MM-Locate-News: Multimodal Focus Location Estimation in News

Golsa Tahmasebzadeh\textsuperscript{1,2}, Eric Müller-Budack\textsuperscript{1,2}, Sherzod Hakimov\textsuperscript{3}, and Ralph Ewerth\textsuperscript{1,2}

\textsuperscript{1} TIB–Leibniz Information Centre for Science and Technology, Hannover, Germany
\textsuperscript{2} L3S Research Center, Leibniz University Hannover, Germany
\textsuperscript{3} University of Potsdam, Germany
{golsa.tahmasebzadeh, eric.mueller, ralph.ewerth}@tib.eu,
sherzod.hakimov@uni-potsdam.de

Abstract. The consumption of news has changed significantly as the Web has become the most influential medium for information. To analyze and contextualize the large amount of news published every day, the geographic focus of an article is an important aspect in order to enable content-based news retrieval. There are methods and datasets for geolocation estimation from text or photos, but they are typically considered as separate tasks. However, the photo might lack geographical cues and text can include multiple locations, making it challenging to recognize the focus location using a single modality. In this paper, a novel dataset called Multimodal Focus Location of News (MM-Locate-News) is introduced. We evaluate state-of-the-art methods on the new benchmark dataset and suggest novel models to predict the focus location of news using both textual and image content. The experimental results show that the multimodal model outperforms unimodal models.

Keywords: multimodal geolocation estimation, news analytics, computer vision, natural language processing

1 Introduction

With the rapidly growing amount of information on the Web and social media, news articles are increasingly conveyed in multimodal form; apart from text, images and videos have become inseparable from news to report on events. Besides, there is an increasing research interest in the geographical content of news in fields like information retrieval or Web science. Although a news article might mention many locations it is important to determine the location of the main story. The concept of focus location is a piece of metadata which indicates the geographic focus of a news article \cite{1}. The focus locations are vital, for instance, to identify the location of a natural disaster for crisis management or to analyse political or social events across the world. Even if the news article focuses on one particular event, the text usually mentions various locations which are not the focus locations (see Figure 1a). On the other hand, a news image alone usually
Colombians shock government, rejecting peace deal. President Juan Manuel Santos admitted Jimenez, vowed his side too. He said in a speech in Havana, Cuba.

Fig. 1: Samples from the MM-Locate-News dataset. (a) The proposed visual model fails due to lack of visual geo-representative clues. (b) The proposed textual model fails due to various locations in the text. In both cases, the combination of visual and textual features helps specify the correct location. Images are replaced with similar ones due to license restrictions.

Most of previous approaches for geolocation estimation depend on either text or image. Text-based approaches assign geo-coordinates to the focus location at the document level or based on specific events. On the other hand, the majority of image-based methods focus on locating photos from certain environments such as cities. In recent years, several deep learning approaches have been introduced for geolocation estimation at global scale without any prior assumptions. However, information from image and text are required for a more robust focus location estimation in multimodal news, as the examples in Figure illustrate. But, so far, only a few methods have considered multiple modalities for geolocation estimation. Overall, these methods suffer from two major limitations: (1) They do not make use of recent multimodal transformer models or other unimodal state-of-the-art approaches to extract rich textual and visual information. (2) Although there are various large-scale datasets for text-based geolocation estimation of news at document level or based on events, they do not provide the accompanying images and thus do not allow for multimodal focus location.

Some multimodal datasets, such as the MediaEval Placing Task benchmark datasets or Multiple Languages and Modalities (MLM), contain image-text pairs with the corresponding geo-coordinates, but do not cover news articles and consequently do not address the related challenges such as multiple entities in the text or the lack of geo-representative features in the images. To the best of our knowledge, BreakingNews is the biggest dataset of multimodal news articles with corresponding location information. However, the locations are primarily extracted based on the Rich Site Summary (RSS) and, if not available, on heuristics such as the publisher location or story text. As a result, each document is provided with a number of potential locations that do not necessarily correspond to the focus location, are inaccurate, or even
wrong in some cases. In summary, there is a clear need for multimodal datasets of news articles with image and text and high-quality focus location labels. In addition, a multimodal approach is required which benefits from state-of-the-art approaches to provide rich visual and textual features for news documents.

In this paper, we address the task of focus location estimation; specifically, we target the gap between lack of visual geo-representative clues in news images and multiple location mentions in the body text of news. Our contributions are summarized as follows. (1) We present the MM-Locate-News (Multimodal Focus Location of News) dataset that consists of 6395 news articles covering 237 cities and 152 countries across all continents as well as multiple domains such as health, environment, and politics. The dataset is collected in a weakly-supervised manner, and multiple data cleaning steps are applied to remove articles with potential inaccurate geolocation information. The acquired dataset addresses drawbacks of other datasets such as BreakingNews [27] as it considers multimodal content of news to label the corresponding location. Furthermore, we provide a test set of 591 manually annotated news articles. (2) We propose several neural network models that, unlike previous work [17,18,4,27,29], exploit state-of-the-art image-based [24,34], text-based [6], and multimodal approaches [26] to extract embeddings for multimodal geolocation estimation of news. (3) Experimental results and comparisons to existing solutions on the proposed dataset and on BreakingNews [27] demonstrate the feasibility of the proposed approach.

The remainder of the paper is structured as follows. Section 2 describes the dataset acquisition, while the proposed model for multimodal geolocation estimation is presented in Section 3. We discuss the experimental results in Section 4. Section 5 concludes the paper and outlines potential directions for future work.

2 MM-Locate-News Dataset

In this section, we present our novel dataset called Multimodal Focus Location of News (MM-Locate-News). As discussed in Section 1, existing datasets for geolocation estimation are either not related to news [21,31,32], do not provide images along with news texts [22,29], or contain unreliable labels for locations extracted from the news feed [27]. In contrast, MM-Locate-News includes image-text pairs of news and the location focus at different geospatial resolutions, such as city, country, and continent. In the sequel, data collection and cleaning steps (Figure 2) as well as annotation process and dataset statistics are presented.

**Data Collection:** The dataset has been collected in a weakly-supervised fashion. To cover a variety of locations from all six continents, we extract all countries, capitals, highly populated cities (minimum population of 500,000; minimum area of 200 km²), and medium populated cities (population between 20,000 and 500,000; area between 100 km² and 200 km²) from Wikidata. For each location, we query EventRegistry for events between 2016 and 2020 from the following categories: sports, business, environment, society, health, and politics.

---

4 Source code: https://github.com/TIBHannover/mm-locate-news
5 http://eventregistry.org/
Note that EventRegistry automatically clusters news articles reporting on the same (or similar) events and that the news title of the cluster centroid represents the event name. To ensure the quality, we filter out events that do not include the location in their name or when their category relevance and query relevance scores, provided by EventRegistry, are below the average scores of all events per query location. The intuition behind this step is that an event with a location mentioned in its name more likely provides news articles focusing on the queried location. Finally, we extract all news articles from the remaining event clusters.

Data Filtering: We apply the following steps to remove irrelevant samples.

1) Named Entity – Query Location Match: We assume that an article is related to a query location if it is geographically close to at least one named entity. Following related work [25], we extract the named entities using spaCy [13] and use Wikifier [3] to link them to Wikidata for disambiguation. We extract coordinate location (Wikidata Property P625), which is available primarily for locations (e.g., landmarks, cities, or countries). For persons, we extract the place of birth (Wikidata Property P19) as they likely act in the respective country (or even city). We compute the Great Circle Distance (GCD) between the geographical coordinates of the query location and the extracted entity locations. We keep news articles that include at least one named entity whose GCD from the query location is smaller than $k\sqrt{a}$ where $a$ is the area (Wikidata Property P2046) of the query location, and $k$ is a hyperparameter as defined in Section 4.

2) Event – News Article Distance: Each news article in EventRegistry is assigned a similarity measure that represents the closeness to an event. We discard articles with a lower similarity than the average similarity of all articles of the same cluster to keep the news articles that are most related to the respective event.

3) Redundancy Removal: We compute the similarity between news articles using TF-IDF vectors (Term Frequency; Inverse Document Frequency) and discard one of the articles when the similarity is higher than 0.5 to remove redundancy.

4) Filtering of Rare Locations: After applying the filtering step 1-3, we remove rare locations (and corresponding articles) with less than five articles as they contain too few articles for training.

Dataset Statistics: In total, we queried 853 locations and extracted 13,143 news articles. After the data cleaning steps, we end up with 6395 news articles for 389 locations (237 cities and 152 countries). We divided the MM-Locate-News dataset into train, validation and test data splits by equally distributing news...
Table 1: Distribution of train, validation and test samples in MM-Locate-News among continents (AF: Africa, SA: South America, EU: Europe, AS: Asia, NA: North America, OC: Oceania)

|       | AF  | SA  | EU  | AS  | NA  | OC  | Total |
|-------|-----|-----|-----|-----|-----|-----|-------|
| Train | 854 | 216 | 1604| 1842| 589 | 161 | 5266  |
| Val   | 93  | 24  | 147 | 160 | 88  | 23  | 535   |
| Test  | 84  | 29  | 179 | 202 | 71  | 26  | 591   |

Table 2: Manual annotation criteria (C) for the MM-Locate-News test set variants (T). Answers “yes”, “no” or “unsure” are denoted as “✓”, while “u” and “✓” denote “unsure” and “yes”.

|       | T1  | T2  | T3  |
|-------|-----|-----|-----|
| C1:   | ✓   | ✓   | u   |
| C2:   | ✓   | ✓   | ✓   |
| C3:   | ✓   | ✓   | ✓   |
| Number of samples | 591 | 65 | 154 |

3 Multimodal Focus Location Estimation

We define focus location estimation as a classification problem where for each article, i.e., image-text pair, the total number of $n$ query locations (country or city) are considered as target classes. The $n$-dimensional one-hot encoded ground truth vector $y = (y_1, y_2, \ldots, y_n) \in \{0, 1\}^n$ represents the query location. We extract textual and visual features as follows.

Visual Features: The visual Scene descriptor (representing a place in a general sense) is based on ResNet-152 model [12] to recognize 365 places (pre-trained on the Places365 dataset [34]). The Location descriptor, $base(M, f^*)$, is taken from a state-of-the-art photo geolocation estimation approach [24]. The Object descriptor utilizes the ResNet-152 model [12] pre-trained on the Image-
Net dataset \[5\]. The CLIP descriptor is extracted using CLIP (Contrastive Language-Image Pretraining) \[26\] image encoder.

**Textual Features**: A pre-trained BERT (Bidirectional Encoder Representations from Transformers) \[6\] model is used to extract two distinct feature vectors. The first one represents the whole body text (BERT-Body), while the other refers to mentions of named entities and averages the individual word embeddings of all named entities (BERT-Entities).

**Multimodal Architecture**: As shown in Figure 3, the textual and visual embeddings are concatenated and passed to the fully connected (FC) layers with an output size of 1024, followed by a ReLU layer. Next, the outputs are fed to another FC layer followed by a \(\tanh(u_i)\) layer. The norm function with clamp \(\min = 10^{-12}\) is applied to extract the visual \(\hat{y}_v\) and textual vector \(\hat{y}_t\) with size of \(n = 389\) (number of locations). We obtain the multimodal output \(\hat{y}_m\) using the maximum probabilities of the visual \(\hat{y}_v\) and textual \(\hat{y}_t\) outputs.

**Network Optimization**: During training, the models are optimized as follows. For unimodal models, the cross-entropy loss between the ground truth one-hot encoded vector \(y\) to one (\(\hat{y}_v\) or \(\hat{y}_t\)) is optimized. For multimodal models, the same procedure is applied, but on both vectors \(\hat{y}_v\) and \(\hat{y}_t\). We freeze the weights of the visual models during training.

4 Experimental Setup and Results

In this section, the evaluation setup (Section 4.1) and experimental results including a comparison with state-of-the-art approaches on the MM-Locate-News (Section 4.2) and BreakingNews \[27\] (Section 4.3) datasets are reported.

4.1 Evaluation Setup

**Evaluation Metrics**: We use the geographical location of the class with highest probability in the network output (\(\hat{y}_v\), \(\hat{y}_t\), or \(\hat{y}_m\)) as the prediction and measure the Great Circle Distance (GCD) to the ground truth location. As suggested by Hays and Efros \[11\], the models are evaluated with respect to the percentage of samples predicted within different accuracy levels: city, region, country, and continent with maximum tolerable GCD to the ground location of 25, 200, 750, and
Table 3: Accuracy [%] of focus location estimates for Cliff-clavin [7], Mordecai [10], ISNs [24] and baselines on different test variants of MM-Locate-News (best results for each modality is highlighted per GCD accuracy levels). Text features: BERT-Body (B-Bd), BERT-Entities (B-Et). Visual Features: Location (Lo), Object (Ob), Scene (Sc), CLIP. GCD accuracy levels: City (CI, max. 25 km GCD to the ground truth location), Region (RE, max. 200 km), Country (CR, max. 750 km), Continent (CT, max. 2500 km)

| Approach                  | Modality | CI   | RE  | CR  | CT  | CI   | RE  | CR  | CT  | CI   | RE  | CR  | CT  |
|---------------------------|----------|------|-----|-----|-----|------|-----|-----|-----|------|-----|-----|-----|
| Mordecai [10] (only country-level) | Textual  | -    | -   | -   | -   | -    | -   | -   | -   | -    | -   | -   | -   |
| Cliff-clavin [7]          | Textual  | 36.9 | 53.1| 71.2| 86.5| 38.5 | 56.9| 66.2| 87.7| 33.1 | 48.7| 72.7| 85.1|
| B-Bd                      | Textual  | 21.8 | 27.4| 41.1| 68.4| 33.8 | 35.4| 49.2| 70.8| 19.5 | 23.4| 40.9| 63.6|
| B-Et                      | Textual  | 48.1 | 53.5| 66.3| 79.5| 58.5 | 64.6| 75.4| 81.5| 42.9 | 46.8| 61.0| 77.9|
| B-Bd + B-Et               | Textual  | 42.5 | 47.9| 60.4| 78.5| 60.0 | 64.6| 73.8| 89.2| 37.0 | 41.6| 53.2| 71.4|
| ISNs [24]                 | Visual   | 2.5  | 4.4 | 12.0| 31.0| 16.9 | 26.2| 40.0| 55.4| 0.6  | 1.9 | 7.8 | 29.9|
| CLIP                      | Visual   | 4.6  | 6.3 | 15.4| 41.1| 4.6  | 4.6 | 13.8| 41.5| 5.2  | 8.4 | 19.5| 42.2|
| Lo + Ob                   | Visual   | 3.2  | 3.7 | 8.5 | 27.7| 10.8 | 10.8| 13.8| 27.7| 3.9  | 4.5 | 9.7 | 27.3|
| Sc + Ob                   | Visual   | 1.2  | 1.5 | 6.4 | 20.6| 3.1  | 4.6 | 9.2 | 21.5| 1.9  | 2.6 | 9.1 | 26.0|
| Lo + Sc                   | Visual   | 2.4  | 3.0 | 9.0 | 27.1| 6.2  | 6.2 | 13.3| 33.8| 1.9  | 3.2 | 9.1 | 26.0|
| Lo + Sc + Ob              | Visual   | 2.5  | 3.4 | 9.1 | 31.0| 7.7  | 7.7 | 13.8| 30.8| 3.2  | 5.2 | 11.7| 33.8|
| Lo + CLIP                 | Visual   | 3.7  | 5.6 | 12.9| 36.9| 4.6  | 7.7 | 13.8| 41.5| 5.2  | 7.1 | 16.2| 37.7|
| Sc + CLIP                 | Visual   | 3.9  | 5.6 | 12.7| 36.4| 3.1  | 4.6 | 13.8| 44.6| 5.8  | 8.4 | 16.9| 40.3|
| Lo + Sc + CLIP            | Visual   | 2.5  | 3.7 | 10.5| 33.8| 4.6  | 4.6 | 10.8| 33.8| 3.2  | 5.2 | 11.7| 43.5|
| CLIP + B-Bd + B-Et        | Textual+Visual | 64.6 | 68.9 | 78.5 | 86.6 | 69.2 | 70.8 | 84.6 | 90.3 | 61.0 | 64.9 | 74.7 | 83.8|
| CLIP + Lo + B-Bd + B-Et   | Textual+Visual | 61.4 | 66.0 | 76.8 | 86.3 | 66.2 | 70.8 | 81.5 | 90.8 | 61.0 | 64.9 | 74.0 | 81.8|
| CLIP + Sc + B-Bd + B-Et   | Textual+Visual | 63.1 | 68.0 | 78.0 | 86.0 | 63.1 | 67.7 | 76.9 | 86.2 | 63.6 | 68.8 | 77.3 | 83.8|
| Lo + Sc + B-Bd + B-Et     | Textual+Visual | 65.1 | 69.5 | 78.7 | 84.8 | 70.8 | 75.4 | 81.5 | 86.2 | 63.6 | 68.2 | 77.9 | 80.5|
| CLIP + Lo + Sc + B-Bd + B-Et | Textual+Visual | 65.5 | 70.0 | 81.2 | 88.7 | 72.3 | 76.9 | 83.1 | 90.8 | 63.6 | 69.5 | 81.2 | 85.7|

2500 kilometers, respectively. For the evaluation on BreakingNews (Section 4.3), we use mean and median GCD metrics as suggested by Ramisa et al. [27].

**Hyperparameters:** We use the Adam optimizer, learning rate of $10^{-4}$, and batch size of 128 for optimization. We train the models for 500 epochs and perform validation after every epoch. The best model is used based on the highest mean accuracy in city and region GCD thresholds on the validation set. We set the parameter for data filtering to $k = 6$ (Section 2) based on an empirical evaluation of a small subset (150 samples) of the dataset.

**Compared Systems:** We evaluate different combinations of the proposed model based on the feature modalities. We also compare against two popular text-based methods Cliff-clavin [7], Mordecai [10] and one image-based state-of-the-art model (ISNs: Individual Scene Networks [24]).

### 4.2 Experimental Results on MM-Locate-News

The results are reported in Table 3 and discussed below.

**Textual Models:** For smaller GCD thresholds specifically city and region, in T2, the combination B-Et + B-Bd improves the performance, and in T1 and T3 the B-Et model provides the best results. When used separately, B-Et has a more substantial impact than B-Bd, indicating that named entities and their frequency play a vital role to predict the focus location in the news. While Mordecai and...
achieve the best results at country and continent level for $T_1$ and $T_3$, respectively, these baselines are either not applicable (Mordecai) or achieve worse results (Cliff-clavin) compared to our models on more fine-grained levels.

**Visual Models:** The results show that CLIP$_i$ performs well in all test variants providing the best results on $T_1$ and $T_3$, and that combinations with scene (Sc + CLIP$_i$) and location features (Lo + Sc + CLIP$_i$) can further improve the results. ISNs specifically trained for photo geolocalization achieve superior results on $T_2$ where images depict the query location and provide enough geographical cues. Unlike CLIP$_i$, ISNs do not generalize well on other test variants.

**Multimodal Models:** The combination of CLIP$_i$ with multimodal information drastically improves the results compared to unimodal models in all test data variants and distance thresholds. Even though our visual models do not outperform ISNs in $T_2$, they considerably improve the results when combined with textual features (Lo + Sc + B-Bd + B-Et). These results suggest that a multimodal architecture is beneficial for focus location estimation in news.

**Qualitative Results:** Figure 4 presents the sample outputs for different models. As expected, it is easier for the textual model to predict the correct focus location when there are multiple mentions of the ground truth location in the text (Figure 4a.1), or person names related to the focus location such as Figure 4a.2, which mentions Masis Mayilian the foreign minister of Artsakh. However, the textual model fails when there are multiple locations (Figure 4b.2) or person names (Figure 4b.1) which are irrelevant to the focus location. It is very challenging for the visual model to predict the correct focus location due...
Table 4: Mean and median GCD (km) using the best models per modality on the BreakingNews test set (mean and median values are divided by 1000).

| Approach                                           | Mean | Median |
|----------------------------------------------------|------|--------|
| CLIP                                               | 2.84 | 1.3    |
| B-Bd + B-Et                                        | 1.93 | 0.88   |
| CLIP + Lo + Sc + B-Bd + B-Et                      | 1.88 | 0.83   |
| Places [27]                                        | 3.40 | 0.68   |
| W2V matrix [27]                                    | 1.92 | 0.90   |
| VGG19 + Places + W2V [27]                         | 1.91 | 0.88   |

As shown in the table, CLIP has the highest mean and median GCD values among the models evaluated. This indicates that CLIP alone does not perform well in estimating focus locations in news images.

### 4.3 Experimental Results on BreakingNews

We also evaluate our proposed focus location estimation architectures on BreakingNews [27] for comparison. We divide the dataset into 33,376, 11,210, and 10,581 samples for train, validation, and test splits. However, please note that the quality of ground truth labels in BreakingNews is much lower compared to our proposed MM-Locate-News dataset as the test data are not annotated manually for focus locations. For a better comparison to the approaches presented by Ramisa et al. [27], we train a regression model by minimizing the GCD between the ground truth (the first geo-coordinate as mentioned in the original paper [27]) and the predicted geographical coordinates based on the various types of state-of-the-art visual and textual features presented in Section 3.

Table 4 shows the comparison of the regression models provided by Ramisa et al. [27] with our approaches using Mean and Median GCD values (divided...
by 1000). From our proposed models, as expected the multimodal model $CLIP_i + Lo + Sc + B-Bd + B-Et$ achieves the best performance. In comparison with $VGG19 + Places + W2V$ matrix from Ramisa et al. [27], our model achieves better results with a difference of 30 km and 50 km in terms of mean and median GCD, respectively. Regarding the proposed unimodal models $B-Bd + B-Et$ outperforms $CLIP_i$ in terms of both mean and median. The proposed textual model $B-Bd + B-Et$ achieves slightly better results in comparison with $W2V$ matrix [27] regarding the median value (with difference of 20 km). Regarding the visual models, $CLIP_i$ outperforms Places [27] in terms of mean value by 560 km. However, in terms of median value it is worse with a margin of 620 km. This may be explained by more (bad) outlier predictions of the Places model. Overall, the results suggest that a multimodal architecture is beneficial for the task of location estimation in news.

5 Conclusion

In this paper, we have introduced a novel dataset called $MM-Locate-News$, which is a multimodal collection of image-text pairs for focus location estimation of news articles. A weakly-supervised method has been suggested to collect news with focus locations. We manually annotated the test data to acquire three different test data variants. We have proposed various baselines and multimodal approaches using state-of-the-art transformer-based architectures for the task of focus location estimation. The experimental results have shown that the combination of textual and visual features for geolocation estimation outperforms the compared approaches that rely on features from a single modality.

In future work, we will extend the dataset by considering additional languages and locations. We will also investigate multimodal geolocation estimation on different news domains and apply further noise removal to the images.

Acknowledgements. This work was partially funded by the EU Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no. 812997 (CLEOPATRA ITN), and by the Ministry of Lower Saxony for Science and Culture (Responsible AI in digital society, project no. 51171145).

References

1. Andogah, G., Bouma, G., Nerbonne, J.: Every document has a geographical scope. Data & Knowledge Engineering 81, 1–20 (2012)
2. Armitage, J., Kacupaj, E., Tahmasebzadeh, G., Swati, Maleshkova, M., Ewerth, R., Lehmann, J.: MLM: A benchmark dataset for multitask learning with multiple languages and modalities. In: International Conference on Information and Knowledge Management, CIKM. pp. 2967–2974 (2020). https://doi.org/10.1145/3340531.3412783
3. Brank, J., Leban, G., Grobelnik, M.: Semantic annotation of documents based on wikipedia concepts. Informatica (Slovenia) 42(1) (2018). http://www.informatica.si/index.php/informatica/article/view/2228
4. Crandall, D.J., Backstrom, L., Huttenlocher, D.P., Kleinberg, J.M.: Mapping the world’s photos. In: International Conference on World Wide Web, WWW. pp. 761–770 (2009). https://doi.org/10.1145/1526709.1526812

5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)

6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 4171–4186 (2019), https://doi.org/10.18653/v1/n19-1423

7. D’Ignazio, C., Bhargava, R., Zuckerman, E., Beck, L.: Cliff-clavin: Determining geographic focus for news articles. In: NewsKDD: Data Science for News Publishing Workshop co-located with ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2014)

8. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: European Conference on Computer Vision, ECCV. pp. 241–257 (2016). https://doi.org/10.1007/978-3-319-46466-4_15

9. Gritta, M., Pilehvar, M.T., Collier, N.: Which melbourne? augmenting geocoding with maps. In: 56th Annual Meeting of the Association for Computational Linguistics, ACL. pp. 1285–1296 (2018), https://aclanthology.org/P18-1119/

10. Halterman, A.: Mordecai: Full text geoparsing and event geocoding. The Journal of Open Source Software 2(9) (2017). https://doi.org/10.21105/joss.00091

11. Hays, J., Efros, A.A.: IM2GPS: estimating geographic information from a single image. In: Conference on Computer Vision and Pattern Recognition, CVPR (2008)

12. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision, ECCV. pp. 630–645 (2016). https://doi.org/10.1007/978-3-319-46493-0_38

13. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing (2017). https://spacy.io

14. Imani, M.B., Chandra, S., Ma, S., Khan, L., Thuraisingham, B.M.: Focus location extraction from political news reports with bias correction. In: International Conference on Big Data, BigData. pp. 1956–1964 (2017). https://doi.org/10.1109/BigData.2017.8258141

15. Izbicki, M., Papalexakis, E.E., Tsotras, V.J.: Exploiting the earth’s spherical geometry to geolocate images. In: European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD. pp. 3–19 (2019), https://doi.org/10.1007/978-3-030-46147-8_1

16. Kim, H.J., Dunn, E., Frahm, J.: Learned contextual feature reweighting for image geo-localization. In: Conference on Computer Vision and Pattern Recognition, CVPR. pp. 3251–3260 (2017). https://doi.org/10.1109/CVPR.2017.346

17. Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, I.: Geotagging text content with language models and feature mining. Proceedings of the IEEE 105(10), 1971–1986 (2017). https://doi.org/10.1109/JPROC.2017.2688799

18. Kordopatis-Zilos, G., Popescu, A., Papadopoulos, S., Kompatsiaris, Y.: Placing images with refined language models and similarity search with pca-reduced VGG features. In: MediaEval 2016 Workshop. vol. 1739 (2016). http://ceur-ws.org/Vol-1739/MediaEval_2016_paper_13.pdf
19. Krippendorff, K.: Computing krippendorff’s alpha-reliability (2011), https://repository.upenn.edu/asc_papers/43

20. Kulkarni, S., Jain, S., Hosseini, M.J., Baldridge, J., Ie, E., Zhang, L.: Multi-level gazetteer-free geocoding. In: International Workshop on Spatial Language Understanding and Grounded Communication for Robotics. pp. 79–88 (2021)

21. Larson, M.A., Soleymani, M., Gravier, G., Ionescu, B., Jones, G.J.F.: The benchmarking initiative for multimedia evaluation: Mediaeval 2016. IEEE MultiMedia 24(1), 93–96 (2017), https://doi.org/10.1109/MMUL.2017.9

22. Lieberman, M.D., Samet, H., Sankaranarayanan, J.: Geotagging with local lexicons to build indexes for textually-specified spatial data. In: International Conference on Data Engineering, ICDE. pp. 201–212 (2010), https://doi.org/10.1109/ICDE.2010.5447903

23. Lin, T., Belongie, S.J., Hays, J.: Cross-view image geolocalization. In: Conference on Computer Vision and Pattern Recognition CVPR. pp. 891–898 (2013), https://doi.org/10.1109/CVPR.2013.120

24. Müller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: European Conference on Computer Vision, ECCV. pp. 575–592 (2018), https://doi.org/10.1007/978-3-030-01258-8_35

25. Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Hakimov, S., Ewerth, R.: Multimodal news analytics using measures of cross-modal entity and context consistency. International Journal of Multimedia Information Retrieval 10(2), 111–125 (2021), https://doi.org/10.1007/s13735-021-00207-4

26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, ICML, pp. 8748–8763 (2021), http://proceedings.mlr.press/v139/radford21a.html

27. Ramisa, A., Yan, F., Moreno-Noguer, F., Mikolajczyk, K.: Breakingnews: Article annotation by image and text processing. IEEE Transactions in Pattern Analysis and Machine Intelligence 40(5), 1072–1085 (2018), https://doi.org/10.1109/TPAMI.2017.2721945

28. Seo, P.H., Weyand, T., Sim, J., Han, B.: Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In: European Conference on Computer Vision, ECCV. pp. 544–560 (2018), https://doi.org/10.1007/978-3-030-01249-6_33

29. Sendyukov, P., Murdoch, V., van Zwol, R.: Placing flickr photos on a map. In: SIGIR Conference on Research and Development in Information Retrieval, SIGIR. pp. 484–491 (2009), https://doi.org/10.1145/1571941.1572025

30. Uffelder, J., Schrodt, P.: Political instability task force worldwide atrocities event data collection codebook. version 1.0 b2 (2009)

31. Uzkent, B., Sheehan, E., Meng, C., Tang, Z., Burke, M., Lobell, D.B., Ermon, S.: Learning to interpret satellite images using wikipedia. In: International Joint Conference on Artificial Intelligence, IJCAI. pp. 3620–3626 (2019), https://doi.org/10.24963/ijcai.2019/502

32. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014), https://doi.org/10.1145/2629489

33. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2 - A large-scale benchmark for instance-level recognition and retrieval. In: Conference on Computer Vision and Pattern Recognition, CVPR. pp. 2572–2581 (2020)
34. Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(6), 1452–1464 (2018), https://doi.org/10.1109/TPAMI.2017.2723009