Multimodal Entity Tagging with Multimodal Knowledge Base

Hao Peng\textsuperscript{1}, Hang Li\textsuperscript{2}, Lei Hou\textsuperscript{1}, Juanzi Li\textsuperscript{1}, Chao Qiao\textsuperscript{2}

\textsuperscript{1}Department of Computer Science and Technology, Tsinghua University, Beijing, China
\textsuperscript{2}ByteDance AI Lab, Beijing, China
peng-h21@mails.tsinghua.edu.cn

Abstract

To enhance research on multimodal knowledge base and multimodal information processing, we propose a new task called multimodal entity tagging (MET) with a multimodal knowledge base (MKB). We also develop a dataset for the problem using an existing MKB. In an MKB, there are entities and their associated texts and images. In MET, given a text-image pair, one uses the information in the MKB to automatically identify the related entity in the text-image pair. We solve the task by using the information retrieval paradigm and implement several baselines using state-of-the-art methods in NLP and CV. We conduct extensive experiments and make analyses on the experimental results. The results show that the task is challenging, but current technologies can achieve relatively high performance. We will release the dataset, code, and models for future research.

1 Introduction

Multimodal knowledge base (MKB) or multimodal knowledge graph (MKG) is an important area for AI technologies because humans’ information processing is inherently multimodal. For example, it is believed that humans learn and utilize concepts such as “Eiffel Tower” through the processing of multimodal data (Bergen, 2012). Construction and utilization of MKB both need to be intensively investigated. In this paper, we propose a new task named Multimodal Entity Tagging (MET) with an MKB and study the problem empirically.

Suppose that we have an MKB containing a vast number of entities and each entity has a large number of texts and images associated. Given a new pair of text and image, MET is to identify the entity described in the given text-image pair with an MKB as shown in Figure 1. The task is crucial, we believe, as a step in multimodal information processing. Note that recognizing whether an image consists of an entity is still a challenging problem in CV (Joseph et al., 2021), and here we expect that in addition to an image, its paired text is also used. This is the first work on the issue, as far as we know.

MET is very challenging because of the diversity, sparsity and ambiguity in MKB: 1) Texts and images are diverse. 2) Many entities contain limited information. 3) An image (or a text) may associate with multiple entities. We perform the task by using the information retrieval paradigm. Given a pair of text and image, candidates (entities in MKB) are first retrieved using retrieval models. Then we utilize intra-modality and inter-modality matching models to rank the candidates. We create a dataset of MET from a large MKB called VisualSem (Alberts et al., 2020) and employ state-of-the-art technologies in NLP and CV to conduct extensive experiments on the dataset. We demonstrate to what extent the existing methods work and provide a foundation for future research on the problem.

2 Task and Dataset

2.1 Task Definition

The task is to identify the entity described in a given text-image pair with an MKB, where entities are depicted in texts and images in the MKB. Formally, suppose that there is a multimodal knowledge base having $k$ entities $\mathcal{K} = \{e_1, e_2, \ldots, e_k\}$. Each entity $e_i$ is associated with $i_t$ texts and $i_m$ images, $\mathcal{E}_i = \{t_1^{(i)}, t_2^{(i)}, \ldots, t_{i_t}^{(i)}\}, \{v_1^{(i)}, v_2^{(i)}, \ldots, v_{i_m}^{(i)}\}$ where $t_k^{(i)}$ and $v_k^{(i)}$ denote a text and an image respectively. Further suppose that there is a pair of text and image $(t, v)$. MET aims to find the corresponding entity $e_i \in \mathcal{K}$ that is described by $(t, v)$. Here, we expect that a model is automatically learned from the MKB, and the task is performed
The capital and largest city of France.

Paris is the capital and most populous city of France.

The international center of culture and commerce of France.

A wrought iron tower 300 meters high that was constructed in Paris in 1889.

The Eiffel Tower is a wrought-iron lattice tower on the Champ de Mars in Paris, France.

An iron tower beside the River Seine in Paris which was built on the occasion of the Exposition Universelle in 1889.

Figure 1: Multimodal entity tagging with multimodal knowledge base. An input contains an image and its text description, such as ”An iron tower in Paris”. An MKB contains a vast number of entities which have various images and texts associated. We decide whether the input text and image pair describes the entity “Eiffel Tower”, by taking it as query, and retrieve relevant entities in the MKB, and ranking the candidate entities.

There are several challenges for the task regarding learning and utilization of the model. First, the scale of the MKB is large, and the content of the MKB is diverse. As shown in Figure 1, images associated with entity “Paris” can be landmark buildings, city flag, city emblem, and map. Diversity of the data poses challenges to multimodal information understanding and utilization. Second, the information might be insufficient for identifying the entities. For example, many entities in the MKB contain a few images and texts associated. Third, the data might also be ambiguous. For example, in Figure 1, the same input image contains multiple entities (e.g., “Eiffel Tower” and “Paris”) which may confuse the model. Section C in appendix provides more details.

### 2.2 Dataset Creation

We derive a new dataset from the multimodal knowledge base VisualSem (Alberts et al., 2020). VisualSem contains 101,244 entities, and each entity has on average 15.2 images and 2.9 glosses (texts) associated. There is no need to annotate data manually as MKB is naturally “labeled” data. We filter out entities that have less than three glosses or images. We split the data into knowledge-base, training, development, and test sets. We ensure that the entities in the knowledge-base, training, development, and test sets do not have common glosses or images, and thus there is no “information leak”. Each instance in the training, development, and test sets consists of a randomly combined pair of gloss and image of an entity. Table 1 shows the statistics of the data.

### 3 Method

We view MET with MKB as an information retrieval problem. The input is a text-image pair \((t, v)\) as the query. 1) **Retrieval** We first separately retrieve the most relevant \(N\) texts and their entities with text \(t\), and the most relevant \(M\) images and their entities with image \(v\). Thus, there are at most \(N+M\) entities. Next, each retrieved text/image is paired with some image/text of the same entity.¹ 2) **Ranking** We then rank the entities based on the relevance between the query \((t, v)\) and the retrieved texts and images. There are multiple matching scores between the query and the retrieved texts and images, and all of them are taken as features of the ranking model. Finally, the top-ranked entities are selected as output.

#### 3.1 Retrieval

**Text Retrieval** We employ the elastic search system² to conduct term-based text retrieval with the

---

¹We give details of this procedure in section B in appendix.

²https://www.elastic.co/
text \( t \) as query. Once the most relevant \( N \) texts \( \{t_{j_1}^{(i_1)}, \ldots, t_{j_N}^{(i_N)}\} \) are retrieved, the corresponding \( N \) entities \( \{e_{i_1}, \ldots, e_{i_N}\} \) are obtained and \( N \) images \( \{v_{j_1}^{(i_1)}, \ldots, v_{j_N}^{(i_N)}\} \) are selected to pair with retrieved texts.\(^2\)

**Image Retrieval** We utilize a ResNet152 model (He et al., 2016) to encode all images as real-valued vectors of dimension 2048. We employ the nearest neighbor search technique\(^3\) to perform vector-based image retrieval with the image \( v \) as query. Once the most relevant \( M \) images \( \{u_{k_1}^{(l_1)}, \ldots, u_{k_M}^{(l_M)}\} \) are retrieved, the corresponding \( M \) entities \( \{e_{l_1}, \ldots, e_{l_M}\} \) are obtained and \( M \) texts \( \{t_{k_1}^{(l_1)}, \ldots, t_{k_M}^{(l_M)}\} \) are selected to pair with retrieved images.\(^2\)

### 3.2 Ranking

The most relevant entities are ranked at the top based on the relevance between the query and the retrieved texts and images. The ranking model is a linear combination of matching scores where the weights are tuned using the development data. The matching scores are calculated with different matching models. Formally, \( e_o = F((t, v, (t_o^p, v_o^p))) \), \( e_o \) donates the matching score and \( F \) denotes the matching model.

**Intra-modality Matching** 1) Text Bi-encoder Matching (TBM). The bi-encoder model transforms the query text and a retrieved text into their representations with two encoders and calculates the similarity (relevance) between the representations (Wu et al., 2020; Thakur et al., 2021). The bi-encoder model is usually more efficient. We train the two encoders by fine-tuning a BERT model (Devlin et al., 2019). The two encoders share tied parameters as in Reimers and Gurevych (2019). 2) Text Cross-encoder Matching (TCM). The cross-encoder transforms the concatenation of the query text and a retrieved text into a representation with only one encoder (Wu et al., 2020), and decides the matching degree (relevance) between the two texts. The cross-encoder model is more accurate, because interactions between the two texts are captured in the encoder. We train the encoder by fine-tuning the BERT model. 3) Image Bi-encoder Matching (IBM). The “image bi-encoder matching” model has two encoders (tied parameters), one for encoding the query image and the other for encoding a retrieved image. It uses cosine to represent the similarity (relevance) between the representations from the two encoders. We implement each of the two encoders using ResNet152 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009).

**Inter-modality Matching** CLIP is an image-text matching model proposed by Radford et al. (2021). CLIP is pre-trained on 400 million image text pairs and demonstrates strong performances on downstream image classification tasks, especially in few-shot or zero-shot settings. We adopt CLIP\(^4\) as a method for inter-modality matching.

### 4 Experiments

#### 4.1 Experimental Settings and Results

We conduct experiments to investigate the hardness of the MET problem as well as the performance of the methods described above. We use Hits@N as evaluation measure, which is the percentage of correct entities at the top N positions. In our experiments, we set \( M = N = 100 \). We use grid search to tune the weights of ranking model (Full Model) with the development set.

Table 2 shows the experimental results. The Hits@1 score of image retrieval is only of 7.8%, indicating that using image data alone would not achieve high performance in retrieval. This is due to the diversity and ambiguity of images. In contrast, the Hits@1 of text retrieval is as high as 41.4%, which indicates that it is more effective to use text data to carry out retrieval.

Table 2 also shows the results of ranking. We make the following observations. 1) The results indicate that the full model of using TCM, TBM, IBM, and CLIP as matching models performs the best in terms of Hits@1. 2) The text cross-encoder matching model (TCM) makes a large performance

| Stage | Model | Hits@1 | Hits@3 | Hits@10 |
|-------|-------|--------|--------|---------|
| Retrieval | Text | 41.4 | 51.3 | 62.5 |
| | Image | 7.8 | 11.8 | 17.1 |
| Ranking | TBM | 41.5 | 52.9 | 66.8 |
| | TCM | 58.4 | 69.4 | 78.0 |
| | IBM | 9.3 | 12.5 | 17.1 |
| | CLIP | 16.3 | 27.9 | 45.3 |
| Full Model | | 61.2 | 71.4 | 79.4 |

Table 2: Experiment results (%). Full Model utilizes all matching (TBM, TCM, IBM, CLIP) scores to rank entities.

---

\(^2\)https://github.com/nmslib/hnswlib

\(^3\)https://github.com/openai/CLIP
| Model        | Hit@1 | Hits@3 | Hits@10 |
|--------------|-------|--------|---------|
| Full Model   | 61.2  | 71.4   | 79.4    |
| w/o IBM      | 59.8  | 70.6   | 79.0    |
| w/o CLIP     | 60.0  | 70.5   | 78.9    |
| w/o TBM      | 60.0  | 70.8   | 79.1    |
| w/o TCM      | 43.5  | 57.1   | 69.8    |

Table 3: Ablation study results (%).

We conduct an ablation study and examine the contributions from different matching models. As shown in Table 3, all models make contributions and the performance will drop if any of them is removed. Text information is essential, as excluding it (w/o TBM or w/o TCM) brings a significant performance decrease. Though identifying entities in images is challenging (9.3% Hits@1 for IBM), images still provide helpful information in multimodal entity tagging, because excluding image information (w/o IBM or w/o CLIP) hurts the performance. In conclusion, the task needs multimodal information and powerful multimodal models.

4.2 Ablation Study

We randomly sample 100 text-image pairs that are incorrectly tagged. We find three types of errors: “Noisy”, “Hard” and “Wrong”. “Noisy” means that noise in the dataset misleads models. “Hard” means that it is not easy for the models to resolve the ambiguity of texts (e.g., texts are general and simple) and images (e.g., entities in images are rare). “Wrong” means that the multimodal information in the input pair and MKB is clear but models fail to utilize the information to recognize entities. It turns out that 46% of the errors are Hard cases, 42% are Wrong cases, and 12% are Noisy cases. This indicates that the state-of-the-art models still cannot accomplish the task satisfactorily.

5 Related Work

There are several multimodal knowledge bases. Xie et al. (2017) create WN9-IMG, which consists of a subset entities of WordNet (Miller, 1995) and images from ImageNet (Deng et al., 2009). Mousselly-Sergieh et al. (2018) develop FB-IMG, which consists of entities from Freebase (Bollacker et al., 2008) and images from the web. Liu et al. (2019) construct MMKG, containing three knowledge bases DBpedia15K, YAGO15K and Freebase15K. Wang et al. (2020) build Richpedia, a large-scale multimodal knowledge base, which consists of “KG entities” and “image entities” on three topics. Recently, Alberts et al. (2020) develop a multimodal and multilingual knowledge base named VisualSem. VisualSem consists of 101,244 entities, 1,539,244 images and multilingual texts. We derive the new dataset from VisualSem because it is the largest MKB publically available.

5 Conclusion

We propose multimodal entity tagging (MET) with multimodal knowledge base (MKB), which is to identify the most related entity in a given text and image pair, using the information in an MKB. The new task is important for enhancing research on construction and utilization of MKB. We solve the problem by using the information retrieval paradigm. We construct a new large-scale dataset for the task and conduct intensive experiments. Experimental results indicate that the task is still challenging, and more powerful models for multimodal representation learning are needed.
References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020a. Building a multimodal entity linking dataset from tweets. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4285–4292.

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020b. Multimodal entity linking for tweets. Advances in Information Retrieval, 12035:463.

Houla Alberts, Teresa Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2020. Visualsem: a high-quality knowledge graph for vision and language. arXiv preprint arXiv:2008.09150.

Benjamin K Bergen. 2012. Louder than words: The new science of how the mind makes meaning. Basic Books (AZ).

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247–1250.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

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

KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. 2021. Towards open world object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5830–5840.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In European Semantic Web Conference, pages 459–474. Springer.

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

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

Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 225–234.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 296–310.

Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. Big Data Research, 22:100159.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Draane, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45. Online. Association for Computational Linguistics.
We conduct grid search (in \{0, 0.1, \ldots, 0.9\}) to tune the weights of ranking model (Full Model) with the development set. The weights of TCM, TBM, IBM, CLIP matching scores are \(\{v_{k1}, v_{k2}, \ldots, v_{KM}\}\) associated. After retrieval, \(N + M\) entities \(\{e_{i1}, \ldots, e_{iN}, e_{i1}, \ldots, e_{iM}\}\) are retrieved. However, images are not retrieved in text retrieval procedure and vice versa. We select an image \(v_0^{(i)}\) of the same entity for each \(t_j^{(i)}\) in \(N\) retrieved texts randomly at training time and select an image \(v_0^{(i)}\) (i.e., the first image of the entity)\(^7\) at inference time. The same procedure is used to select texts for retrieved images. The final retrieval results are \(\{e_{i1}, \ldots, e_{iN}, e_{i1}, \ldots, e_{iM}\}\) with \(\{(t_1^{(i_1)}, v_1^{(i_1)}), \ldots, (t_1^{(i_N)}, v_1^{(i_N)}), \ldots, (t_{KM}^{(i_M)}, v_{KM}^{(i_M)})\}\) associated.

### C.2 Data Sparsity

Many entities in the multimodal knowledge base do not have sufficient information. About 9.7% of the entities contain no more than three images in the dataset and 12.2% of the entities contain only one text. We conduct experiments on sparse text and sparse image data. We can observe from the results in Table 6. 1) On sparse text data, the performance of text matching models (TBM and TCM) drops significantly. 2) On sparse image data, the performance of image matching model (IBM) also

\(^5\)We find the first image (or text) of an entity contains little noise.
Table 4: Ambiguity of texts. The texts are general, simple and thus ambiguous.

| Text                                                        | # of Entities |
|-------------------------------------------------------------|---------------|
| A province of Indonesia                                      | 25            |
| One of the moons of Jupiter                                  | 20            |
| State of Mexico                                              | 19            |
| Disease                                                     | 19            |
| Genus of reptiles (fossil)                                   | 19            |
| A city of Japan                                              | 14            |
| American musician                                            | 14            |
| Year                                                        | 13            |
| ISO 3166-1 country code                                      | 13            |
| Medical specialty                                            | 13            |
| Male given name                                              | 13            |

Table 5: Assembling results (%). Underlines indicate assembling improves the performance.

| Model             | Hits@1  | Hits@3  | Hits@10 |
|-------------------|---------|---------|---------|
| TBM               | 41.5    | 52.9    | 66.8    |
| w/ assemble       | 41.0    | 49.5    | 64.2    |
| TCM               | 58.4    | 69.4    | 78.0    |
| w/ assemble       | 59.4    | 67.2    | 78.3    |
| IBM               | 9.3     | 12.5    | 17.1    |
| w/ assemble       | 4.6     | 7.1     | 12.7    |
| CLIP              | 16.3    | 27.9    | 45.3    |
| w/ assemble       | 20.6    | 31.0    | 48.2    |
| Full Model        | 61.2    | 71.4    | 79.4    |
| w/ assemble       | 63.9    | 70.9    | 79.5    |

C.3 Ambiguity

An image may contain several entities. About 21.6% of the images in the dataset are associated with two or more entities. Furthermore, entities in images may be rare. Therefore, recognizing entities from the images using visual information alone is challenging. Table 10 shows examples indicating the ambiguity of images. The animal in the first image is “Genet”. However, considering more general species categories, it can be “Viverrine” or “Chordate” (“Procyon” is a mistakenly associated entity). There are several entities in the second and third images. Depending on the texts, the recognized entities might be different. The entities in the third image are rare. Therefore, the utilization of texts is essential to recognize the entities in query. In the meantime, texts are also ambiguous. For example, the descriptions are general and simple. Table 4 gives the top ten frequent texts, and the texts are ambiguities. Although ambiguous texts are not common (1.7% of the texts in the MKB have two and more entities associated), it still brings challenges to the task.

D Experimental Results

D.1 Assembling of Ranking Model

After retrieval, each retrieved text/image is paired with the first image/text of the same entity at inference time as mentioned in Section B. However, considering the diversity of MKB, assembling more instances (an instance is a text or an image) when computing matching scores for entities is a straightforward idea to enhance the performance. We conduct additional experiments to assemble multiple instances in ranking procedure at inference time. Specifically, for each candidate \( e_{o} \) in retrieved entities \( \{ e_{i1}, \ldots, e_{iN}, e_{l1}, \ldots, e_{lM} \} \), we select \( K \) instances including the retrieved instance \( (v_{p}^{(o)} \text{ or } t_{p}^{(o)}) \) to compute matching scores with input instance (image \( v \) or text \( t \) considering different matching models) and then average them as the final score. The assemble of Full Model is linear combination of separate assemble models and the weights of TCM, TBM, IBM, CLIP matching scores are 0.1, 0.3, 0.1, 0.9 respectively.8 We set \( K = 3 \) in our experiments and larger \( K \) may bring further improvement which we leave as future work.

Table 5 illustrates the assembling results. We can observe that: 1) Hits@1 of assembled ranking model (Full Model) improves significantly, indicating more images (or texts) bring more information to distinguish and recognize the corresponding entities. 2) Hits@1 of CLIP improves a lot, indicating inter-modality matching models benefit more from the diversity of MKB. 3) The assemble performance of IBM and TBM drops and Hits@1 of IBM drops about 50% compared to original Hits@1. The main reason is that the additional instances (images or texts) may not be retrieved by retrieval models which means the relevance between instances and entities can not be learned easily, especially for bi-encoder models which are less powerful and more difficult to train.

In conclusion, utilization of more instances (im-

---

8We conduct grid search with development set.
ages or texts) may bring performance improvement to models especially inter-modality matching models due to the diversity of MKB. Meanwhile, more instances may confuse models and hurt the performance, especially for less powerful bi-encoder matching models.

**D.2 More Experimental Results**

Table 7 reports the results on both development and test sets. One can see that the results on development and test sets are generally consistent. Hits@100 achieves 86%, about 20% higher than Hits@1, indicating that there is still room for performance improvement and more powerful models for multimodal representation are needed.
### Table 6: Results on sparse data (%). All models perform worse on sparse data, especially on sparse text data. This shows sparsity is still a challenge to the new task MET.

| Model | Sparse Text Data | | Sparse Image Data | |
|-------|------------------|------------------|------------------|
|       | Hits@1 | Hits@3 | Hits@10 | Hits@1 | Hits@3 | Hits@10 |
| TBM   | 30.5  | 11.0   | 40.3  | 12.6 | 54.5  | 12.3   | 41.1  | 4.9   | 51.2  | 1.7   | 61.7  | 5.1   |
| TCM   | 46.2  | 12.2   | 57.3  | 12.1 | 67.5  | 10.5   | 55.2  | 2.2   | 66.1  | 3.3   | 74.2  | 2.8   |
| IBM   | 8.3   | 3.0    | 12.3  | 0.2  | 17.2  | 0.1    | 4.4   | 4.9   | 5.2   | 7.3   | 6.5   | 10.6  |
| CLIP  | 15.3  | 1.0    | 25.0  | 2.9  | 39.7  | 5.6    | 11.7  | 4.6   | 19.8  | 8.1   | 32.7  | 12.6  |
| Full Model | 50.3 | 10.9   | 60.6  | 10.8 | 69.4  | 10.0   | 58.1  | 3.1   | 66.5  | 4.9   | 75.0  | 4.4   |

Table 7: All experiment results (%). The results on development and test sets are generally consistent which indicates that development set is suitable for model selection.

| Stage | Model | Dev | Test |
|-------|-------|-----|------|
|       | Hits@1 | Hits@3 | Hits@10 | Hits@100 | Hits@1 | Hits@3 | Hits@10 | Hits@100 |
| Retrieval | Text | 38.8 | 50.4 | 63.1 | 79.6 | 41.4 | 51.3 | 62.5 | 78.9 |
|         | Image | 7.9  | 12.7 | 19.1 | 32.6 | 7.8  | 11.8 | 17.1 | 31.3 |
| Ranking | TBM   | 41.0 | 52.3 | 65.0 | 82.1 | 41.5 | 52.9 | 66.8 | 81.7 |
|         | TCM   | 59.0 | 70.2 | 78.5 | 85.9 | 58.4 | 69.4 | 78.0 | 85.4 |
|         | IBM   | 10.5 | 14.0 | 18.9 | 34.0 | 9.3  | 12.5 | 17.1 | 34.1 |
|         | CLIP  | 18.5 | 29.9 | 46.6 | 82.9 | 16.3 | 27.9 | 45.3 | 82.3 |
|         | Full Model | 61.6 | 71.7 | 79.6 | 86.1 | 61.2 | 71.4 | 79.4 | 85.5 |

Table 8: Error analysis. 1) "**Noisy**". "Electric Outlet" has power plugs images associated which misleads models. 2) "**Hard**". The input image of "Province of Sassari" is ambiguous and hard to recognize the corresponding city. 3) "**Wrong**". The input text of "Grumman" is similar to "Ingalls Shipbuilding" and text matching models fail to recognize the correct entity "Grumman" which indicates more powerful models for multimodal representation are needed. One can click entity ids to see details of the entities.

| Type  | Input | Entity | Prediction |
|-------|-------|--------|------------|
| Noisy | Type of power plugs standardized by the National Electrical Manufacturers Association | bn:03357573n | bn:00030158n |
|       | Nema Connector | Electric Outlet |
| Hard  | A town, and associated province in Sardinia, Italy. | bn:00665773n | bn:00665687n |
|       | Province of Sassari | Province of Cagliari |
| Wrong | A former aircraft maker, now PART of Northrop Grumman | bn:01455566n | bn:01359190n |
|       | Grumman | Ingalls Shipbuilding |
| ID          | Name               | Description                                                                 |
|-------------|--------------------|-----------------------------------------------------------------------------|
| bn:00068195n| Romanesque Architecture | • A style of architecture developed in Italy and western Europe between the Roman and the Gothic styles after 1000 AD;  
• Characterized by round arches and vaults and by the substitution of piers for columns and profuse ornament and arcades.  
• Romanesque architecture is an architectural style of medieval Europe characterized by semi-circular arches.  
• Architecture of Europe which emerged in the late 10th century and lasted to the 13th century. |
| bn:00015540n| Paris               | • The capital and largest city of France; and international center of culture and commerce.  
• Paris is the capital and most populous city of France, with an estimated population of 2,148,271 residents as of 2020, in an area of more than 105 square kilometres.  
• Designated nouvelle Rome at various stages of history between the reigns of Philip IV and Louis XIV.  
• The capital and largest city of France.  
• Capital of France. |
| bn:00036225n| Francisco de Goya | • Francisco José de Goya y Lucientes was a Spanish romantic painter and printmaker.  
• Spanish painter well known for his portraits and for his satires (1746-1828).  
• 18th and 19th-century Spanish painter and printmaker.  
• Francisco Goya, a Spanish painter. |

Table 9: Examples showing the diversity of data. The examples show typical entity categories: 1) Conceptual entities containing various instances and images. 2) Location entities containing diverse images related to the location (e.g., landmarks, maps, symbols). 3) Person entities containing diverse images related to the person (e.g., portraits, works).
| Image | Text | Entity |
|-------|------|--------|
| ![Image](79x499 to 120x553) | The type genus of the family Procyonidae: raccoons. Small cat-like predatory mammals of warmer parts of the Old World. Any animal of the phylum Chordata having a notochord or spinal column. Agile Old World viverrine having a spotted coat and long ringed tail. | bn:00039337n (Procyon) bn:00080161n (Viverrine) bn:00018748n (Chordate) |
| ![Image](77x458 to 122x492) | Large black-and-white herbivorous mammal of bamboo forests of China. Woody tropical grass having hollow woody stems; mature canes used for construction and furniture. | bn:00002174n (Giant Panda) bn:00008254n (Bamboo) |
| ![Image](77x395 to 122x426) | A situation or topic as if viewed from an altitude or distance. A wrought iron tower 300 meters high that was constructed in Paris in 1889; for many years it was the tallest man-made structure. The capital and largest city of France; and international center of culture and commerce. An art movement launched in 1905 whose work was characterized by bright and nonnatural colors and simple forms; influenced the expressionists. | bn:00010616n (Bird’s Eye View) bn:00029980n (Eiffel Tower) bn:00015540n (Paris) bn:00033829n (Fauvism) |

Table 10: Examples showing the ambiguity in the data. One image may contain 1) various entities. The second image contains entities “Giant Panda”, “Bamboo”. The third image contains entities “Eiffel Tower”, “Paris”, “Fauvism”. If not specified, it is hard to recognize the corresponding entity. 2) an object corresponding to various entities. The first image contains an animal object. If considering different species categories level, the entity can be “Viverrine”, “Chordate”, “Genet”. 3) rare entities. The third image contains entity “Bird’s Eye View” which is obscure and rare.