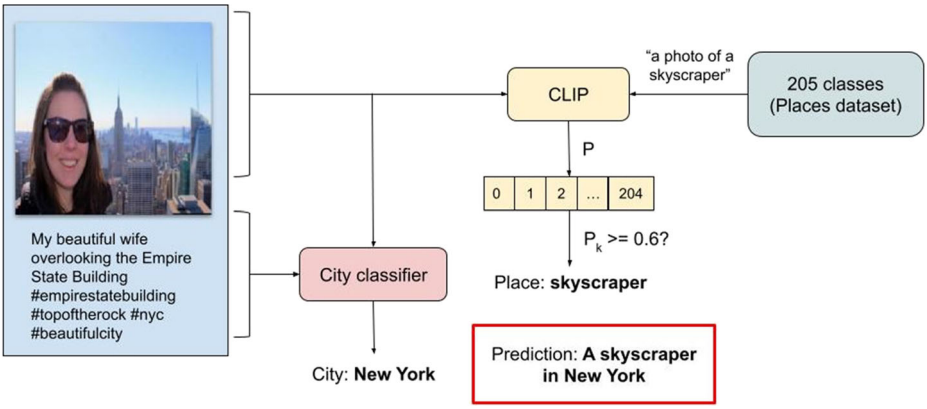

Fig. 2 Components of the final architecture combining the city classifier and the place identifier

Table 1 Accuracy obtained by BERT using only textual information depending on the anonymisation strategy

Anonymisation strategy	Accuracy
Original	0.7661
Categorical	0.6417
Removal	0.6344

possible types of places are defined in the set of labels from the Places dataset [42]. Examples of these classes are "skyscraper", "bridge", and "restaurant", involving both indoor and outdoor locations. In this way, the geolocation component defined in the first stage can be extended to also identify specific places in a city.

The Instacities1M dataset was used again in this stage to evaluate the identification of places. The classification procedure proposed leverages the transfer learning capabilities of the CLIP neural network, a state-of-the-art architecture for classification tasks in zero-shot setting. The complete procedure is described in [23]. Figure 2 shows the integration between the city classifier and the place identifier.

4 Evaluation

This section describes in detail the evaluation carried out on the city classifier and the place identifier, together with a deeper analysis of the performance of the whole system depending on specific cities and places.

4.1 City classifier

Different baselines were established using state-of-the-art models for text and image classification. The goal was to compare the performance of these individual models with respect to the multimodal ensemble used in the final architecture. The text component uses the pre-trained BERT model defined in the Hugginface library.[2] This model consists of a bidirectional Transformer pre-trained on a large corpus using a combination of masked linguistic modelling target and next-sentence prediction tasks. In the current experimental setup, the model was fine-tuned using an output linear layer that transforms the 768 outputs into 10, corresponding to the number of classes (cities) of the problem at hand. The hyperparameters during the training procedure included dropout of 0.3 [11], AdamW optimisation [38], and an early stopping regularisation with patience 3, so that the number of epochs vary depending on the performance over the validation subset.

Taking into account that the occurrence of city names in the text would make it trivial to identify to which city the post belongs, a study was carried out to evaluate the impact of this situation. To this end, three versions of BERT were considered proposing different anonymisation strategies: "Original" , "Categorical", and "Removal". "Original" indicates that the initial content was used and there was no anonymisation at all. "Categorical" implies that occurrences of the names of cities in text were replaced by a special tag `<CITY>`. Finally, "Removal" consists of directly deleting city names from text. The results of the three approaches on the Instacities1M dataset are shown in Table 1. The evaluation was carried out using 10% of the dataset as test (consisting of 100,000 samples evenly distributed among

[2] https://huggingface.co/bert-base-uncased.

Table 2 Accuracy obtained by RESNET18 and ViT models using only visual information

Model	Input size (pixels)	Epochs	Accuracy
RESNET18	256 x 256	7	0.2740
ViT	224 x 224	3	0.1147

the ten cities), whereas the remainder was used for training (80%) and validation (10%). As expected, the best result was obtained using the original content by a large margin, improving almost 20% the second best approach ("Categorical").

Regarding image classification, two different baselines were tested. The first one was RESNET18, a CNN deep learning architecture pre-trained on more than one million images from the ImageNet database [28]. This model encodes rich feature representations for a wide

Fig. 3 Examples of images (and expected class) misclassified by the visual models

range of images and is able to classify them into 1,000 object categories such as “keyboard”, “mouse”, “pencil”, or different breeds of animals. The image input size of the network is 224 x 224 pixels. A final linear layer was added to the original system to obtain 10 outputs corresponding to the 10 cities of the problem addressed. Finally, a fine-tuning process was performed including early stopping, dropout of 0.3, batch size of 32, adaptive learning rate [39] and Adam optimiser.

The second baseline model tested was ViT (Vistion Transformer) [8], a Transformer-based approach that has demonstrated to improve state-of-the-art CNN architectures while requiring substantially fewer computational resources during the training procedure. The hyperparameters during the training phase were the same used with RESNET18, also adding a final layer with 10 outputs.

Table 2 shows the results of these two models using only image information on Instacities1M with the same train, validation and test split used in the text-only experiments. The performance obtained with the visual content is significantly worse than those obtained using only text, with RESNET18 doubling the accuracy of ViT.

These results reveal that it is more difficult to extract relevant features that characterise cities from images than from texts. Figure 3 shows some images misclassified by the visual models.

In these examples, it is evident that models using only image information would hardly be able to make a correct prediction about them. The text corresponding to these images is shown in Table 3. There are references in the text that can help the model to make the correct prediction. For instance, image 2 shows clear references in the text to the region and country where the city to be predicted is located. On the contrary, image 1 is an example of misclassification. In this case, there is a reference to the country where the city is located, but

Table 3 Examples of texts not anonymised, expected classes, and predictions made by BERT (text) and RESNET18 (image) for the images shown in Fig. 3

	Expected class	Prediction using text	Prediction using image
Image 1	losangeles	newyork	singapore
#paradise #amazing #classy#chic #beautiful #wonderful#love #pretty #beautiful #nails#art #fashion #time #good#great #divine #sublime#shinny #skinny #cute #sweet#makeup #losangeles #usa#trendy #fancy			
Image 2	toronto	toronto	singapore
Our Ultra Rare Strawberry Las Vegas Bubba Kush! Grown by #sincityseeds#homeofthedank #slvbk #strawberrylasvegasbubbakush #medicalmarijuana #cannabiscommunity #cannabis #weed #stoners#blaze #marijuana #ganja#stoner #maryjane #alberta#quebec #canadianstoner#surrey #toronto #montreal#ontario #ottawa#britishcolumbia #calgary#canadian			
Image 3	newyork	toronto	chicago
Tired from today’s preparation and shooting but it was worth it. Shout out to my team member @kissmybootyque for the makeup, shout to @jazzi_juice out for modeling and being patient with the body painting and all, and really be shout to @bohemian.photo for pointers on lighting and a great studio experience. Pics coming soon Mua/ assistant: @kissmybootyque Model: @jazzi_juice #ART#artistikmind #fashion#fashion week #newyork#vogue #blackgirlmagic#superbowl #nymodel#beautiful #afropunk#fashionillustrator...			
Image 4	sanfrancisco	sanfrancisco	sydney
Happy Easter! Vegan for the animals, for the environment & for my health# House made chips tossed in a creamy salsa roja with cashew cream & cilantro. Mariposa bakery toast dipped in plantain batter & grilled topped with maple syrup, citrus cashew whipped cream, pecans & fresh fruit Tempeh chorizo, caramel onions, red peppers, roasted potatoes, black beans and cashew chipotle aioli served with smoked tomato salsa and ranchera. #vegan #instavegan#vegansofig #veganaf#vegansofinstagram#vegangirl #plants ...			

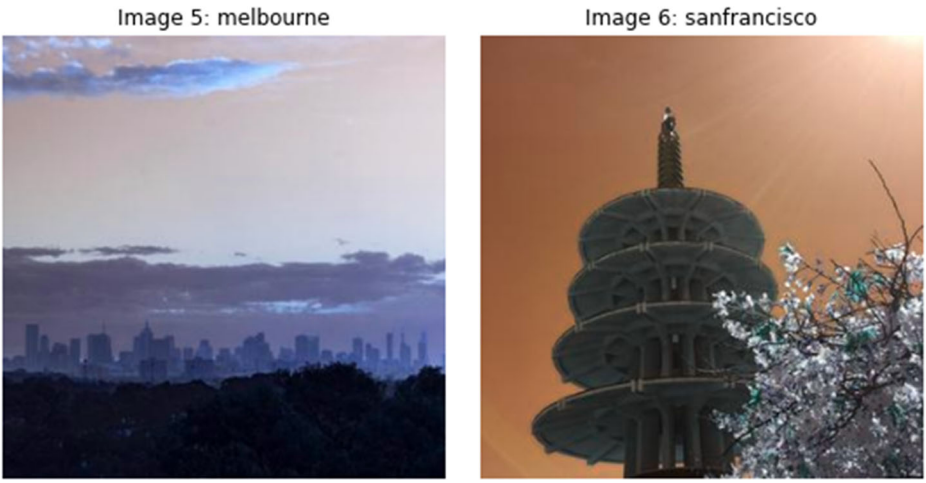

Fig. 4 Examples of images correctly classified by the RESNET18 image model

since there are more cities in the same country among the expected classes, the prediction ends up being incorrect.

Nevertheless, there are images that provide enough visual clues of the target city to be correctly classified by the image models. Figure 4 shows two examples of images where the target class was correctly identified by the image model.

After evaluating the image and text models individually, the next step consisted in testing the performance of the multimodal ensemble strategies proposed, where the outputs of BERT (text) and RESNET18 (image) were combined to provide the final prediction of the system.

Table 4 shows the results of all the ensembles on the InstaCities1M test subset ordered by accuracy, including the two individual baseline models. In the column *Top 1 accuracy* the result is considered to be correct only if it matches the highest probability class given

Table 4 Accuracy for the five ensemble strategies proposed and the text (BERT) and image (RESNET18) baselines

Ensemble	Top 1 accuracy	Top 3 accuracy
P = T + I	0.6639	0.8464
P = 2*T + I	0.6621	0.8455
P = 0.8*T + 0.2*I	0.6558	0.8417
L1 normalisation	0.6447	0.8339
Top predicted	0.6419	0.8293
BERT (text only)	0.6417	0.8294
RESNET18 (image only)	0.2740	0.5243

by the models. In *Top 3 accuracy* the result is correct if it matches any of the three classes with the highest probabilities.

All the multimodal ensembles proposed improved the accuracy obtained with the individual models. The best result was obtained by adding the prediction vectors of both models without normalisation ($P = T + I$). The second and third best results were obtained using the weighted sum of image and text predictions.

In order to check the robustness of the models with respect to variations in the input instances, the test subset was randomly split into 10 folds and the models were evaluated for each of them separately. The goal of this experiment was to calculate the standard deviation between folds, that is, to determine if the models performed equally disregarding the subset of input instances. Note that the average accuracy does not change with respect to using the entire test dataset.

Results show that the average standard deviation for accuracy of the proposed models was 0.0025 in *Top 1 accuracy* and 0.0016 in *Top 3 accuracy*. The maximum deviations were found in the *Top predicted* ensemble (0.0030) and *RESNET18* model (0.0038) for *Top 1 accuracy* and *Top 3 accuracy*, respectively. The minimum deviations were found in the ensemble $P = 0.8*T + 0.2*I$ (0.0020) for *Top 1 accuracy* and $P = 2*T + I$ (0.0008) for *Top 3 accuracy*. These low standard deviation values reveal that the models behave robustly, with very similar results regardless of the input subset.

Finally, a last analysis was carried out by generating a set of confusion matrices in order to better understand the geolocating performance of the best multimodal ensemble ($P = T + I$) with respect to the text-only and image-only models. The test subset of 100,000

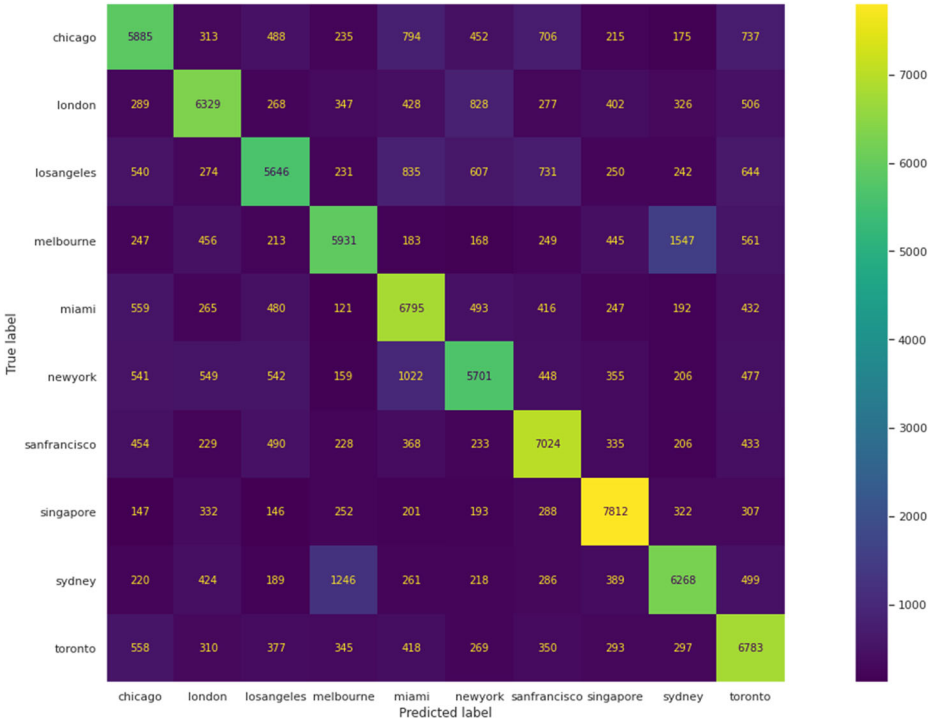

Fig. 5 Confusion matrix of the text-only model (BERT)

samples (10,000 for each city) of InstaCities1M was used in this analysis. Figure 5 shows the confusion matrix corresponding to the text-only model. It is interesting to highlight the number of times that Melbourne and Sydney are mistaken for each other (1,246 and 1,547 occurrences). There is also a significant number of errors when the model is classifying New York as Miami (1,022 occurrences). Figure 6 shows the confusion matrix of the image-only model. In this case, the error is generally higher with respect to the text-only version, as the accuracy values already suggested. The cities where the best accuracy was obtained are Singapore and Miami, but nevertheless the accuracy is less than 50% in both cases (4,346 and 4,010 out of 10,000 samples).

The last confusion matrix, shown in Fig. 7, corresponds to the results obtained using the best multimodal ensemble proposed. Results are similar to those obtained by the text-only model. The ensemble obtained better results in all cities except Sydney, where the text-only model correctly identified the class 82 times more (0.02% improvement). The aforementioned example of New York and Miami, which showed large confusion in the text-only model (1,022 cases), is still a problem for the ensemble with a small increase in the number of incorrect classifications (1,053 cases). In the case of Melbourne and Sydney, the ensemble provides a large improvement with respect to the text-only version in cases where Melbourne was mislabelled as Sydney, reducing the confusion from 1,547 to 1,138 (27% error reduction). On the contrary, the cases where Sydney was mislabelled as Melbourne increased from 1,246 to 1,713 (an increase of 37%).

In the light of these results, it can be concluded that the inclusion of image information does not have a significant impact on the average performance of the system, but it can

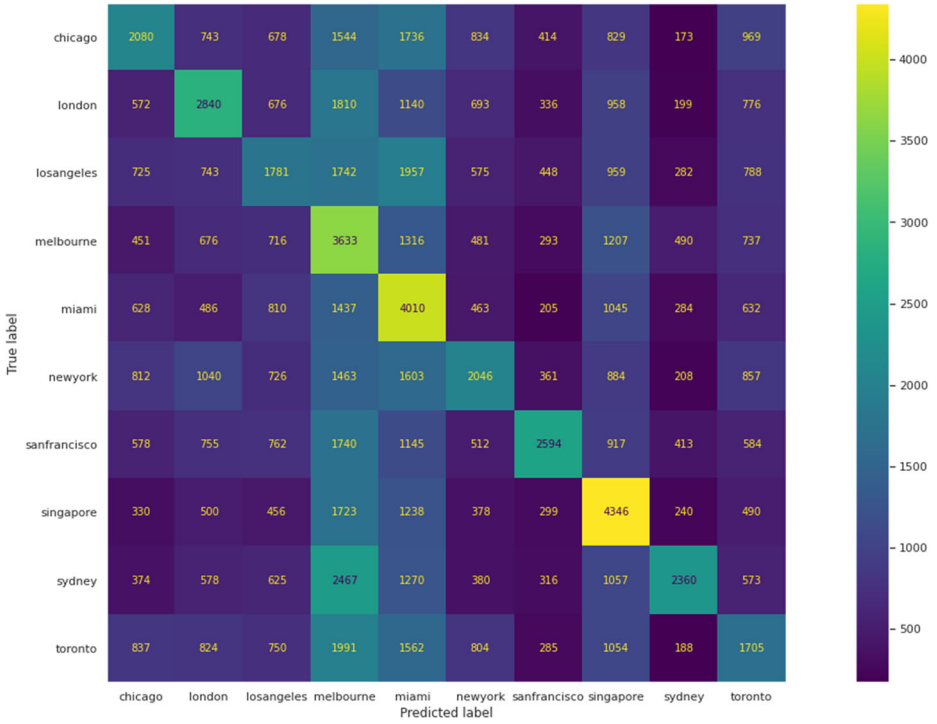

Fig. 6 Confusion matrix of the image-only model (RESNET18)

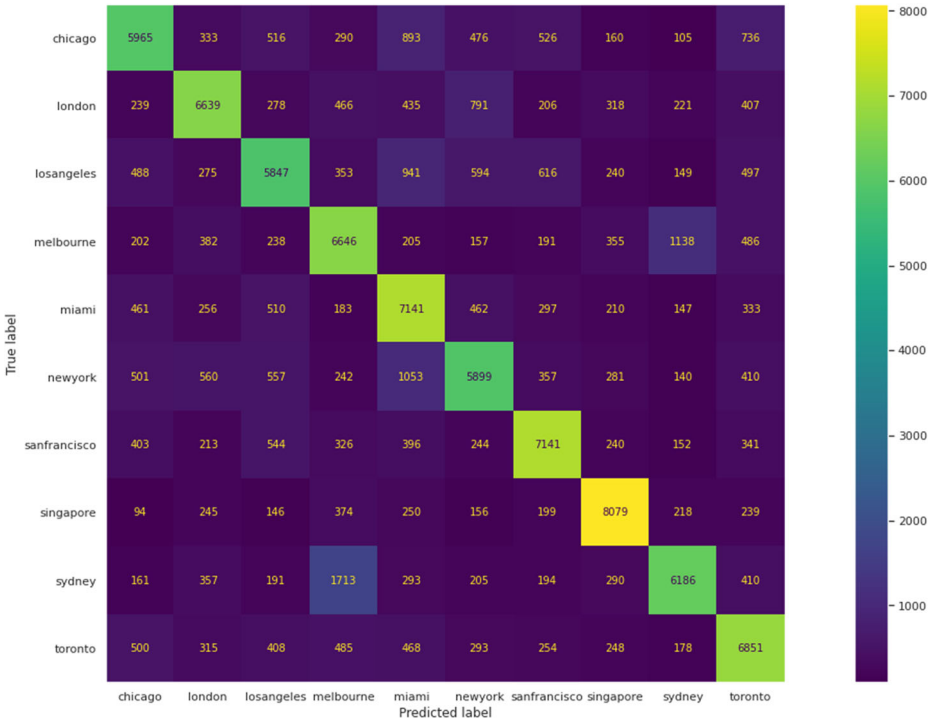

Fig. 7 Confusion matrix of the $P = T + I$ ensemble

noticeably change the performance of specific categories as seen in the case of Sydney and Melbourne. In this example, the number of samples labelled as Melbourne with respect to Sydney significantly increased. This leads to the idea that images from Melbourne have visual features that make the classifier better identify their origin and complement textual information, but also that Sydney images share these visual cues, which makes sense taking into account that these two cities are close, both geographically and culturally.

4.2 Place identifier

The test subset of 100,000 samples from Instacities1M was used again in this experiment to evaluate the accuracy of the place identifier. As mentioned before, this part of the system is based on CLIP and the set of 205 tags defined in the Places dataset. The procedure followed is the same described in [23]. In this work, the accuracy of the place identifier varied significantly depending on the type of place, being higher than 90% for outdoor places such as "plaza", "bridge" and "skyscraper", but lower than 20% for indoor places such as "shower", "hospital" or "beauty salon". Using the text associated with each image did not improve the accuracy of the classifier that used only images.

The output of the classifier is a set of probabilities (adding up to 1 by using a softmax layer) for each possible place. A place was considered to be identified in a post if the probability assigned by the classifier was over the 0.6 threshold established. Out of the 100,000 items analysed, the system was able to identify a place (i.e. to provide a provability over

Table 5 Five places with the most occurrences for each city in the InstaCities1M test subset, according to the city classifier and place identifier systems

City	Place 1	Place 2	Place 3	Place 4	Place 5
Chicago	beauty salon	art gallery	baseball stadium	skyscraper	martial arts gym
London	beauty salon	phone booth	coffee shop	shoe shop	art gallery
Los Angeles	beauty salon	martial arts gym	art gallery	coffee shop	ice cream parlor
Melbourne	coffee shop	beauty salon	martial arts gym	sky	ice cream parlor
Miami	plaza	beauty salon	martial arts gym	sandbar	boat deck
New York	skyscraper	beauty salon	bridge	plaza	art gallery
San Francisco	bridge	residential neighborhood	beauty salon	baseball stadium	sea cliff
Singapore	beauty salon	martial arts gym	coffee shop	outdoor of a swimming pool	east asia temple
Sydney	coffee shop	sea cliff	beauty salon	outdoor of a swimming pool	sky
Toronto	beauty salon	coffee shop	martial arts gym	skyscraper	ice cream parlor

0.6) in 10,384 cases. From this subset, the city classifier was able to correctly predict the location in 7,227 cases, achieving 0.6960 accuracy.

Table 5 shows the five places with the most occurrences for each city in the InstaCities1M test subset, according to the output of the city classifier and place identifier systems. Section 4.3 provides additional information on the accuracy of these predictions.

Figure 8 shows a heatmap with the ten most identified places in the whole test subset from InstaCities1M. Subsequently, the city classifier was used to locate these places in the ten cities of the dataset. It is worth noting that, for example, that the most common place identified by far in San Francisco is “bridge”.

4.3 Classification of specific places in cities

Finally, a last analysis was carried out to study how the city classifier behaves with respect to different types of places. Table 6 shows a list of the ten most common places found in the 100,000 samples test subset used before. The table includes the total number of samples for each place, the number of correctly classified samples, and the accuracy obtained. Results are shorted from highest to lowest accuracy values. The best performance was achieved on “bridge” (91.30%), followed by “baseball stadium” (90.91%), “indoor of a stage” (79.71%) and “skyscraper” (77.07%). The worst results were achieved for “coffee shop” (66.14%), “ice cream parlor” (62.70%) and “martial arts gym” (57.60%). It is interesting to note that the performance of the place identifier is significantly better in outdoor locations (such as a bridge, a stadium or a skyscraper) than indoor places (such as coffee shops or gyms). This is an expected result, since outdoor venues are often associated to points of interest and iconic touristic attractions in cities.

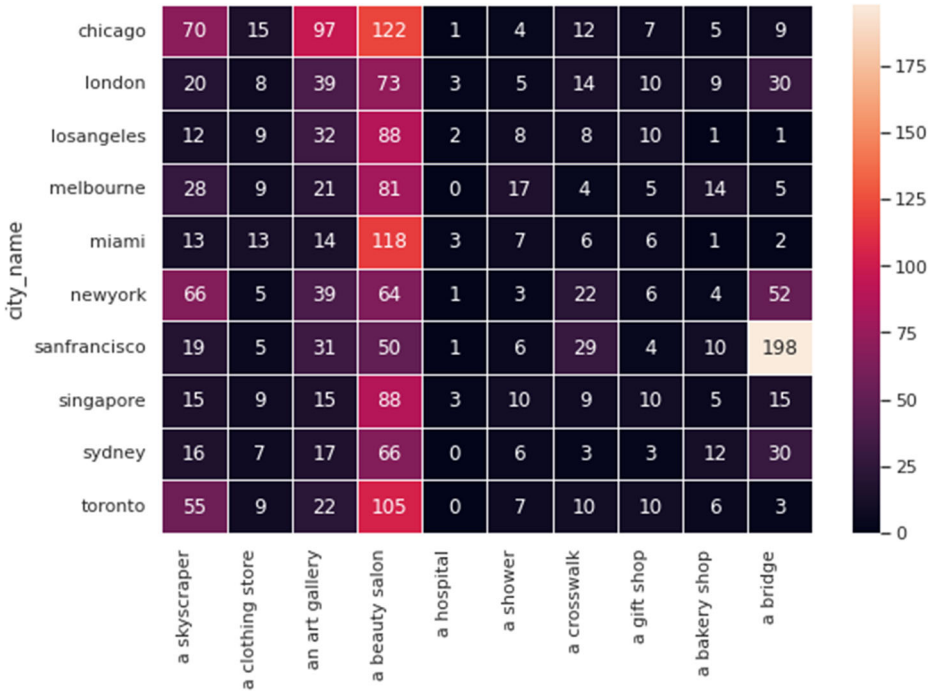

Fig. 8 Predicted number of occurrences of the ten most common places in the InstaCities1M test subset for each location

The previous table was focused on the ten most common places in the dataset. Tables 7 and 8 show respectively the best and worst accuracy achieved by the place identifier for the 205 possible categories in the Places dataset. As mentioned above, places corresponding to outdoor locations achieve the best results (94.29% for "market"), whereas indoor places perform significantly worse (52.78% for "locker room").

Table 6 Accuracy of the place identifier on the ten most common places in the test subset

Place	Total	Correct	Accuracy
Bridge	345	315	0.9130
Baseball stadium	176	160	0.9091
Indoor of a stage	207	165	0.7971
Skyscraper	314	242	0.7707
Art gallery	327	252	0.7706
Plaza	360	254	0.7056
Beauty salon	855	589	0.6889
Coffee shop	505	334	0.6614
Ice cream parlor	252	158	0.6270
Martial arts gym	467	269	0.5760

Table 7 Places for which the best accuracy was obtained by the place identifier component

Place	Total	Correct	Accuracy
Market	35	33	0.9429
Pagoda	16	15	0.9375
Baseball stadium	173	160	0.9249
Bridge	345	315	0.9130
Phone booth	98	89	0.9082
Formal garden	21	19	0.9048
Residential neighbourhood	164	147	0.8963
Train station platform	19	17	0.8947
Football stadium	62	55	0.8871
Driveway	17	15	0.8824

Figure 9 shows a heatmap in the line of that provided in Fig. 8 (which included the frequencies of places by city) but this time displaying the accuracy obtained instead of the frequency. In this heatmap, places with 10 or less occurrences for a city were removed as these results are not considered as significant due to the small number of samples involved. Such is the case of "gift shop" or "hospital", whose occurrences for any city range from 0 to 3 and 3 to 10 respectively.

The most relevant aspect of this heatmap is that it shows specific places and cities where high accuracy is achieved, while there are others where it is not. For example, "bridge" in London, New York and San Francisco is correctly geolocated in more than 90% of the cases, reaching up to 99% in the case of San Francisco. On the other hand, the worst accuracy was obtained on average for "beauty salon", probably for being an indoor location, which adds extra difficulty to the classification process. The exception in this particular case is Singapore, with 91% success rate.

Table 8 Places for which the worst accuracy was obtained by the place identifier component

Place	Total	Correct	Accuracy
Hotel room	45	25	0.5556
Dam	27	15	0.5556
Fairway	20	11	0.5500
Reception	102	56	0.5490
Dinette home	188	103	0.5479
Galley	22	12	0.5455
Home office	24	13	0.5417
Hot spring	13	7	0.5385
Bowling alley	13	7	0.5385
Locker room	36	19	0.5278

Fig. 9 Accuracy obtained depending on the place identified and the city predicted

5 Conclusions and future work

This work presented an approach to geolocate multimodal (image and text) information from social networks. A set of baselines were evaluated using only textual and image information. The experiments revealed that the text-only model outperformed significantly the image-only model.

These baselines were then combined following five different ensemble strategies to leverage the information of the individual models. The multimodal ensemble models improved in all cases the results obtained with the individual (text-only and image-only) models. In the experiments carried out, the accuracy of the best multimodal ensemble improved 2% the results obtained using the only-text model.

The geolocation model was also combined with a place identifier to provide a more fine-grained system. The results revealed that the model worked better with outdoor places such as bridges, skyscrapers or stadiums, than with indoor places such as coffee shops or hotel rooms.

Although the ensemble models demonstrated to outperform the baselines, there is room for improvement by proposing additional strategies to combine text and image information. For instance, it is necessary to exploit the conditional ensemble part by balancing the weight given to text and image inputs. Other alternatives should also be explored, such as replacing the ensemble with a mixed model that could be trained in conjunction with the text-only and image-only models.

Another interesting path to explore is the inclusion of other state-of-the-art models for text and image such as RoBERTa [22] and Capsule Neural Networks [29], which have demonstrated high performance in different classification tasks. Another path that deserves attention is to explore the use of multimodal neural network models such as VisualBERT [21] and LXMERT [32], and their adaptation to multiclass classification. These models were not tested in this work since they were designed for other tasks such as image question answering.

Also a promising avenue of research is the integration of this system with a sentiment analysis model on the textual part in order of catalogue feelings about types of places in specific cities, which could be useful for the tourism sector [24]. Sentiment analysis in noisy datasets, as is the case in social networks, can be improved by using text processing algorithms such as those described in [26]. In this way, geolocated touristic places in a region or city could not only be detected, but it would be possible also to obtain as complementary data the sentiments of people about these places and compare opinions between different places and cities.

Finally, it is important to note that trustworthy artificial intelligence will become one of the most important research topics in this area in the future [14]. Regarding the machine learning models used in this work, it is often difficult to make sense of outcomes calculated by them (black box effect). Moreover, these models can suffer from stability issues, since changes in the input data or in the initialisation may impact the final results. For these reasons, explainability and robustness are two main topics that must be addressed in the future in order to promote reliability and trust in artificial intelligence solutions.

Supplementary information The source code created is available at: https://github.com/matrox1000/geolocation/.

Acknowledgements This research has been partially funded by project "Desarrollo de un ecosistema de datos abiertos para transformar el sector turístico" (GVA-COVID19/2021/103) funded by Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital de la Generalitat Valenciana, "A way of making Europe" European Regional Development Fund (ERDF) and MCIN/AEI/10.13039/501100011033 for supporting this work under the "CHAN-TWIN" project (grant TED2021-130890B-C21) and the HORIZON-MSCA-2021-SE-0 action number: 101086387, REMARKABLE, Rural Environmental Monitoring via ultra wide-ARea networKs And distriButed federated Learning. We also would like to thank Nvidia for their generous hardware donations that made these experiments possible.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability The datasets analysed during the current study are available at: https://gombru.github.io/2018/08/01/InstaCities1M/.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Afyouni I, Aghbari ZA, Razack RA (2022) Multi-feature, multi-modal, and multi-source social event detection: a comprehensive survey. Inf Fusion 79(2021):279–308. https://doi.org/10.1016/j.inffus.2021.10.013
2. Arora G, Pavani PL, Kohli R, Bibhu V (2016) Multimodal biometrics for improvised security. 2016 1st Int Conf Innovation Challenges in Cyber Secur, ICICCS 2016 (Iciccs):1–5. https://doi.org/10.ICICCS.2016.7542312
3. Chang M-W, Ratinov L, Roth D, Srikumar V (2008) Importance of semantic representation: dataless classification. In: Proceedings of the 23rd national conference on artificial intelligence - vol 2. AAAI'08. AAAI press, pp 830–835
4. Cheng J, Fostiropoulos I, Boehm B, Soleymani M (2021) Multimodal phased transformer for sentiment analysis. EMNLP 2021 - 2021 conference on empirical methods in natural language processing, proceedings, pp 2447–2458. https://doi.org/10.18653/v1/2021.emnlp-main.189
5. Cho J, Lei J, Tan H, Bansal M (2021) Unifying vision-and-language tasks via text generation. arXiv:2102.02779
6. Choi JH, Lee JS (2019) Embracenet: a robust deep learning architecture for multimodal classification. Inf Fusion 51(2018):259–270. arXiv:1904.09078. https://doi.org/10.1016/j.inffus.2019.02.010
7. Devlin J, Chang M, Lee K, Toutanova K (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, NAACL-HLT. Association for computational linguistics, pp 4171–4186. https://doi.org/10.18653/v1/n19-1423
8. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929
9. Duong CT, Lebret R, Aberer K (2017) Multimodal classification for analysing social media. arXiv:1708.02099
10. Dzabraev M, Kalashnikov M, Komkov S, Petiushko A (2021) MDMMT: multidomain multimodal transformer for video retrieval. IEEE Comput Society Conf Comput Vis Pattern Recognit Workshops:3349–3358. https://doi.org/10.1109/CVPRW53098.2021.00374
11. Fan A, Grave E, Joulin A (2019) Reducing transformer depth on demand with structured dropout, vol 103, pp 1–15. arXiv:1909.11556
12. Gomez R, Gomez L, Gibert J, Karatzas D (2019) Learning to learn from web data through deep semantic embeddings. Lect Notes Comput Sci (including subseries Lecture Notes Artif Intell Lecture Notes in Bioinformatics) 11134 LNCS:514–529. arXiv:1808.06368. https://doi.org/10.1007/978-3-030-11024-6_40
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Comput Society Conf Comput Vis Pattern Recognit:770–778. https://doi.org/10.1109/CVPR.2016.90
14. Holzinger A (2021) The next frontier: AI we can really trust. In: Machine learning and principles and practice of knowledge discovery in databases. Springer, pp 427–440. https://doi.org/10.1007/978-3-030-93736-2_33
15. Huang J, Tao J, Liu B, Lian Z, Niu M (2020) Multimodal transformer fusion for continuous emotion recognition. In: ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3507–3511. https://doi.org/10.1109/ICASSP40776.2020.9053762
16. Jaegle A, Gimeno F, Brock A, Vinyals O, Zisserman A, Carreira J (2021) Perceiver: general perception with iterative attention. In: Proceedings of the 38th international conference on machine learning. Proceedings of machine learning research. PMLR, vol 139, pp 4651–4664
17. Kumar P, Ofli F, Imran M, Castillo C (2020) Detection of disaster-affected cultural heritage sites from social media images using deep learning techniques. J Comput Cultural Heritage, vol 13(3). https://doi.org/10.1145/3383314
18. Kumar A, Singh JP, Dwivedi YK, Rana NP (2020) A deep multi-modal neural network for informative twitter content classification during emergencies. Annals Oper Res:(0123456789). https://doi.org/10.1007/s10479-020-03514-x
19. Li Z, Li Z, Zhang J, Feng Y, Zhou J (2021) Bridging text and video: a universal multimodal transformer for audio-visual scene-aware dialog. IEEE/ACM Trans Audio Speech Language Process 29:2476–2483. https://doi.org/10.1109/TASLP.2021.3065823
20. Li P, Lu H, Kanhabua N, Zhao S, Pan G (2019) Location inference for non-geotagged tweets in user timelines [Extended Abstract]. Proc Int Conf Data Eng 2019-April(6):2111–2112. https://doi.org/10.1109/ICDE.2019.00250

21. Li LH, Yatskar M, Yin D, Hsieh C-J, Chang K-W (2019) VisualBERT: a simple and performant baseline for vision and language:(2), pp 1–14. arXiv:1908.03557
22. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach, pp 1–13, coRR arXiv:1907.11692. https://doi.org/10.48550
23. Lucas L, Tomás D, Garcia-Rodriguez J (2022) Exploiting the relationship between visual and textual features in social networks for image classification with zero-shot deep learning. In: Sanjurjo gonzález H, Pastor López I, García Bringas P, Quintián H, Corchado E (eds) 16th International conference on soft computing models in industrial and environmental applications (SOCO 2021). Springer, pp 369–378
24. Lucas L, Tomás D, Garcia-Rodriguez J (2022) Sentiment analysis and image classification in social networks with zero-shot deep learning: applications in tourism. In: Sanjurjo gonzález H, Pastor López I, García Bringas P, Quintián H, Corchado E (eds) 16th International conference on soft computing models in industrial and environmental applications (SOCO 2021). Springer, pp 419–428
25. Miller SJ, Howard J, Adams P, Schwan M, Slater R, Miller S, Howard J, Adams P, Schwan M, Slater R (2020) SMU data science review multi-modal classification using images and text multi-modal classification using images and text, vol 3(3)
26. Petz G, Karpowicz M, Fürschuß H, Auinger A, Stříteský V, Holzinger A (2015) Reprint of: computational approaches for mining user's opinions on the web 2.0. Inf Process Manag 51(4):510–519. https://doi.org/10.1016/j.ipm.2014.07.011
27. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision. In: Proceedings of the 38th international conference on machine learning. Proceedings of machine learning research. PMLR, vol 139, pp 8748–8763. Accessed Dec 2022. https://proceedings.mlr.press/v139/radford21a.html
28. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252. arXiv:1409.0575. https://doi.org/10.1007/s11263-015-0816-y
29. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Advances in neural information processing systems, vol 30, pp 3856-3866. Curran Associates, Inc., USA
30. Saquete E, Tomás D, Moreda P, Martínez-Barco P, Palomar M (2020) Fighting post-truth using natural language processing: a review and open challenges. Expert Syst Appl 141:112943
31. Singh B, Sharma DK (2022) Predicting image credibility in fake news over social media using multi-modal approach. Neural Comput Appl 34(24):21503–21517. https://doi.org/10.1007/s00521-021-06086-4
32. Tan H, Bansal M (2019) LXMErt: learning cross-modality encoder representations from transformers. In: EMNLP-IJCNLP 2019 - 2019 conference on empirical methods in natural language processing and 9th international joint conference on natural language processing, proceedings of the conference, pp 5100–5111. arXiv:1908.07490. https://doi.org/10.18653/v1/d19-1514
33. Tomás D, Ortega-Bueno R, Zhang G, Rosso P, Schifanella R (2022) Transformer-based models for multimodal irony detection. J Ambient Intell Humanized Comput:1–12. https://doi.org/10.1007/s12652-022-04447-y
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30, pp 5998-6008. Curran Associates, Inc., USA
35. Wang L, Li Y, Lazebnik S (2015) Learning deep structure-preserving image-text embeddings. arXiv:1511.06078. https://doi.org/10.48550
36. Xu P, Zhu X, Clifton DA (2022) Multimodal learning with transformers: a Survey:1–23. arXiv:2206.06488
37. Yao S, Wan X (2020) Multimodal transformer for multimodal machine translation, pp 4346–4350. https://doi.org/10.18653/v1/2020.acl-main.400
38. You Y, Li J, Reddi S, Hseu J, Kumar S, Bhojanapalli S, Song X, Demmel J, Keutzer K, Hsieh C-J (2019) Large batch optimization for deep learning: training BERT in 76 minutes. arXiv:1904.00962
39. You K, Long M, Wang J, Jordan MI (2019) How does learning rate decay help modern neural networks. arXiv:1908.01878
40. Yu J, Li J, Yu Z, Huang Q (2020) Multimodal transformer with Multi-View visual representation for image captioning. IEEE Trans Circuits Syst Video Technol 30(12):4467–4480. https://doi.org/10.1109/TCSVT.2019.2947482
41. Zhao B, Gong M, Li X (2022) Hierarchical multimodal transformer to summarize videos. Neurocomputing 468:360–369. https://doi.org/10.1016/j.neucom.2021.10.039

42. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database - supplementary materials. NIPS'14 Proc 27th Int Conf Neural Inf Process Syst 1:487–495
43. Zhou F, Qi X, Zhang K, Trajcevski G, Zhong T (2022) Metageo: a general framework for social user geolocation identification with few-shot learning. IEEE Trans Neural Netw Learn Syst 1:1–15. https://doi.org/10.1109/TNNLS.2022.3154204