Visualizable or Non-visualizable? Exploring the Visualizability of Concepts in Multi-modal Knowledge Graph

Xueyao Jiang[1], Ailisi Li[1], Jiaqing Liang[1], Bang Liu[2], Rui Xie[3], Wei Wu[3], Zhixu Li[1(✉)], and Yanghua Xiao[1,4(✉)]

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China
[2] Mila & DIRO, Université de Montréal, Montréal, Québec, Canada
[3] Meituan, Shanghai, China
[4] Fudan-Aishu Cognitive Intelligence Joint Research Center, Shanghai, China
{xueyaojiang19,alsli19,zhixuli,shawyh}@fudan.edu.cn,l.j.q.light@gmail.com, bang.liu@umontreal.ca, rui.xie@meituan.com,wuwei19850318@gmail.com

Abstract. An important task in image-based Multi-modal Knowledge Graph construction is grounding concepts to their corresponding images. However, existing research omits the intrinsic properties of different concepts. Specifically, there are some concepts that can not be characterized visually, such as *mind*, *texture*, *session cookie* and so on. In this work, we define concepts like these as non-visualizable concepts (NVC) and the others like *dog* that have clear and specific visual representations as visualizable concepts (VC). And, we propose a new task of distinguishing VCs from NVCs, which has rarely been tackled by the existing efforts. To address this problem, we propose a multi-modal classification model combining concept-related features from both texts and images. Due to the lack of enough training samples especially for NVC, we select concepts in ImageNet as the instances for VC, and also propose a webly-supervised method to get a small set of instances for NVC. Based on the small training set, we modify the basic two-step positive-unlabeled learning strategy to train the model. Extensive evaluations demonstrate that our model significantly outperforms a variety of baseline approaches.

Keywords: Visualizable Concept · Multi-modal Knowledge Graph.

1 Introduction

Nowadays, multi-modal data (mainly images) is introduced into Knowledge Graph to enrich the representation of concepts, and increasing efforts are focused on grounding entities or concepts with their corresponding images to construct image-based Multi-modal Knowledge Graph (MMKG) [6–10]. However, not all concepts in Knowledge Graphs can be characterized accurately using images, such as *mind*. Although we may associate it with images of *brain*, as shown in Fig 1, what these images depicted are not exactly *mind* itself but something relevant with it.

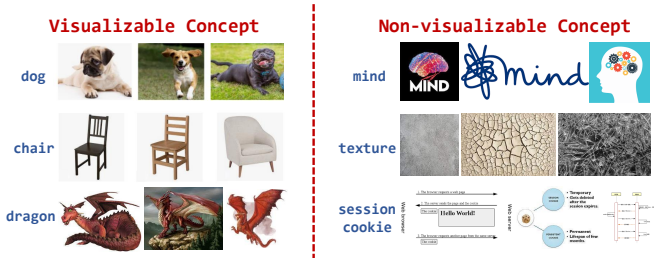

Fig. 1: Examples of Visualizable Concepts and Non-visualizable Concepts.

Thus, we propose to classify concepts by whether they can be characterized visually. We name concepts that do not have a clear visual representation and can not be characterized visually as **Non-visualizable Concepts (NVCs)**. Oppositely, concepts that have clear and specific visual representations are called as **Visualizable Concepts (VCs)**, such as *dog*.

Most previous MMKG construction efforts focused on grounding entities or concepts with their corresponding images without carefully differentiating between VCs and NVCs. For example, TinyImage [13] (which is a MMKG built on WordNet [4]) simply eliminated all abstract concepts in the hierarchical taxonomy of WordNet while it is rough and inaccurate according to our experiments in Section 3. [14] analyzed the problem of non-visualizable concepts in the *person* subtree of ImageNet, but they rely on crowdsourcing to annotate the imageability of synsets which requires a lot of manpower and material resources and is difficult to apply to large-scale MMKG. Several previous research has also addressed this problem implicitly in a learning-based way. [3] proposed to filter out non-visualizable concepts based on the visually salient score of concepts. However, they merely use webly searched images to evaluate the visually salient score, which easily suffers from the noise or bias of web data.

In this paper, we explore the visualizability of concepts by classifying VCs versus NVCs. Particularly, we design a visualizable concept classifier with multimodal information as features, namely text and images. The classifier takes the concept description and online images of the concept (collect from the image search engine) as input and then output whether the concept is a VC or NVC. Besides, our classifier still faces the following challenges:

1) **Lack of labeled data.** We propose to automatically construct a partially annotated dataset by contrasting a symbolic knowledge base (i.e. WordNet) and an annotated image dataset organized according to this knowledge base (i.e. ImageNet [1]).
2) **Learning under PU setting.** The partially annotated dataset contains a small set of positive data and a large set of unlabeled data. Due to the severe imbalance between the number of positive data and the number of unlabeled data, we adopt the two-step PU Learning technique to tackle this problem.

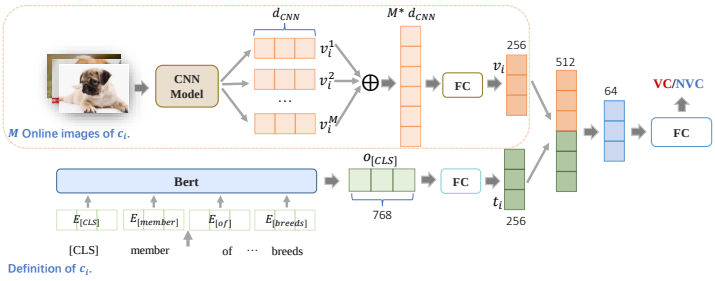

Fig. 2: Overview architecture of Visualizable Concept Classifier.

2 Methodology

In this paper, we model this problem as a binary classification problem and propose a multi-modal visualizable concept classification model to solve it, and the Two-step PU Learning strategy is leveraged to tackle the PU setting challenge in the dataset.

2.1 Multi-modal Visualizable Concept Classifier

As shown in Fig 2, we leverage two separate streams to get the text representation and visual representation, respectively. The representations of two modalities are further concatenated together to get the fusion of two features, which are fed into a binary classification network to classify VC and NVC.

- **Text embedding** We use BERT to get the text embedding. The concept definition d_i is first processed into the input format of BERT as: "[CLS] d_i [SEP]" and then fed into BERT. The embedding of mark "[CLS]" is then further encoded into a 256-dim vector $t_i \in R^{256}$.
- **Image embedding** Several pretrained image classification models are leveraged to get the image representations. We input M images into the pretrained image classification model respectively, and get M feature vectors $v_i^j \in R^{d_{CNN}} (1 \leq j \leq M)$ which is the output of the final pooling layer. d_{CNN} denotes the dimension of CNN model's average pooling layer and differs by the CNN models. These M vectors are concatenated together and fed into a fully-connected layer to get the final visual feature vector $v_i \in R^{256}$.
- **Classifier** The embeddings of text and image are then concatenated into a 512-dim vector and then fed into a binary classifier (contains a 64-dim hidden layer and an output layer) to generate the probability that the concept is a VC or not.

2.2 Training Under PU Setting

Due to the small amount of positive data and lack of negative data, we leverage Two-step PU Learning strategy to train the Visualizable Concept Classifier. As

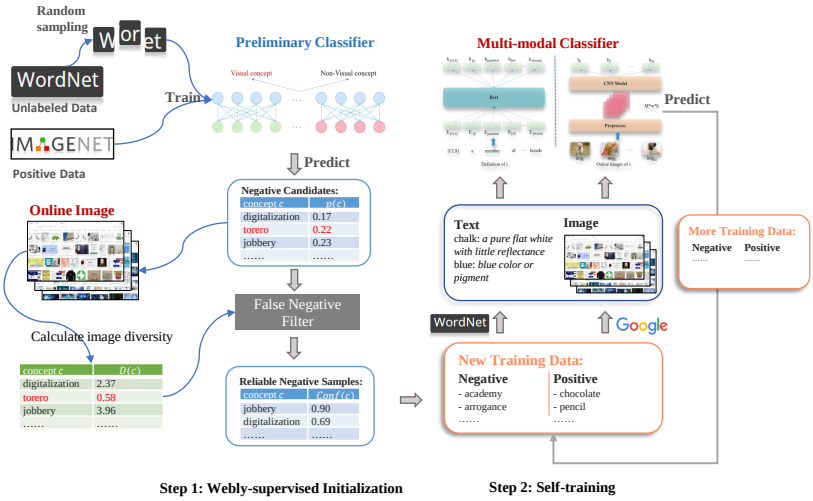

Fig. 3: Overview of the Framework.

shown in Fig 3, to complement the training dataset for binary classification, we design a Webly-supervised Initialization step that screens out a reliable negative set with the same size as that of the positive set from unlabeled data and then train the Multi-modal classifier iteratively with the self-training strategy.

3.2.1 Webly-supervised Initialization

We propose a webly-supervised automatic method to construct a high quality negative set without manual labeling. It includes two steps: 1) Preliminary Classifier; 2) False Negative Filter.

1) Preliminary Classifier Some unlabeled data is randomly sampled out as pseudo negatives. Together with the labeled positive data, they are used to train a preliminary classifier of VC and NVC. And then the preliminary classifier is used to predict the unlabeled data and output the confidence score of a concept c to be NVC (denoted by $p(c)$).

2) False Negative Filter This filter is designed based on the following assumption: *Concepts with diverse online images are more likely to be non-visualizable.* We measure the image diversity of concept c by calculating the standard deviation of corresponding images' feature vectors (denoted by $D(c)$), then we recalculate the confidence score of a concept c to be NVC as $Conf(c) = softmax(1 - p(c)) + softmax(D(c))$.

3.2.2 Self-training

After getting the reliable negative data, we iteratively train our multi-modal classifier. In each iteration, we train the classifier and then use it to predict

unlabeled data. Predicted data with high confidence score will be added to the training set to serve next iteration's training.

However, noise is unavoidable by adopting automated approach to sample negative data in the first step of PU Learning. Self training under such setting will face the challenge of label drift. We design two small tricks to avoid it:

- **randomly sample** Taking sampling negative data as example, instead of using top k candidates as the pseudo labeled data, we randomly sample k candidates according to their confidence scores. For example, the concept c's confident score is 0.8, then its probability of being selected is 0.8.
- **extra visual information** We reuse the priori knowledge that non-visualizable concepts' online images have higher diversity which can be measured by their standard deviation. The diversity score is added to predicted confidence score.

3 Experiment

3.1 Datasets and Settings

Dataset We conduct our experiments on WordNet noun set. And we use ILSVRC dataset [5] to label positive concepts and remain the reset in WordNet unlabeled. For each concept, we collect two modalities of data. One is definition text in WordNet, the other is online images which are retrieved from Google search engine. The search query is in the form of "c d". c is the name of a concept and d is the definition from WordNet to disambiguate concepts with the same name. We manually labeled 600 concepts randomly extracted from WordNet as test set including 322 positive samples and 278 negative samples.
Experimental settings We apply BERT [2] to gain the text features. We conduct experiments to extract image features with three different CNN models including InceptionV3 [12], Resnet50 [5] and VGG16 [11]. During training, batch size is set to 64 and the learning rate adjustment schedule is set as [2] in each iteration while the initial learning rate of each iteration is set to 1e-3, 1e-4, 1e-6, 1e-8.

3.2 Main Results

We are the first to explicitly propose to distinguish visualizable concepts and non-visualizable concepts automatically, so there are few work that we can compare to. We compare our method with 3 baselines:
Full set As most previous work of the MMKG construction ignore to distinguish VC and NVC, we design a baseline approach that regards all concepts as VC.
TinyImage [13] TinyImage is a MMKG that is constructed based on WordNet hierachy. It simply regards all abstract concepts as NVCs and the others as VCs.

[5] http://image-net.org/challenges/LSVRC/2012/browse-synsets

Table 1: Comparison result

Model	acc	rec_{VC}	$prec_{VC}$	rec_{NVC}	$prec_{NVC}$
Full set	0.537	**1.0**	0.537	0	-
TinyImage	0.605	0.491	0.684	0.737	0.556
LEVAN	0.542	0.991	0.538	0.007	0.250
Ours	**0.828**	0.84	**0.830**	**0.810**	**0.820**

LEVAN [3] LEVAN proposed to filter out NVC based on the visual salience score of concepts. Specifically, they train a binary image classifier for each concept on a dataset in which online images of the concept are labeled positive and background images are labeled negative. They regard a concept as visual salient if the well trained classifier can reach a threshold of accuracy on the validation set that has the same setting with the training set.

We compare our method with 3 baselines mentioned above. As shown in Tab 1, our framework outperforms the other 3 methods in acc, $prec_{VC}$, rec_{NVC} and $prec_{NVC}$. The reason of *Full set* method has the highest recall of VC is that it regards all concepts as VC. Besides, it predicts no NVC, so the precision of NVC is not calculable. We finally use the optimal model to predict the whole unlabeled set and get 35,481 NVCs and 37,702 VCs.

3.3 Ablation Study

In this section, we provide ablation studies on Webly-supervised Initialization, multi-modal classification model and PU Learning.

Webly-supervised Initialization We conduct ablation experiments for both two submodules in the Webly-supervised Initialization step (step 1) by removing one of these two submodules. Besides, as for Preliminary Classifier, we test two classification models: **text based model** (TM) and **multi-modal model** (MM). The MM is same as depicted in Section 2.1. TM refers to the model that removes the image embedding structure (depicted in the dotted box in Fig 2) from MM. In conclusion, we design five experiments: (1) only FF; (2) MM w/o FF; (3) MM with FF; (4) TM w/o FF; (5) TM with FF.

To measure the quality of the datasets that outputed by above 5 combinations, we train the classifier on them and evaluate the accuracy on validation set. The experiment results are given in Tab 2 (step 1). The combination of text based model (TM) and Flase Negative Filter results in a best negative set on which the multi-modal classifier is trained to achieve a highest accuracy of **0.78**. As shown, False Negative Filter brings around 7% absolute improvement.

Multi-modal Classification For Self-training step (step 2), we also conduct experiments on TM and MM. As is shown in Tab 2 (step 2), the training of classifier benefits a lot from multi-modal information. The removal of multi-modal information in self-training leads to a drop of around 12% in accuracy

Table 2: Ablation experiments of Webly-supervised initialization

Test step	Model	False Negative Filter	val accuracy
step 1	-	√	0.65
	MM	×	0.67
	MM	√	0.65
	TM	×	0.71
	TM	√	**0.78**
step 2	+ TM	√	0.77
	+ MM(InceptionV3)	√	0.80
	+ MM(ResNet50)	√	0.80
	+ MM(Vgg16)	√	**0.83**

after 4 iterations. As for influence of different CNN models, the accuracy of these three models are respectively 0.797(Inception V3), 0.804(ResNet50) and **0.828(VGG16)**, in which VGG16-based model achieves the best result.

Two-step PU Learning We compare the performance of our framework with classifiers directly trained on dataset constituting of randomly sampled negative data and labeled positive data. We trained the classifier 5 times and the average accuracy of these classifiers is 0.683, while our framework can reach an accuracy of **0.828**, which is approximately 7% higher.

4 Related Work

The visualization of concepts is not a new topic in computer vision. LEVAN [3] proposed to recognize "visual salient" words during constructing image dataset. In this work, the authors believed that images of a visual salient ngrams can be easily distinguished from background images and have small inter-class variances. A classifier based on SVM is trained to distinguish online images of a ngram and background images. The classifier's accuracy of VC should exceed a threshold. TinyImage [13] is a MMKG constructed based on WordNet and contains 75k noun concepts with 1,052 images per concept in average. The authors regarded all the abstract concepts as not proper to be matched with images and removed them by dropping all the hyponyms of the word "abstraction", while according to our experiments, such strategy is rough and inaccurate.

5 Conclusion

In this work, we propose a new task: *distinguishing visualizable concepts from non-visualizable concepts* and model this problem as a binary classification problem. We automatically generate a partially labeled dataset and propose a novel two-step PU learning framework to train the classifier on such dataset. Besides, multi-modal information of concepts is used to enhance the performance of the visualizable concept classifier. Extensive experimental results show that our solution achieves the state-of-the-art results compared to several baselines.

Acknowledgement

This work is supported by National Key Research and Development Project (No.2020AAA0109302), Shanghai Science and Technology Innovation Action Plan (No.19511120400), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103) and National Natural Science Foundation of China (Grant No. 62072323).

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Divvala, S.K., Farhadi, A., Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3270–3277 (2014)
4. Fellbaum, C.: Wordnet. The encyclopedia of applied linguistics (2012)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision 123(1), 32–73 (2017)
7. Li, M., Zareian, A., Lin, Y., Pan, X., Whitehead, S., Chen, B., Wu, B., Ji, H., Chang, S.F., Voss, C., et al.: Gaia: A fine-grained multimedia knowledge extraction system. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 77–86 (2020)
8. Mitchell, T., Fredkin, E.: Never ending language learning. In: Big Data (Big Data), 2014 IEEE International Conference on. pp. 1–1 (2014)
9. Perona, P.: Vision of a visipedia. Proceedings of the IEEE 98, 1526 – 1534 (09 2010)
10. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. International journal of computer vision 77(1-3), 157–173 (2008)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
12. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
13. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE transactions on pattern analysis and machine intelligence 30(11), 1958–1970 (2008)
14. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 547–558 (2020)