Multimodal Geolocation Estimation of News Photos

Golsa Tahmasebzadeh[1,2], Sherzod Hakimov[3], Ralph Ewerth[1,2], and Eric Müller-Budack[1,2]

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

Abstract. The widespread growth of multimodal news requires sophisticated approaches to interpret content and relations of different modalities. Images are of utmost importance since they represent a visual gist of the whole news article. For example, it is essential to identify the locations of natural disasters for crisis management or to analyze political or social events across the world. In some cases, verifying the location(s) claimed in a news article might help human assessors or fact-checking efforts to detect misinformation, i.e., fake news. Existing methods for geolocation estimation typically consider only a single modality, e.g., images or text. However, news images can lack sufficient geographical cues to estimate their locations, and the text can refer to various possible locations. In this paper, we propose a novel multimodal approach to predict the geolocation of news photos. To enable this approach, we introduce a novel dataset called Multimodal Geolocation Estimation of News Photos (*MMG-NewsPhoto*). *MMG-NewsPhoto* is, so far, the largest dataset for the given task and contains more than half a million news texts with the corresponding image, out of which 3000 photos were manually labeled for the photo geolocation based on information from the image-text pairs. For a fair comparison, we optimize and assess state-of-the-art methods using the new benchmark dataset. Experimental results show the superiority of the multimodal models compared to the unimodal approaches.

Keywords: Multimodal Photo Geolocalization · News Analytics · Information Retrieval

1 Introduction

Multimedia data have been growing exponentially on the Web and social media in the last decade. To convey information more efficiently, many news articles appear in a multimodal format, i.e., using both image and text. However, along with the proliferation of news articles, fake news has gathered momentum. Hence, it is essential to organize, analyze, and contextualize the image content.

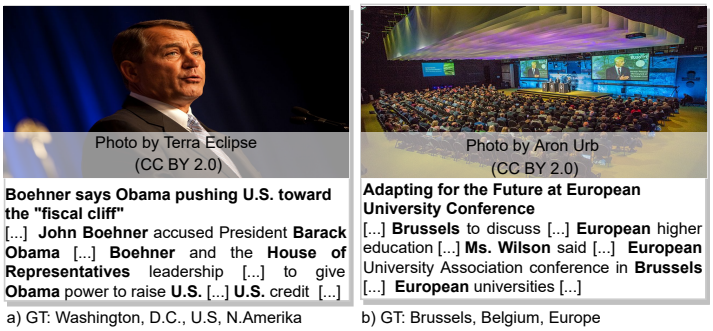

Fig. 1: Samples from the *MMG-NewsPhoto* dataset. GT: Ground Truth location. Photos are replaced with similar ones due to license restrictions.

The estimation of the geolocation of news images is an important aspect for various real-world applications. Example applications are news retrieval [1], image verification [10], and misinformation detection in news [34].

Most previous approaches for geolocation estimation of photos depend solely on visual information [16,17,25], and only a few methods process more than one modality [20,21]. Existing image-based methods are mainly focused on specific environments such as cities [5,17] or landmarks [2,7,42]. However, the image-based methods are unable to represent news-related geographic features such as *public personality* (Fig. 1 a) or an *event* (Fig. 1 b). Most of the multimodal approaches are based on the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset [36], and depend on the tags provided with the images. However, they do not make use of rich textual information provided in news body that indicates possible photo locations (Fig. 1 b). A multimodal dataset of news articles is *BreakingNews* [31], where the geolocation labels are provided by the news feed primarily taken from the RDF (Resource Description Framework) Site Summary (RSS) or, if not available, inferred using heuristics such as the publisher location or the story text [31]. However, the extracted geolocations can be inaccurate or even wrong. Another drawback of the *BreakingNews* dataset is that the labels of the test split are derived in the same way. Overall, there is a considerable need for a multimodal dataset of news articles that provides geolocation labels specifically for images, as well as multimodal solutions for geolocation estimation of news photos.

In this paper, we define the task of photo geolocalization as a multimodal problem. We propose a multimodal approach that considers visual and textual information from the news photo and body text that integrates hierarchical information of different granularities (spatial resolutions). The main contributions are summarized as follows: (1) We introduce the *MMG-NewsPhoto* (Multimodal Geolocation Estimation of News Photos) dataset that contains more than half a million news articles. The proposed dataset covers more than 14,000 cities and 241 countries across all continents within multiple news domains such as *Health*, *Business*, *Society*, and *Politics*; (2) We provide extensive annotation guidelines

and define news-specific visual concepts that represent the photo geolocation; (3) We propose a multimodal approach that leverages state-of-the-art visual and textual features for multimodal photo geolocalization; (4) We evaluate our proposed method on two datasets, including *MMG-NewsPhoto* and compare it against state-of-the-art methods, including some baseline re-implementations. The source code, dataset, and annotation guidelines are publicly available[4].

The remainder of the paper is structured as follows. Section 2 describes the related work. In Section 3, the acquisition of the dataset is explained. The proposed model for multimodal geolocation estimation is presented in Section 4, while the experimental setup and results are discussed in Section 5. Section 6 concludes the paper and outlines future directions.

2 Related Work

There are two main criteria to classify the approaches for geolocation estimation of photos: the environment target and the data type, i.e., images and multimodal data [9]. In this section, we briefly review related work on photo geolocation estimation and primarily focus on multimodal approaches, existing datasets, and their drawbacks.

Image-based Approaches. Many existing methods based on image geolocalization focus on urban [5,17] and natural environments, such as mountains [4,37]. Some attempts estimate photo location at global scale without any prior assumptions about the environment. Most of them treat geolocation estimation as a classification problem [25,32,35,43]. Improvements were made, for example, by exploiting a retrieval approach and a large geo-tagged image database [40], using overlapping sets of visually similar cells [32], incorporating a hierarchical cell structure as well as environmental scene context [25], or leveraging the advantages of contrastive learning [19]. However, while these approaches achieve promising results solely based on visual information, news provides textual information that can further increase the performance, particularly in the absence of distinct geographical cues (Fig. 1 b).

Multimodal Approaches. There are only few methods [11,20,21,31,33] that address geolocation estimation as a multimodal problem most of which rely on constructing large-scale geographical language models by generating a probabilistic model based on mentions of textual tags across the globe [20,21,33]. Crandall et al. [11] combine image content and textual metadata at two granularity levels, at a city level (≈100 km) and landmark level (≈100 m). Trevisiol et al. [38] process the textual information of a set of videos to determine their georelevance and to find frequent matching items. In case of lack of such information, they resort to visual features. Later, a multimodal approach was proposed by Ramisa et al.[31] where they combine visual features with text using the nearest neighbor method and Support Vector Regression (SVR).

[4] Source code & dataset: https://github.com/TIBHannover/mmg-newsphoto

Multimodal Datasets. Most multimodal approaches are based on the *YFCC 100M* dataset [36] or the *MediaEval Placing Task* benchmark datasets [23] including images, videos, and metadata. Another dataset proposed by Uzkent et al. [39] contains images and text from Wikipedia combined with satellite images. More recently, a dataset called *Multiple Languages and Modalities* (MLM) [1] has been introduced, which includes images along with multilingual texts from *Wikidata* [41] for multimodal location estimation and information retrieval [1]. Unlike the previous datasets, the *BreakingNews* introduced by Ramisa et al. [31] contains multimodal news articles and is the most relevant for our work. It includes image, text, caption, and metadata (such as geo-coordinates and popularity) and covers various domains such as *Sports*, *Politics*, and *Health*. The provided geolocation labels for both training and evaluation are extracted from the RSS, publisher, or news text. But as discussed in Section 1, these automatically derived locations can be inaccurate or even wrong. Instead, we provide high-quality manually annotated photo geolocations for fair and reliable evaluation. In addition, our proposed *MMG-NewsPhoto* dataset includes more than half a million samples (*BreakingNews* only includes around 60,000 samples with geolocations) from 241 countries and more than 14,000 cities across continents.

3 MMG-NewsPhoto Dataset

In this section, we explain the dataset creation (Section 3.1) and annotation process (Section 3.2) of the proposed *MMG-NewsPhoto* dataset for multimodal geolocation estimation of news images.

3.1 Dataset Creation

Datasets. We use the collection of articles provided by the *Good News* [6] and *CC-News* [24] datasets. *Good News* [6] is an image captioning dataset comprising 466,000 image-caption pairs. Based on web links to the news articles, we extract all articles with a body text, title, image link(s) with corresponding caption(s), and domain label(s). *CC-News* [24] includes 44 million documents written in English extracted from around 30,000 unique news sources. We sort the sources based on the number of news articles and scrape news documents from the top-20 sources in the same way as mentioned above. Finally, we download all the images and discard the ones with corrupted or inaccessible images. As a result, we end up with circa 10 million data samples, including body text, and at least one image-caption pair per sample acquired from both news sources.

Initial Removal. We remove redundant documents (except one) based on the cosine similarity (normalized to [0, 1]) of the body texts using *TF-IDF* (Term Frequency; Inverse Document Frequency) above a threshold of 0.5. Next, we manually group the domain labels into ten categories such as *Health*, *Business*, and *Politics* (see full list in Fig. 2, left). Some domains such as *Art* and *Technology* include various invalid images for the task, i.e., ads or stock photos. We

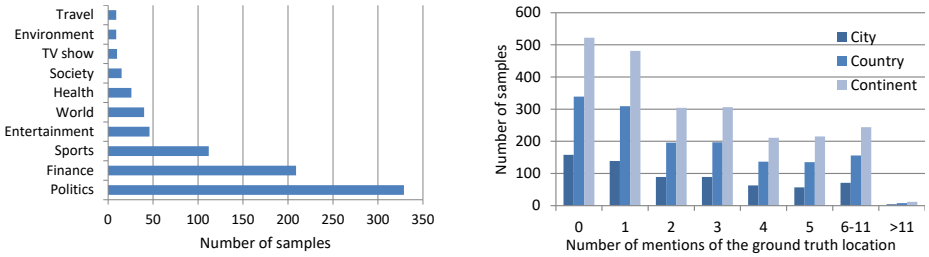

Fig. 2: Left: Test data distribution among domains. Right: Frequency of ground truth location mentions in the body text for the test split.

discard these types of images as they typically lack geographic content or do not correspond to the locations mentioned in the body text of news.

Location Linking. We assume that locations mentioned in a caption are good candidates for photo geolocation. We apply named entity recognition and disambiguation to extract all locations in the captions. Following related work [27], we use *spaCy* [15] to extract the named entities and use *Wikifier* [8] to link them to *Wikidata* entities. We only keep entities of type *Location* with valid geocoordinates (latitude, longitude) extracted from the *Wikidata* Property *P625*.

Photo Location Assignment. The location entities extracted from the captions do not always indicate the photo locations and can, for example, also refer to entity attributes, e.g., "*U.S. President Biden*". Thus, captions are tokenized to extract certain prepositions, e.g., "across", "along", and "in", which combined with a location mention, are more likely to refer to the photo location. We keep samples for which the distance of one of 37 prepositions[4] to the *claimed photo location* is at most two tokens. Furthermore, samples with more than one unique location are removed, resulting in exactly one *claimed photo location*.

Location Enrichment. We apply reverse-geocoding to map around 50,000 fine-grained locations (i.e., city, road, building, etc.) extracted from the captions to cities using *Nominatim* [29]. Next, we extract associated country (*Property P17*), continent (*Property P30*) and geo-coordinates (*Property P625*) from *Wikidata*.

Data Sampling. For manual annotation, 3000 samples are selected to construct the test dataset. To avoid bias, the samples are selected (1) from all domains, (2) from all continents, (3) from highly populated cities (minimum population of 500,000) and medium populated cities (population: 20,000 to 500,000), (4) with at least three unique locations mentioned in text, and (5) with different number of mentions of the ground-truth location in the body text. The latter ensures that simple cases with frequent mentions of the ground truth and complex cases,

i.e., many locations mentioned in the text with somewhat equal frequencies, are included. For simple cases, a textual approach that leverages the frequency of named entities can already achieve high performance without even considering the image. Based on complex cases, we can analyze the direct impact of the image for multimodal geolocation estimation. The statistics for the test split are visualized in Fig. 2, right. From the remaining samples, 10% are randomly chosen for validation, and the rest is used for training.

3.2 Data Annotation Process and Guidelines

We give an in-depth explanation of the guidelines used for the manual annotation of the test split, which is aimed at making the assessment fair and transparent. The exact guidelines used during annotation are provided on our *GitHub* page[4].

Geo-representative Concepts. For photo geolocation estimation, a *geographically representative image* depicts concepts that help to identify its location. We group *geo-representative concepts* into two types: *strong* and *weak concepts*. A *strong concept* is a unique identity of a location, e.g., the appearance of the *Eiffel Tower* in an image that can unambiguously be assigned to the city *Paris*, country *France*, and continent *Europe*. A *weak concept*, on the other hand, provides clues for one or even a few specific locations but without sufficient evidence on its own. For example, a certain *President* is an identity of a country but can travel to different locations. Only multiple *weak concepts*, all of which correspond to the same location, in an image can lead to the identification of the geolocation of news photos. For instance, multiple *car plates* or *groups of people* can represent the corresponding country. As shown in Table 1, we define *strong* or *weak* visual concepts based on the following eight categories: *building*, *clothing*, *event*, *group of people*, *natural scenery*, *object*, *public personality* and *scene text*.

Annotation Questions (Q). Given an image-caption pair and the linked location of the caption, we ask each annotator the following questions:
Q1: ***Is it a valid sample?*** To determine whether a sample is valid for the identification of the photo geolocation, an annotator selects *"no"* if (1) the image is an ad, a stock photo, a web page, a map, or a data visualization, (2) the linked location is wrong, not a location, or not the *claimed photo location* (see paragraph *Photo Location Assignment*) of the caption. Otherwise, *"yes"* is chosen.
Q2: ***Which weak and strong concepts are shown in the image?*** The annotator selects the strong or weak concepts (Table 1) depicted in the image.
Q3: ***Is the linked city (Q3.1), country (Q3.2), continent (Q3.3) shown in the image?*** These questions are asked to obtain the ground-truth location at various granularities. A user selects *"yes"* if (1) at least one *strong concept* is visible, (2) a single *weak concept* occurs in high frequency (e.g., multiple *car plates*), (3) a combination of at least two distinct *weak concepts* is shown, or (4) a single *weak concept* with valid proof (e.g., a Web page that proves the

Table 1: Strong and weak visual concepts used in the annotation process.

Strong geo-representative concepts			
Category	**City**	**Country**	**Continent**
Building	Buildings, landmarks	-	-
Clothing	-	Public service uniforms	-
Event	Social movements, sports competitions	Social movements, sports competitions, natural disasters, country elections, wars	Sports competitions, natural disasters
Group of people	-	-	-
Natural scenery	City-specific natural landmarks	Country-specific natural landmarks	Continent-specific natural landmarks
Object	Logos of events, organizations, etc.	Public service vehicles	-
Public personality	-	-	-
Scene text	Street signs with mentions of cities	Country names in signs	-
Weak geo-representative concepts			
Category	**City**	**Country**	**Continent**
Building	-	Buildings with specific architectures	-
Clothing	Uniforms of sport clubs	Uniforms of soldiers, cultural costumes, national sport team uniforms	-
Event	-	-	-
Group of people	-	Residents of a country, common activity	-
Natural scenery	-	-	Land forms, flora, fauna
Object	-	Personal cars and/or car plates, flag, logo	-
Public personality	-	Politicians, athletes, celebrities	-
Scene text	-	Text in specific language	-

location) is provided. Otherwise, *"no"* is selected. If *"yes"* is given as an answer, a confidence level is selected: *"very confident"*, *"confident"*, and *"not confident"*.
Q4: ***What is the environmental setting of the image?*** The user selects one of the following categories: *"indoor"*, *"outdoor urban"*, *"outdoor nature"* to indicate the environment in which an image was taken.
Q5: ***Is it a closeup?*** Since locations are usually difficult to predict for closeups, we asked the annotators to identify whether the image shows a closeup or not.
Q6: ***Did you need external resources for Q3?*** The final question determines whether or not the annotator needed external resources to decide on Q3. If "Yes" is selected, we asked the annotators to provide the links.

Annotator Training. We employed four graduate students with computer science backgrounds who were paid 10 EUR per hour (slightly above the minimum wage in Germany in early 2022) for annotations. Furthermore, three experts (doctoral and postdoctoral researchers) with a research focus on computer vision and multimodal analytics provided annotations. All annotators were trained based on the annotation guidelines[4]. We performed two dry runs using 100 samples and discussed the results to refine the guidelines.

Annotation Process. The annotation task was performed in two steps as follows. (1) All annotators were asked to validate the 3000 samples according to Q1. Using majority voting, 1700 valid samples were obtained. (2) For each *valid*

Table 2: Data distribution for continents (top) and top-10 countries (bottom).

	Europe	**N.America**	**Asia**	**Oceania**	**Africa**	**S.America**	**Total**
Train	190,064	188,175	121,045	20,468	21,096	13,920	554,768
Validation	21,041	20,675	13,120	2147	2,331	1,579	60,893
Test$_{city}$	196	189	215	13	27	20	660
Test$_{country}$	235	212	274	13	35	25	794
Test$_{continent}$	235	215	278	13	37	27	805
Total	211,769	209,466	134,932	22,654	22,526	15,573	617,920

	U.S.	**U.K.**	**India**	**China**	**Australia**	**France**	**Japan**	**Germany**	**Spain**	**Russia**
Train	173,584	82,917	27,435	18,390	17,018	16,347	15,669	14,477	13,702	9,330
Validation	19,076	9,253	3,024	2,007	1,805	1,766	1,732	1,569	1,459	1,055
Test$_{country}$	190	82	121	11	11	8	17	24	11	15

sample, Q2 to Q6 were annotated by three annotators, and majority voting was applied to select samples where two users agreed on the answer per question. Based on selected answers for Q3.1 to Q3.3, we obtained the final annotations. For all questions, the answer should be *"yes"*, with a confidence level of either *"very confident"* or *"confident"*. Samples, where at least two annotators selected the confidence level *"not confident"* were re-annotated by an expert. As a result, we obtained final annotations for Q3.1, Q3.2, and Q3.3, where the answers correspond to the granularity of the geolocation of images. These granularities are turned into three variants of the test data: Test$_{city}$, Test$_{country}$, Test$_{continent}$. Please note that finer granularity samples are subsets of coarser granularities.

Annotation Study Findings. Krippendorff's alpha [22] was used to calculate inter-annotator agreements for Q3. The agreements are 0.41 for *city*, 0.41 for *country*, and 0.51 for the *continent*, which we consider low to moderate. Responses to Q4 and Q5 indicated that 40.2% of the images are closeups and 37.7% are indoor images, both of which typically depict few weak geo-representative concepts and are challenging for the photo-geolocation task. For 49.7% of the samples, annotators needed external resources (Q6) to decide whether the image showed the linked location. Overall, these numbers demonstrate the difficulty of the task for humans and explain the moderate inter-coder agreement for Q3.

Dataset Statistics. The *MMG-NewsPhoto* dataset includes 554,768 training, 60,893 validation, and 2259 test samples (sum for all granularities). The dataset contains 14,331 cities, 241 countries, and 6 continents. Table 2 shows data distribution among continents and top-10 countries. Since 1700 test samples and thus about 57% of the test samples are valid, we assume that train and validation sets contain a similar proportion of valid samples.

4 Multimodal Photo Geolocation Estimation

We define multimodal geolocation estimation of news photos as a classification task, where the photo location is predicted based on the visual content and con-

textual information from the accompanied body text. The number of $|\mathbb{C}_g|$ locations available in the dataset for a granularity g (e.g., city, country, or continent) are considered as target classes. The $|\mathbb{C}_g|$-dimensional one-hot encoded vector $\mathbf{y}_g = \langle y_1, y_2, \ldots, y_{|\mathbb{C}_g|} \rangle \in \{0,1\}^{|\mathbb{C}_g|}$ represents the ground-truth location. In the remainder of this section, we define the features incorporated from state-of-the-art approaches and describe the multimodal architecture and loss function.

Textual Features. The pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) [12] is employed to extract two distinct types of textual features, each with 768 dimensions, from the body text of the news article. (1) We average the embeddings extracted with BERT of each sentence to create a single vector B-Bd $\in \mathbb{R}^{768}$ to encode the global contextual information. (2) To create an entity-centric embedding, denoted as B-Et $\in \mathbb{R}^{768}$, we follow related work [27] and combine *spaCy* [15] and *Wikifier* [8] to link location, person, and event entities to *Wikidata*. The BERT embeddings for these entities are extracted based on their *Wikidata* label. Finally, we compute the average of the entity vectors taking into account multiple mentions of the same entity, as they may be more important for the geolocation of the photo.

Visual Features. To represent the *geo-representative visual concepts*, we rely on CLIP (Contrastive Language-Image Pretraining) [30]. We use ViT-B/32 image encoder to extract visual features with 512 dimensions denoted as CLIP$_i \in \mathbb{R}^{512}$.

Network Architecture. In our proposed model architecture, we aim to combine textual and visual features to predict photo geolocations on various granularities, i.e., city, country, and continent levels. Since the feature dimension of visual and textual features differ, we first encode each feature vector using l_e fully-connected (FC) layers with n_e neurons each. Next, we concatenate these embeddings and feed them into l_o output FC-layers. In the hidden output layers, we use n_o neurons, and in the last output layer, the number of neurons corresponds to the number of locations $|\mathbb{C}_g|$ for a given granularity g. To leverage the hierarchical information, we employ individual classifiers for each granularity in city, country, and continent level to output probabilities $\hat{\mathbf{y}}_g \in \mathbb{R}^{|\mathbb{C}_g|}$ of size $|\mathbb{C}_{city}| = 14,331$, $|\mathbb{C}_{country}| = 241$, and $|\mathbb{C}_{continent}| = 6$. Please note that we use the *Rectified Linear Unit (ReLU)* activation function [28] for all layers except the last output layer that uses a *softmax*. More details are provided on GitHub[4].

Loss Function. To aggregate the granularity classifiers and highlight the hierarchical attribution, we build a multi-task learning loss function as follows:

$$\mathcal{L} = \sum_g \lambda_g \mathcal{L}_g(\mathbf{y}_g, \hat{\mathbf{y}}_g), \text{ with } g \in \{\text{city, country, continent}\}, \quad (1)$$

$$\mathcal{L}_g(\mathbf{y}_g, \hat{\mathbf{y}}_g) = -\mathbf{y}_g \log \hat{\mathbf{y}}_g - (1 - \mathbf{y}_g) \log(1 - \hat{\mathbf{y}}_g), \quad (2)$$

where λ_g are the relative weights learned during training for the different granularities, considering the difference in magnitude between losses by consolidating the log standard deviation. The cross-entropy loss $\mathcal{L}_g$ for a single granularity $g \in \{\text{city, country, continent}\}$ is defined according to Equation (2).

5 Experimental Setup and Results

This section presents the experimental setup, comparison of different architectures on the proposed *MMG-NewsPhoto* dataset as well as on *BreakingNews* [31].

5.1 Experimental Setup

Evaluation Metrics. We use the Great Circle Distance (GCD) between the geocoordinates of the predicted and ground-truth location at several tolerable error radii [13]. These values are 25, 200, and 2500 kilometers for city, country, and continent, respectively. Furthermore, we measure the Accuracy@k that indicates whether the ground-truth location is within the top-k model predictions.

Hyperparameter Settings. To extract textual features, we limit the text to 500 tokens. We set the number of FC-layers to $l_e = 2$, $l_o = 2$ and choose $n_e = 1024$, $n_o = 512$ neurons (Section 4). While *single-task learning* model variants (denoted with stl) are optimized using a single granularity g, the remaining models use the multi-task loss presented in Equation (1) to learn from hierarchical geographical information. We use the *Adam* optimizer [18], a learning rate of 10^{-5}, batch size of 256, weight decay of 0 for optimization, ReLU activation $max(0, x)$ [28], and norm [3] with a clamp $min = 1 \times 10^{-12}$. Before each layer, a dropout with a ratio of 0.1 is applied. We train all the models for 100 epochs and clip gradients with a max norm of 5. The model with the lowest loss on the validation set is used for evaluation.

Baselines. The model architectures that are used to compare against our proposed models, and the experimental results are as follows. Note that we did not fine-tune these models and used their official models or implementations.

$base(M, f^*)$ [25] is a state-of-the-art model for photo geolocation estimation model based on ResNet-101 [14] pre-trained on a subset of *YFCC100M* [36].

$T\text{-}base(M, f^*)$ is an extension of $base(M, f^*)$ where its predictions are reduced to mentioned locations in the news body to include textual information.

T-Freq is based on language models for geo-tagging text [20,23,33]. We employ a statistical model based on frequency of entities per city using the train set. More details are provided in the supplemental material on *GitHub*[4]. The predicted location per sample is the one with the highest probability.

VT_{CM} is based on the cross-modal entity consistency of image and text [26] based on persons, locations, and events. To get predictions per test image, we sort *Cross-modal Location Similarity (CMLS)* values and get the top k locations.

Table 3: Fraction of samples [%] localized within a GCD of at most 25 km (**CI:** city level), 200 km (**CR:** country level), and 2500 km (**CT:** continent level) on MMG-NewsPhoto.

Approach	CI	CR	CT
$base(M, f^*)$ [25]	10.3	20.2	40.9
CLIP_i	**30.6**	**65.5**	**78.3**
T-Freq	12.6	31.5	49.9
B-Bd	31.5	73.4	85.6
B-Et	31.4	73.7	83.5
B-Bd ⊕ B-Et	**32.1**	**74.7**	**84.6**
$T\text{-}base(M, f^*)$	31.2	58.8	70.7
VT_{CM} [26]	22.3	50.1	60.1
CLIP_i ⊕ B-Bd ⊕ B-Et	**43.0**	**76.7**	**83.4**

Table 4: Mean and median GCD divided by 1000 km on city level for the BreakingNews test set. Models trained on MMG-NewsPhoto are evaluated in a zero-shot setting on BreakingNews and MMG → BN means that the model is finetuned on BreakingNews.

Approach	Training	Mean	Median
CLIP_i	MMG	3.67	1.37
CLIP_i	MMG → BN	3.22	0.92
B-Bd ⊕ B-Et	MMG	2.26	**0.47**
B-Bd ⊕ B-Et	MMG → BN	**2.25**	0.51
CLIP_i ⊕ B-Bd ⊕ B-Et	MMG	2.70	0.63
CLIP_i ⊕ B-Bd ⊕ B-Et	MMG → BN	2.38	**0.50**
Places [31]	BN	3.40	**0.68**
W2V matrix [31]	BN	1.92	0.90
VGG19 + Places + W2V matrix [31]	BN	**1.91**	0.88

Table 5: Accuracy@k (A@k) for different test sets (number of samples in brackets) of MMG-NewsPhoto. Textual features: BERT-Body (B-Bd), BERT-Entities (B-Et). Visual Features: CLIP_i. All models are based on the multi-task loss function unless *stl* is mentioned which stands for single task learning.

Approach	Modality	Test_{city} (660)				$\text{Test}_{country}$(794)				$\text{Test}_{continent}$ (805)	
		A@1	A@2	A@5	A@10	A@1	A@2	A@5	A@10	A@1	A@2
$base(M, f^*)$ [25]	Visual	8.3	11.2	15.9	19.8	12.7	16.9	23.2	30.1	51.1	73.7
CLIP_i (stl)	Visual	**29.1**	**38.9**	**48.5**	57.4	**61.5**	70.3	**81.0**	**88.0**	77.4	89.2
CLIP_i	Visual	27.9	37.7	**48.5**	**58.0**	**61.5**	**70.9**	80.4	85.0	**78.1**	**90.8**
T-Freq	Textual	10.5	14.1	19.2	24.5	31.1	38.4	48.0	54.3	55.8	70.8
B-Bd	Textual	27.9	38.2	49.2	60.2	69.5	76.2	84.3	88.7	85.0	92.8
B-Et	Textual	28.2	**40.5**	52.3	62.7	70.3	**79.5**	86.8	89.9	83.0	**92.8**
B-Bd ⊕ B-Et (stl)	Textual	**28.9**	39.2	50.9	**62.9**	70.4	78.0	86.0	91.1	83.6	**92.8**
B-Bd ⊕ B-Et	Textual	28.6	40.2	**52.9**	62.0	**70.8**	78.6	**87.3**	**91.2**	**84.1**	92.0
$T\text{-}base(M, f^*)$	Multimodal	27.1	36.5	43.3	44.2	62.8	74.2	79.5	80.1	75.4	86.0
VT_{CM} [26]	Multimodal	11.4	20.3	36.1	42.6	40.4	63.5	84.0	87.7	53.8	81.4
CLIP_i ⊕ B-Bd ⊕ B-Et (stl)	Multimodal	37.9	50.9	62.7	71.2	73.6	**82.2**	89.5	92.2	81.9	90.3
CLIP_i ⊕ B-Bd ⊕ B-Et	Multimodal	**39.5**	**52.1**	**64.5**	**72.7**	**73.3**	81.1	**90.1**	**92.6**	**82.9**	**92.7**

5.2 Results on MMG-NewsPhoto

Comparison of the Unimodal Models. As Table 3 shows, regarding the visual models, CLIP_i noticeably outperforms the baseline $base(M, f^*)$ [25]. Regarding the textual models, the B-Bd ⊕ B-Et surpasses the individual features. It indicates that both the contextual information and named entities and their frequencies play a vital role in the geolocation estimation of a news photo. Table 5 reports the results for Accuracy@k and shows that the CLIP_i visual model is superior at the country and continent levels, but in the city-level CLIP_i (stl) is slightly better. Among the textual models, the B-Bd ⊕ B-Et outperforms the rest at the country and continent levels, but it is not significantly better than B-Bd ⊕ B-Et (stl) in the city level.

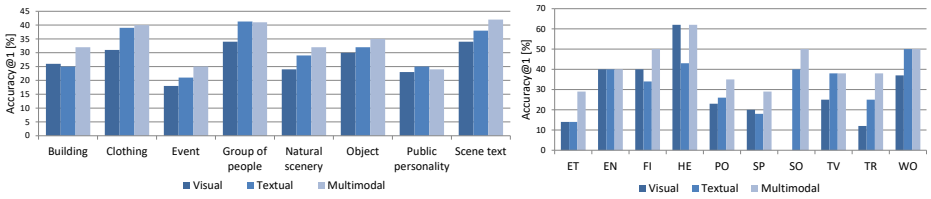

Fig. 3: Accuracy@1 [%] of the best performing visual, textual and multimodal models per concept (left) and per domain (right). ET: Entertainment, EN: Environment, FI: Finance, HE: Health, PO: Politics, SP: Sports, SO: Society, TR: Travel, TV: TV show, WO:World.

Comparison of the Multimodal Models. As presented in Table 3, the combination of the best unimodal features, $\text{CLIP}_i \oplus$ B-Bd $\oplus$ B-Et significantly outperforms all the other models in all granularity levels. Regarding Accuracy@k, Table 5 confirms the same results. For the multi-task setting, it was effective in all the granularities. In conclusion, the hierarchical information propagated from the larger granularity levels not only improves the performance in the smaller granularities, such as city but also in the country and the continent levels.

Comparison of Different Domains. Fig. 3, right presents the Accuracy@1 per domain for different models. As shown, the multimodal model outperforms in most of the domains. In domains like *Finance*, *Health*, and *Sports*, the visual model outperforms the textual model. In *TV show* and *World*, adding visual information does not help, and in the *Health* domain, additional textual information does not impact the performance.

Comparison of Different Concepts. Fig. 3, left shows the Accuracy@1 per concept (see Table 1). As presented, the proposed multimodal model outperforms the rest in all the concepts except *public personality* and *group of people*. Also, it is observed that, based on the multimodal model, the concept *event* results in the lowest, and *scene text* results in the highest performance.

Qualitative Results. Fig. 4 illustrates the results of different models. As expected, the visual model fails when there are only weak geo-representative concepts (Fig. 4 a). However, it succeeds when: (1) there is a strong concept (such as a landmark in Fig. 4 b), or (2) a weak concept occurs in high frequency, e.g., *soldier* in Fig. 4 d. The textual model fails when: (1) no relevant location is mentioned (Fig. 4 b), (2) various irrelevant entities are mentioned, e.g., *U.S.* in Fig. 4 d. As expected textual model succeeds if there are many relevant entities to the location (Fig. 4 a, c). When the text mentions many topics irrelevant to the image, the multimodal model fails (Fig. 4 d). Conversely, the multimodal model succeeds in either of the following conditions: (1) the text provides rich

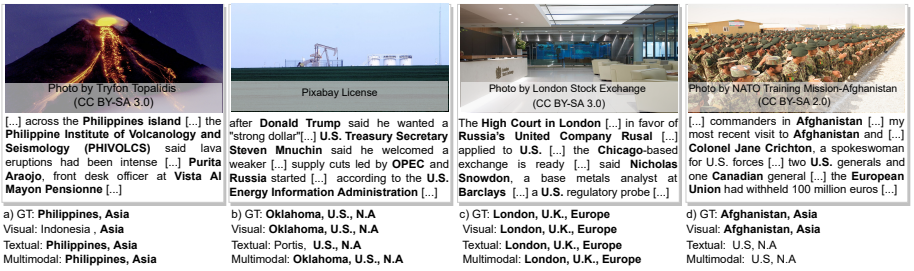

Fig. 4: Sample outputs from the *MMG-NewsPhoto* dataset with the predicted locations using best-performing textual, visual and multimodal models. Predictions written in bold are correct and correspond to the ground-truth (GT) locations. Images are replaced with similar ones due to license restrictions.

information (both in terms of entities and content) such as Fig. 4 a, c, or (2) the image illustrates strong visual concepts, such as Fig. 4 b.

5.3 Results on *BreakingNews*

Although the image locations provided by *BreakingNews* [31] can be inaccurate (discussed in Section 1), we perform experiments on the dataset for comparison. *BreakingNews* includes 33,376, 11,209, and 10,580 samples for train, validation, and testing. Ramisa et al. [31] treat the task as a regression problem where their models output the geo-coordinates. In our case, we handle the problem as a classification task to predict a specific city, country, or continent. Thus, we mapped the geo-coordinates to the closest city, country and continent classes in *MMG-NewsPhoto* based on GCD. Table 4 presents the comparison of the proposed models with *BreakingNews* (abbreviated with BN) [31] approaches. The comparison is based on the Mean and Median GCD values [31]. We evaluate our approach in two settings. In the zero-shot setting, the model was trained on *MMG-NewsPhoto* and tested on *BreakingNews* without further optimization. In the second configuration, the best model on *MMG-NewsPhoto* is both fine-tuned and tested on *BreakingNews*. The B-Bd $\oplus$ B-Et model has the lowest Median value (470 km) in the zero-shot setting and outperforms VGG19 + Places + W2V matrix [31] (880 km). In general, the comparison confirms the feasibility of applying the proposed models to unseen examples. In the second setting (MMG $\rightarrow$ BN), CLIP$_i$ $\oplus$ B-Bd $\oplus$ B-Et outperforms all the *BreakingNews* baselines by 180 to 380 km of the median value. As observed, our models perform better using the median metric, i.e., our models are better for the majority of samples.

6 Conclusions and Future Work

This paper proposes a novel multimodal approach for geolocation estimation of news photos that integrates the hierarchical information of different granularities (spatial resolutions). For this purpose, we have introduced a novel dataset

called *MMG-NewsPhoto* that contains more than half a million image-text pairs for more than 14,000 cities and 241 countries. We manually annotated 3000 samples for the evaluation to acquire different data variants at the granularity levels of city, country, and continent. We have compared our approach with several state-of-the-art approaches and baselines. Experiments showed that the combination of textual and visual features outperforms the compared models that rely only on features from a single modality. In future work, visual concepts (e.g, car plates, events, etc.), including scene text (e.g., on buildings, street signs, etc.), could be extracted for an improved geolocalization. Furthermore, the impact of photo geolocation estimation on tasks such as fake news detection, news recommendation, and cross-modal retrieval could be investigated.

Acknowledgements This work was partially funded by the EU Horizon 2020 research and innovation program under the Marie Skłodowska-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. 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. ACM (2020), https://doi.org/10.1145/3340531.3412783
2. Avrithis, Y., Kalantidis, Y., Tolias, G., Spyrou, E.: Retrieving landmark and non-landmark images from community photo collections. In: International Conference on Multimedia, MM. pp. 153–162. ACM (2010), https://doi.org/10.1145/1873951.1873973
3. Ba, L.J., Kiros, J.R., Hinton, G.E.: Layer normalization. CoRR (2016), http://arxiv.org/abs/1607.06450
4. Baatz, G., Saurer, O., Köser, K., Pollefeys, M.: Large scale visual geo-localization of images in mountainous terrain. In: European Conference on Computer Vision, ECCV. pp. 517–530. Springer (2012), https://doi.org/10.1007/978-3-642-33709-3_37
5. Berton, G.M., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Conference on Computer Vision and Pattern Recognition, CVPR. pp. 4868–4878. IEEE (2022), https://doi.org/10.1109/CVPR52688.2022.00483
6. Biten, A.F., Gómez, L., Rusiñol, M., Karatzas, D.: Good news, everyone! context driven entity-aware captioning for news images. In: Conference on Computer Vision and Pattern Recognition, CVPR. pp. 12466–12475. Computer Vision Foundation / IEEE (2019), http://openaccess.thecvf.com/content_CVPR_2019/html/Biten_Good_News_Everyone_Context_Driven_Entity-Aware_Captioning_for_News_Images_CVPR_2019_paper.html
7. Boiarov, A., Tyantov, E.: Large scale landmark recognition via deep metric learning. In: Zhu, W., Tao, D., Cheng, X., Cui, P., Rundensteiner, E.A., Carmel, D., He, Q., Yu, J.X. (eds.) Proceedings of the 28th ACM International Conference

on Information and Knowledge Management, CIKM. pp. 169–178. ACM (2019), https://doi.org/10.1145/3357384.3357956
8. Brank, J., Leban, G., Grobelnik, M.: Semantic annotation of documents based on wikipedia concepts. Informatica (Slovenia) (2018), http://www.informatica.si/index.php/informatica/article/view/2228
9. Brejcha, J., Cadík, M.: State-of-the-art in visual geo-localization. Pattern Analysis and Applications pp. 613–637 (2017), https://doi.org/10.1007/s10044-017-0611-1
10. Cheng, J., Wu, Y., AbdAlmageed, W., Natarajan, P.: QATM: quality-aware template matching for deep learning. In: Conference on Computer Vision and Pattern Recognition, CVPR. pp. 11553–11562. Computer Vision Foundation / IEEE (2019), http://openaccess.thecvf.com/content_CVPR_2019/html/Cheng_QATM_Quality-Aware_Template_Matching_for_Deep_Learning_CVPR_2019_paper.html
11. 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. ACM (2009), https://doi.org/10.1145/1526709.1526812
12. 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. Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/n19-1423
13. Hays, J., Efros, A.A.: IM2GPS: estimating geographic information from a single image. In: Conference on Computer Vision and Pattern Recognition, CVPR. IEEE Computer Society (2008)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition, CVPR. pp. 770–778. IEEE Computer Society (2016), https://doi.org/10.1109/CVPR.2016.90
15. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing (2017), https://spacy.io
16. 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. Springer (2019), https://doi.org/10.1007/978-3-030-46147-8_1
17. 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. IEEE Computer Society (2017), https://doi.org/10.1109/CVPR.2017.346
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations, ICLR (2015), http://arxiv.org/abs/1412.6980
19. Kordopatis-Zilos, G., Galopoulos, P., Papadopoulos, S., Kompatsiaris, I.: Leveraging efficientnet and contrastive learning for accurate global-scale location estimation. In: International Conference on Multimedia Retrieval, ICMR. pp. 155–163. ACM (2021), https://doi.org/10.1145/3460426.3463644
20. Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, I.: Geotagging text content with language models and feature mining. Proceedings of the IEEE pp. 1971–1986 (2017), https://doi.org/10.1109/JPROC.2017.2688799

21. 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. CEUR-WS.org (2016), http://ceur-ws.org/Vol-1739/MediaEval_2016_paper_13.pdf
22. Krippendorff, K.: Computing krippendorff's alpha-reliability (2011), https://repository.upenn.edu/asc_papers/43
23. Larson, M.A., Soleymani, M., Gravier, G., Ionescu, B., Jones, G.J.F.: The benchmarking initiative for multimedia evaluation: Mediaeval 2016. IEEE MultiMedia pp. 93–96 (2017), https://doi.org/10.1109/MMUL.2017.9
24. Mackenzie, J.M., Benham, R., Petri, M., Trippas, J.R., Culpepper, J.S., Moffat, A.: CC-News-En: A large english news corpus. In: International Conference on Information and Knowledge Management, CIKM. pp. 3077–3084. ACM (2020), https://doi.org/10.1145/3340531.3412762
25. 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. Springer (2018), https://doi.org/10.1007/978-3-030-01258-8_35
26. Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Ewerth, R.: Multimodal analytics for real-world news using measures of cross-modal entity consistency. In: International Conference on Multimedia Retrieval, ICMR. pp. 16–25. ACM (2020), https://doi.org/10.1145/3372278.3390670
27. 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 pp. 111–125 (2021), https://doi.org/10.1007/s13735-021-00207-4
28. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Fürnkranz, J., Joachims, T. (eds.) International Conference on Machine Learning (ICML). pp. 807–814. Omnipress (2010), https://icml.cc/Conferences/2010/papers/432.pdf
29. Nominatim. https://nominatim.org/release-docs/latest/api/Reverse/, accessed: 2022-05-19
30. 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. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
31. 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 pp. 1072–1085 (2018), https://doi.org/10.1109/TPAMI.2017.2721945
32. 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. Springer (2018), https://doi.org/10.1007/978-3-030-01249-6_33
33. Serdyukov, P., Murdock, V., van Zwol, R.: Placing flickr photos on a map. In: SIGIR Conference on Research and Development in Information Retrieval, SIGIR. pp. 484–491. ACM (2009), https://doi.org/10.1145/1571941.1572025
34. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: A multi-modal framework for fake news detection. In: IEEE International Conference on Multimedia Big Data, BigMM. pp. 39–47. IEEE (2019), https://doi.org/10.1109/BigMM.2019.00-44

35. Theiner, J., Müller-Budack, E., Ewerth, R.: Interpretable semantic photo geolocation. In: Winter Conference on Applications of Computer Vision, WACV. pp. 1474–1484. IEEE (2022), https://doi.org/10.1109/WACV51458.2022.00154
36. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.: The new data and new challenges in multimedia research. CoRR (2015), http://arxiv.org/abs/1503.01817
37. Tomesek, J., Cadík, M., Brejcha, J.: Crosslocate: Cross-modal large-scale visual geo-localization in natural environments using rendered modalities. In: Winter Conference on Applications of Computer Vision, WACV. pp. 2193–2202. IEEE (2022), https://doi.org/10.1109/WACV51458.2022.00225
38. Trevisiol, M., Jégou, H., Delhumeau, J., Gravier, G.: Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In: International Conference on Multimedia Retrieval, ICMR. pp. 1–8. ACM (2013), https://doi.org/10.1145/2461466.2461468
39. 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. ijcai.org (2019), https://doi.org/10.24963/ijcai.2019/502
40. Vo, N.N., Jacobs, N., Hays, J.: Revisiting IM2GPS in the deep learning era. In: International Conference on Computer Vision, ICCV. pp. 2640–2649. IEEE Computer Society (2017)
41. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM pp. 78–85 (2014), https://doi.org/10.1145/2629489
42. 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. IEEE (2020), https://doi.org/10.1109/CVPR42600.2020.00265
43. Weyand, T., Kostrikov, I., Philbin, J.: Planet - photo geolocation with convolutional neural networks. In: European Conference on Computer Vision, ECCV. pp. 37–55. Springer (2016), https://doi.org/10.1007/978-3-319-46484-8_3