MLM: A Benchmark Dataset for Multitask Learning with Multiple Languages and Modalities

Jason Armitage*
University of Bonn
Germany

Endri Kacupaj*
University of Bonn
Germany

Golsa Tahmasebzadeh*
TIB – Leibniz Information Center for Science and Technology
Germany

Swati
Jožef Stefan Institute
Slovenia

Maria Maleshkova
University of Bonn
Germany

Ralph Ewerth
TIB – Leibniz Information Center for Science and Technology
Germany

Jens Lehmann
University of Bonn
Fraunhofer IAIS
Germany

ABSTRACT

In this paper, we introduce the MLM (Multiple Languages and Modalities) dataset - a new resource to train and evaluate multitask systems on samples in multiple modalities and three languages. The generation process and inclusion of semantic data provide a resource that further tests the ability for multitask systems to learn relationships between entities. The dataset is designed for researchers and developers who build applications that perform multiple tasks on data encountered on the web and in digital archives. A second version of MLM provides a geo-representative subset of the data with weighted samples for countries of the European Union. We demonstrate the value of the resource in developing novel applications in the digital humanities with a motivating use case and specify a benchmark set of tasks to retrieve modalities and locate entities in the dataset. Evaluation of baseline multitask and single task systems on the full and geo-representative versions of MLM demonstrate the challenges of generalising on diverse data. In addition to the digital humanities, we expect the resource to contribute to research in multimodal representation learning, location estimation, and scene understanding.

1 INTRODUCTION

At the core of recent research in machine learning is the proposal that systems trained on large sets of data will display forms of the human ability to generalise. Learning algorithms will enable systems to infer abstract representations from raw data and tackle a wide range of tasks [7]. The web provides an abundance of data for training models that enable functions in application pipelines. Data availability then is no impediment to the aims of machine learning in relation to these applications - but there remain two important ways in which systems fail to generalise.

1. The first of these is the inability for systems to cope with diversity in real-world data [6][53]. Even small changes in the distributions of data lead to steep declines in the performance of machine learning models. Information on the web and maintained in archives is present in a diverse range of modalities including language, images, and numerical data. Variety is also observed in each modality taken in isolation. Consider text in documents stored and studied by cultural and academic institutions. A system that is designed to predict contextual information - in this case, locations referenced in the documents [1] - may encounter text written in several languages and writing systems. Additional inputs with relevant information can be stored in databases or as structured text in a knowledge base. Machine learning systems that are robust to disparities between and within modalities can make use of a higher proportion of the information present across a range of materials [5].

The second way in which current systems fail to generalise is in using representations learned during training to perform new tasks. Analysis of machine learning in application workflows reveals a host of disparate models - each trained on homogeneous data to conduct specialised tasks [41]. In this approach, learning is siloed with the result that computation and data required for training and inference multiply - and there is limited sharing of parameters and representations across pipelines. Systems that conduct multiple tasks hold the promise of simplifying application architectures and improving efficiency in the use of resources. In the case of researchers conducting cultural and historical analysis, this also presents the advantages of studying or accessing materials in multiple formats [40] such as textual documents and images that reference the same subject.

The Multiple Languages and Modalities (MLM) dataset consists of text in three languages (English, French, and German), images, location data, and knowledge graph triple classes (ie component entities of triples for the sample entity that are stored in Wikidata). The resource is designed to evaluate multitask systems in relation to single task alternatives when trained simultaneously on sequences of tasks - in this case, a system that performs both

©2020 Copyright held by the authors. This is the author’s version of the work. It is posted here for your personal use. Not for redistribution. The definitive version was published in the Proceedings of the 29th ACM International Conference on Information and Knowledge Management. https://doi.org/10.1145/3340531.3412783
cross-modal retrieval and location estimation. Multitask learning aims to benefit from commonalities in input data [13] and MLM is derived from knowledge graph entities that connect all of the samples in the data. Relationships between entities are further manifested in triple classes.

1.1 Contributions

We identify four contributions resulting from this research.

- The first contribution is the generation process, which provides a consistent and reproducible approach to build a resource of diverse data with semantic relationships. As detailed below, this process will accelerate the release of future versions of MLM.
- The second contribution is the release of a novel dataset with a range of modalities and languages that - to our knowledge - is unique in the research literature. The resource tests the ability for multitask systems to exploit both formal and indirect relationships between samples.
- The third contribution is the specification of a benchmark series of tasks to evaluate multitask systems that characterise applications used and built by researchers and developers in the digital humanities. The resource presents a unique option to this community to develop novel applications that make use of multimodal frameworks for the assessment of cultural and historical phenomena [40].
- The fourth contribution is building a multitask system that combines several methods to perform in concurrence cross-modal retrieval and location estimation on highly diverse data.

1.2 Use Case

We present a use case that applies to the benchmark evaluation system. Multitask learning systems are characterised by homogeneity both in modalities and languages. The Edinburgh Geoparser [1] and Perdido [37] are recent instances of applications that retrieve locations from multilingual literary texts sourced in books and journals. Identifying locations depicted in visual materials is also a core function in applications that georeference historical images on landscapes and urban areas [9,20]. In the multitask system framework presented below (see Figure 1), we enable geoparsing and georeferencing in two novel ways. First the researcher has the option of starting a query either with a text or an image. During the retrieval task (see Figure 1), related materials in the missing modality are returned. This creates the opportunity to conduct a multimodal analysis on the entity of interest [40]. Second a multitask framework returns a predicted location for both inputs.

Figure 1: Cross-modal retrieval and location estimation performed by the Multitask IR+LE framework.

2 IMPACT

In this section, we highlight research areas where the MLM resource enables innovation and assess the benefits of the dataset in relation to existing resources for evaluating multitask systems. We have outlined above new approaches in the digital humanities that are enabled by the dataset and benchmark evaluation tasks. Multitask learning systems that learn on multimodal data are also an active area of research in relation to multimodal representation learning, location estimation, and scene understanding [5] [28]. MLM is further designed to evaluate the ability for multitask systems to leverage relationships between constituent entities in data and knowledge graph properties used in the generation process. Multitask learning systems that exploit these relationships as a signal are positioned to deliver additional benefits to applications that rely on semantic data and knowledge graphs, which include recommender systems, mobile information retrieval, and bioinformatics [2].

Resources for training and evaluating multitask learning systems are characterised by homogeneity both in modalities and languages. The Office-Home and Computer Survey are examples of resources to train text-only systems [55]. Similarly multitask systems for computer vision turn to benchmark datasets where images are the primary modality [12,32]. Multimodal datasets employed for multitask learning consist predominantly of two modalities [30] [16] and a single language. Although ImageCLEF runs challenges (see Table 1) with textual samples in multiple languages, these instances also focus on bimodal learning [49]. The Multiple Language and Modalities dataset, which we propose, approximates the diversity of online data and those stored in digital archives or databases in presenting samples in three modalities with multilingual text and semantic data. To our knowledge, this resource is unique in presenting an opportunity to train multitask systems on visual, textual, spatial and semantic web inputs for multimodal, and multilingual learning.

Geographical location is central to both the resource and task presented in this research. We are motivated by on-going investigations on representation bias in machine learning datasets to present a second geo-representative version of our resource. Representation bias has been found to penalise minority classes...
in classification tasks [45]. MLM-irle-gr is a subset of the full data with balanced sample sizes of entities for the 28 countries in the European Union. This enables governmental and commercial organisations that provide applications and services in this region to minimise the impact of representation bias resulting from training data when evaluating multitask learning systems.

3 MLM - MULTIPLE LANGUAGES AND MODALITIES

In this research we propose a benchmark to evaluate the strengths of multitask systems to generalise on diverse data. This section starts by defining the tasks constituting the evaluation, continues with an introduction of the dataset, and concludes with details on the generation process.

3.1 Benchmark Evaluation

The benchmark evaluation is composed of two tasks that are performed in conjunction to retrieve and locate entities represented by multimodal and multilingual inputs. The first task consists of cross-modal retrieval where a representation in an input modality \( x_i \) returns the target \( y_i \) in a corresponding modality. In this case, the inputs are one from a pair \( P = \{(x_i^D, x_i^C)\}_i \) = 1, where \( x_i^D \) is an image and \( x_i^C \) is a combination of structured \( x_i^T \) and unstructured \( x_i^L \) textual inputs. The objective is to return the corresponding target \( y_i \) at the lowest position in the ranking of returned predictions. The objective of the second task is to use the same pair of inputs \( P = \{(x_i^D, x_i^C)\}_i \) = 1 to perform a classification on a set of geo-cells \( y = (y_1, y_2, \ldots, y_n) \) - each of which groups a set of entities based on location given as geo-coordinates. The final objective of the evaluation is to return results that surpass the baseline system presented below on both tasks. Additional details of the constituent tasks are included in Section 3.1.

3.2 Dataset

The Multiple Languages and Modalities comprises data points on 236k human settlements for evaluating and optimising multitask learning systems. MLM presents a dataset with a high level of diversity in terms of modality and language. For each entity, we have extracted text summaries, images, coordinates, and their respective triple classes. Text summaries are available in three languages (English, French, and German) with each entity having between one and three language entries.

Human settlements from all continents are provided in the full dataset with 72% located in Europe (see Figure 2). MLM-irle is a version of the dataset generated for the benchmark evaluation tasks and features geo-cells required for location estimation. To serve organisations that focus on the European Union, we have also created a version of the dataset - MLM-irle-gr (ie georepresentative) - that provides a geographically balanced coverage of human settlements in this region. MLM-irle-gr contains data on 24k human settlements across the EU weighted in relation to the population count [3] for each of the 28 countries.

Figure 2: MLM: Number of Entities by Continent (000).

3.3 Generation Process

In this first release, a core aim of the research was to develop a generation process that enables fast and consistent scaling for future versions. The process outlined also results from a core requirement for the resource: to evaluate the ability of multitask systems to incorporate information on entities with semantic relations. In this section, we detail stages in the framework for dataset generation. Our selection of the Wikidata Knowledge Graph as the primary source was conditioned on the requirement for data with semantic relations between entities and the ability enabled by SPARQL to target data that met our research aims.

![Figure 3: MLM dataset generation process.](https://population.un.org/wpp/Download/Standard/Population/)
3.3.1 Define base entity type. The first step consisted of defining the base entity type that best supported developing applications that make use of geographical location. We selected the Wikidata entity human settlement (ID Q486972) based on the number of entities returned relative to other candidate base entity types. In the second step, instances of human settlement base type with targeted modalities were extracted resulting in over half a million entities. We applied constraints to extract entities with a Wikipedia article in at least one target language (EN, DE, FR), a minimum of one image, and a set of geocoordinates resulting in 40k human settlements.

3.3.2 Generate a list of classes. We increased the scope and granularity of human settlements by extracting entity types that are subclass of the the base type. Entities are returned for triples up to three hops from the starting entity. An example of additions is the commune of Italy (ID Q747074) - a subclass of the entity municipality (ID Q15284), which is a subclass of the human settlement base type. This stage resulted in a set of 910 unique classes where all entities derive from the base type by either one, two, or three hops.

3.3.3 Extract entities. In the extraction stage, we run queries on all entities that are an instance of at least one of the 910 unique classes in the generated list. This returned a set of 2.5M entities. A final application of the constraints related to targeted data types generated a collection of data that constituted 240k entities.

3.3.4 Download images and summaries. The final stage consists of downloading images and text summaries for the extracted entities. While querying Wikidata we were able to also extract Wikimedia Commons links of images and the corresponding Wikipedia link for each entity. Text summaries were downloaded using the MediaWiki API while images using the WikiMedia link. The process was accelerated by chunking the data and running multiple sessions. Download errors led to a small loss of entities and the final raw dataset is in Table 2.

3.3.5 Geographical cell generation. In this stage, entities along with their images and summaries are assigned to geographical cells. To this end, we first exclude all SVG images since they mostly contain maps and signs which are non-descriptive for our tasks. Next, we divide the earth into non-overlapping geographical cells C using the S2 Geometry Library. In this library, initial cells are generated by projecting the earth’s surface on a cube with six sides. Based on geocoordinates of the entities, a hierarchical subdivision is applied, where each cell is the node of a quad-tree. Starting at the root nodes, the respective quad-tree is subdivided recursively until all cells contain a maximum of \( T_{\text{max}} = 500 \) entities. Finally, in order to have more uniform cells relative to the entities, all generated cells containing less than \( T_{\text{min}} = 20 \) entities are removed.

4 AVAILABILITY

Versions of MLM listed in Table 2 are available for direct download and use. To support findability and sustainability, the MLM dataset is published as an on-line resource at https://doi.org/10.5281/zenodo.3885753. A separate page with detailed explanations and illustrations is available at http://cleopatra.ijs.si/goal-mlm/ to promote ease-of-use. The project GitHub repository contains the complete source code for the system and generation script is available at https://github.com/GOALCLEOPATRA/MLM. Documentation adheres to the standards of FAIR Data principles with all relevant metadata specified to the research community and users. It is freely accessible under the Creative Commons Attribution 4.0 International license, making it reusable for almost any purpose.

4.1 Updating and Reusability

MLM is supported by a team of researchers from the University of Bonn, the Leibniz Information Center for Science and Technology, and Jožef Stefan Institute. The resource is already in use for individual projects and as a contribution to the project deliverables of the Marie Skłodowska-Curie CLEOPATRA Innovative Training Network. In addition to the steps above that make the resource available to the wider community, usage of MLM will be promoted to the network of researchers in this project. Awareness among researchers and practitioners in digital humanities will be promoted by demonstrations and presentations at domain-related events. The range of modalities and languages present in the dataset also extends its application to research on multimodal representation learning, multilingual machine learning, information retrieval, location estimation, and the Semantic Web. MLM will be supported and maintained for three years in the first instance. A second release of the dataset is already scheduled and the generation process outlined above is designed to enable rapid scaling.

5 MULTITASK LEARNING FRAMEWORK

In this section, we specify a multitask learning framework that characterises the use case above by performing cross-modal retrieval and location estimation tasks. The section starts with formulations of the tasks evaluated on MLM and concludes with details of the system and methods.

5.1 Tasks

5.1.1 Cross-modal Retrieval. We build on the approach proposed by Marin et al to perform cross-modal information retrieval on visual u and textual v inputs. Embeddings \( u \) and \( v \) are learned as described below. A mapping from the visual and textual embeddings to a common space is learned using a cosine embedding loss, which minimises the distance between \( u \) and \( v \).
We apply state-of-the-art embedding methods in computer vision where (described in Section 3.3.5) are regarded as classes. To formulate labels, and was introduced for the MediaEval Placing Task 2016 includes around five million geo-tagged images with geolocation Commons 100 Million dataset (YFCC100M) [47]. The subset model is pretrained on a subset of the Yahoo Flickr Creative Places365 dataset [57], where the task is to recognize 365 distinct descriptors is taken from the model [38] based on ResNet101 [21] aimed at predicting the geolocation information of an image. This descriptor is taken from the model [38] based on ResNet [21] model pre-trained on the Images when generating Visual embeddings descriptors are utilized to better represent location information of

\[ L_{cos}(v^n, v^t, l) = \begin{cases} 
1 - \cos(v^n_i, v^t_i), & \text{if } s = 1 \\
\max(0, \cos(v^n_i, v^t_i) - \text{margin}), & \text{if } s = -1
\end{cases} \]  

(1)

where \( v^n \) and \( v^t \) are representations of the visual and textual inputs respectively. At inference time, an image or textual sample is presented with the aim of returning samples for the corresponding modality.

5.1.2 Location estimation. In this architecture, location estimation is regarded as a classification problem. To this end, geo-cells (described in Section 3.3.5) are regarded as classes. To formulate the location estimation task, we have \( q \) geo-cells \( y = (y_1, y_2, \ldots, y_q) \) for each of which there are \( n \) number of entities represented by images, summaries and triples. Each entity is labeled with its corresponding GPS coordinates in (latitude, longitude) pairs. The task is to predict \( y \) given \( v^n \) and \( v^t \) based on a cross-entropy loss:

\[ L(v, y) = -v[y] + \log \left( \sum_j \exp(v[j]) \right) \]  

(2)

5.1.3 Multitask. The multitask learning framework (see Figure 4) trains a combined pipeline to perform cross-modal retrieval and location estimation in conjunction. The system learns over the cross-entropy and cosine embedding losses with a weighted average. In order to account for the difference in magnitude between losses, log variance is calculated over the tasks and converted to log standard deviation \( \log(\sigma) \). This weight is applied to the losses before computing the mean \( \mu \) over the products.

5.2 System

We apply state-of-the-art embedding methods in computer vision and NLP to generate representations of the inputs. Two types of descriptors are utilized to better represent location information of the images when generating Visual embeddings \( v^n \). One of the descriptors is taken from a ResNet [21] model pre-trained on the Places365 dataset [57], where the task is to recognize 365 distinct places - examples include beach, stadium, street. The second descriptor is taken from the model [38] based on ResNet101 [21] aimed at predicting the geolocation information of an image. This model is pretrained on a subset of the Yahoo Flickr Creative Commons 100 Million dataset (YFCC100M) [47]. The subset includes around five million geo-tagged images with geolocation labels, and was introduced for the MediaEval Placing Task 2016 (MP-16) [29]. Each descriptor results in a feature vector of size 2048, so eventually we concatenate both vectors and get a vector of size 4096 for each image. To generate textual embeddings \( v^t \) for summaries and triples we use BERT [19] to extract word vectors for all sentences in text. Since text content of news articles is long, we select maximum 500 characters at a time and extract word vectors from BERT, and average the vectors obtained from each token of the sentence.

An integrated architecture (see Figure 3) takes the visual and text embeddings to perform the two tasks in concurrence. Visual embeddings \( v^n \) are fed into a 1D convolutional layer (conv1d) with ReLU activation \( \max(0, x) \) - and a linear layer with hyperbolic tangent \( \tanh(u^t) \). An LSTM layer replaces the conv1d for text summary \( v^t \) and triple class \( v^t \) embeddings. Regularisation consists of dropout for textual embeddings and norm with a clamp \( \min = 1 + 10^{-12} \) for all modalities. Linear layers with dropout receive the outputs of the norm in location estimation. In both tasks, text summary and triple classes are fused with concatenation \( v^t = [v^n, v^t] \).

6 EXPERIMENTS

Experiments assess the performance of the multitask learning system in relation to single task pipelines. We align with benchmark cross-modal retrieval tasks [35] in selecting median rank (medR) as the primary metric. Median rank also takes into account both the performance of the retrieval and the position of the retrieved instance. To compute medR, a subset of 500 instances are randomly sampled from the test set. Then the experiments are repeated 10 times and report the mean results. Next, for each input image or text, items in the corresponding modality are ranked.

The objective is to retrieve the sample for the target modality at the highest rank. The retrieved lists are further assessed using recall percentage at top K (R@K). This measures the percentage of queries where the matching item is ranked among the top K returned samples. The formulation of location estimation as a multiclass classification task motivates the selection of Precision (P) (Eq.3a), Recall (R) (Eq.3b) and weighted F1 score (Eq.3c) as measures of performance for the second task.

\[ \text{Precision} = \frac{TP}{TP + FP} \]  

(3a)

\[ \text{Recall} = \frac{TP}{TP + FN} \]  

(3b)

\[ F_1^{\text{weighted}} = \frac{1}{Q} \sum_{j=1}^{Q} \frac{Q_j^2}{P_j + R_j} \]  

(3c)
In the above definitions, $TP$, $FP$ and $FN$ stand for True Positive, False Positive and False Negative respectively and $Q$ is the number of geo-cells, and $j$ is the $j$th instance of geo-cells.

6.1 System Configuration and Training

Pipelines for single task and multitask learning are trained for 100 epochs on a single GeForce GTX 1080 Ti GPU, using training/validation sets of 175k/22k and 18k/2k for MLM-irle and MLM-irle-gr respectively. Full details on the dataset splits are available at [http://cleopatra.ijs.si/goal-mlm/](http://cleopatra.ijs.si/goal-mlm/). Although we trained for a number of epochs adopted for comparable systems [55], we note that validation loss was still declining for both multitask and single task systems at the end of training. For all tasks, consistent hyperparameter settings are maintained to enable comparative analysis of the results. The batch size is set at 128 with a learning rate of 0.0001 and a probability of 0.1 in dropout layers.

6.2 Results

In this section evaluation results are discussed for multitask and single task pipelines on the MLM-irle and MLM-irle-gr (ie geo-representative) datasets containing 22k and 2k test instances respectively.

6.2.1 Cross-modal Retrieval In Table 3, we report medR and Recall@K for image to text (first four columns) and text to image retrieval (the remaining columns). The multitask system attained a strong improvement on the single task equivalent in the first of these with medR at 19.9. Qualitative checks on samples consisted of extracting a random sample of 500 entities from the test set. Two examples of triple classes where the multitask system scored strongly relative to the single equivalent are “town of the United States” and “Ortsteil”. This corresponds to increases of 13% on R@5 and 20% points on R@10. Although R@1 also showed a marked increase, scores for both multitask and single task are at low levels. The multitask system also outperformed strongly on the geo-representative dataset with a 26-point improvement on medR relative to the single task pipeline. Multitask and single task systems attained stronger scores on text to image retrieval for both datasets with double the R@K values and medR at 11.9 - underlining the superiority of the multitask approach in cross-modal retrieval.

6.2.2 Location Estimation A summary of performance for multitask and single task systems on location estimation is presented in Table 4. Results are broken down by performance on predicting geo-cells using images and text consecutively. As with cross-modal retrieval, both systems return stronger scores when inputs are textual relative to visual. A notable difference with the first task is that performance improves on the geo-representative dataset. This difference is most marked in image-based estimation: the F1 score for the multitask system is 10 points higher on MLM-irle-gr relative to the larger dataset. Scores for the multitask system on both image- and text-based location estimation are closely aligned with the single task system - indicating there are only modest gains from using the former approach on the second task.

7 RELATED RESEARCH

In the first instance, researchers designing multitask systems that learn from multimodal data turn to existing multimodal resources for benchmarking. A reliance on image captioning datasets has led to multitask learning proposals limited to paired image and text samples [8, 54]. Datasets in the medical and bioinformatics areas that have inspired multitask approaches also use these two modalities [39, 27]. Several proposals pair natural language and knowledge graph embeddings for training multitask systems [33]. In contrast to our research, tasks in this area are concentrated on applications related to NLP and knowledge graphs - and use text-only data. Multitask learning on multimodal data is an area of research in robotics [11, 44] - although the objective of training autonomous agents is distinct from the use cases for MLM. In all of the cited research, text inputs are English-only. Multilingual resources and benchmarks for training multitask architectures are motivated primarily to solve NLP tasks and are correspondingly unimodal [22, 31]. A multitask solution was proposed for the multilingual IKEA dataset with a specific objective in machine translation [58]. As with previously noted cases, this resource is bimodal and so applies to a smaller subset of constituent tasks relative to MLM.

Research on information retrieval for multimodal data has an extensive history and we focus here on resources and benchmark tasks where samples are present in multiple languages. The Cross-Language Evaluation Forum (CLEF) evolved into ImageCLEF (from 2003), GeoCLEF (from 2005) and VideoCLEF (from 2008) tracks. Resources in the LifeCLEF tracks of ImageCLEF have included geocoordinates - but the focus is primarily on images and is aimed at researchers in disciplines related to biology [23]. Information retrieval on video content has extended to multilingual scenarios where either queries, content, or both are in several languages [3]. Datasets in this area are mostly composed of comparatively small numbers of samples. Constituent modalities (ie video, audio) vary from those in MLM and place an emphasis on the spatiotemporal context of the inputs [51]. Video retrieval is also differentiated by an emphasis on domain-specific sub-tasks including shot segmentation [18] and detection of objects or scenes [51]. All of these resources exclude semantic Web elements.

Location estimation methods related to this research breakdown into three main groups [10]: natural [4, 56], city-scale [24, 14] and global [25, 39]. Several methods in the published research rely on a single modality [52, 58, 36] and we focus here on multimodal approaches. MediaEval benchmark placing datasets [29, 15] include more than five million instances with images, videos and metadata used to estimate capturing locations represented in multimedia. Kordopatis-Zilos et al. [25] build on this dataset by combining visual features with a language model based on word frequency in the text for geolocation estimation. In [50], a dataset
We are motivated by this research to address representation and when the test data fails to match the population of interest [45].

A baseline multitask system that makes modalities and languages. Abaselinemultitasksystemthatmakes geoparsing and georeferencing information encoded in multiple parts of the machine learning pipeline, evaluation bias results samples for classes that are under-sampled in the data. In another approach that use geolocation information for image recognition [45, 17] and image understanding [54]. In [54] the authors leverage season and location context with a probabilistic framework to help improve region recognition in images. In [17] the effectiveness of using geolocation on fine-grained recognition is examined. All the aforementioned methods are based on datasets in English only. All these resources are intended for single task scenarios for the corresponding task. In contrast, this paper provides a multilingual dataset with a great number of modalities. The objective also differs in the focus on systems that perform geolocation estimation as one of a sequence of tasks.

Research on different forms of bias in data - and their impact on machine learning systems - underlines the value of resources containing image and text is collected from Wikipedia combined with around 890K satellite images and is used for interpreting satellite images. BreakingNews [42] is a resource with 100K instances in the domains of sports, politics, arts, healthcare or local news. The researchers apply a baseline that combines a Convolutional Neural Network and Recurrent Neural Networks for multimodal geolocation estimation. In addition there are other approaches that use geolocation information for image recognition [45, 17] and image understanding [54]. In [54] the authors leverage season and location context with a probabilistic framework to help improve region recognition in images. In [17] the effectiveness of using geolocation on fine-grained recognition is examined. All the aforementioned methods are based on datasets in English only. All these resources are intended for single task scenarios for the corresponding task. In contrast, this paper provides a multilingual dataset with a great number of modalities. The objective also differs in the focus on systems that perform geolocation estimation as one of a sequence of tasks.

9 ACKNOWLEDGMENTS

The project leading to this publication has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES


8 CONCLUSION AND FUTURE WORK

In this research, we introduced the MLM resource for evaluating multitask systems on diverse data. In addition, we specified a set of benchmark tasks that characterise the specific use case of geoparsing and georeferencing information encoded in multiple modalities and languages. A baseline multitask system that makes use of both cross-modal and knowledge fusion was specified and evaluated against single task systems. The paper details a process that generates a resource comprising linked entities. The generation process also enables consistent and rapid scaling of the resource for future releases. Finally we also ensured that the dataset meets FAIR principles and released a geo-representative version of the dataset to serve cultural and academic institutions with a focus on the European Union. We believe that the resource will improve generalisation in multitask learning systems and so streamline the pipelines of applications that learn from materials stored in digital archives. In upcoming research, we will release future versions with data in additional languages, add entries from other data sources, design systems for specific functions, and develop an application that further exploits contextual data.


