KB-VLP: Knowledge Based Vision and Language Pretraining

Kezhen Chen [1] **Qiuyuan Huang** [2] **Yonatan Bisk** [3] **Daniel McDuff** [2] **Jianfeng Gao** [2]

Abstract

Transformer-based pretraining techniques have achieved impressive performance on learning cross-model representations for various multi-modal tasks. However, off-the-shelf models do not take advantage of commonsense knowledge and logical reasoning that are crucial to many real-world tasks. To this end, we introduce a novel pretraining approach - **K**nowledge **B**ased **V**ision and **L**anguage **P**retraining (KB-VLP) - which uses knowledge graph embeddings extracted from text and detected image object tags to enhance the learning of semantically aligned and knowledge-aware representations, and improve the models generalization, and interpretability. KB-VLP is pretrained on a large image-text corpus and automatically extracted knowledge embeddings, and then finetuned on several downstream vision-language tasks. Experiments show that KB-VLP significantly improves the performance on VQA, GQA, NLVR^2 and OKVQA tasks compared with the baselines.

1. Introduction

Large-scale pretraining models have dramatically improved the quality of natural language processing (NLP) and vision-language models. Although these methods use image and text information as inputs and learn image-text alignments via well designed pretraining tasks, most still lack the external commonsense knowledge necessary for many tasks. The external knowledge is usually hard or impossible to be learned from standard dataset. More specifically, existing models: a) can often disregard the shared and complementary information provided by different modalities; b) largely ignore the structure of the knowledge graph and commonsense reasoning. To overcome these challenges, we argue that models should not only consider multiple modalities (i.e., vision and language) but also the rich structural and logical information embedded in commonsense knowledge. In this paper, we develop a general-purpose vision-language pretraining method, **K**nowledge **B**ased **V**ision and **L**anguage **P**retraining (KB-VLP). We leverage a knowledge graph (Wikipedia) which has more than nine million entities and corresponding relations. The knowledge graph provides rich information that could be useful for many downstream tasks related to vision-language understanding.

To summarize our contribution: i) We develop a knowledge-reasoning self-supervised pretraining approach using a knowledge graph structure to learn multi-modal representations, which includes physical properties, and ontological qualia/relations that might be hard/impossible to recover from text alone; ii) We adapt knowledge-reasoning-patches rather than use text and image bounding box features. Our approach enables the model to identify the types of knowledge, the space of entities, etc. that we are interested in and which may not be captured by the standard objects detected via bounding box approaches. We promote these newer representations to handle a broader space of visual semantics than previous methods. iii) We leverage the Wikipedia knowledge graph as the commonsense knowledge base, which includes entities, corresponding relations and information for general-purpose applications on multi-modal tasks, we present experiments and analysis to demonstrate the effectiveness of our approach.

2. Related Work

Multi-modal representation learning is essential for vision-language tasks, such as image-captioning, visual question answering and visual commonsense reasoning. Large-scale Transformer architectures (Vaswani et al., 2017) have achieved impressive performance by pretraining representations for NLP problems (Peters et al., 2018; Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019; Radford et al., 2019; Brown et al., 2020). Recent work on vision-language pretraining (VLP) has shown that these large-scale pretraining methods can also lead to effective cross-modal representations (Lu et al., 2019a; Tan & Bansal, 2019; Zhou et al., 2019; Chen et al., 2019; Alberti et al., 2019; Li et al., 2020a; 2019; 2020b; Zhang et al., 2021; Kim et al., 2021). Although many researchers are paying attention to the importance of

[1]Northwestern University; [2]Microsoft Research, Redmond; [3]Carnegie Mellon University. Correspondence to: Kezhen Chen <kezhenchen2021@u.northwestern.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

multi-modal pretraining, until recently many had not used external knowledge graph information in these methods. While there are works using knowledge during pretraining of language models (Yu et al., 2020; Xu et al., 2021; Rosset et al., 2021; Zhou et al., 2020; He et al., 2020a; Xiong et al., 2019; He et al., 2020b; Agarwal et al., 2021), most have not been applied to multi-modal transformers (e.g. for vision *and* language). Additionally, several of the proposed architectures are domain-specific and are hard to extend to new tasks. In this paper, we introduce a knowledge-based pretraining model using the transformer architecture for multi-modal understanding and reasoning. The knowledge representations in our method can be easily extracted on massive data. Our proposed KB-VLP architecture shows how the structural knowledge and reasoning information extracted from text and images facilitates learning more robust and knowledge-aware representations for vision-language tasks.

3. KB-VLP Approach for Vision-Language Pretraining

When humans reason about the world, they usually do so via multiple modalities and combine sensory information with external knowledge. Inspired by this idea, we introduce a new pretraining approach, **K**nowledge **B**ased **V**ision and **L**anguage **P**retraining (KB-VLP), which uses a multi-layer transformer model to learn unified representations on external knowledge and vision-language inputs. Given an image-text pair, we extract the knowledge information from the text and image and convert them to knowledge graph embeddings. These embeddings are used as additional inputs for pretraining. Figure 1 shows an illustration of KB-VLP. In this section, we first present how we extract the external knowledge from the knowledge base and then we introduce the details of our pretraining approach.

3.1. Extracting Knowledge

For our experiments we chose the Wikidata knowledge base (Vrandecic & Krotzsch, 2014) as a source of external knowledge. This contains a large number of relevant and important real-world entities. Given a piece of text T with n words $\{w_0, ..., w_n\}$, we first perform named entity recognition on T based on the wikidata knowledge graph and generate an entity set E, which has m named entities $\{e_0, ..., e_m\}$. Each entity has a span in T with length of one or more words. After the named entity recognition, words in T can be separated into two subsets P and Q. The first subset P has p words $\{w_0, ..., w_p\}$, that construct the recognized entities. The second subset Q has q words $\{w_0, ..., w_q\}$, which are the remaining words excluding the recognized entities. Next, these named entities are converted to knowledge graph embeddings. We use the open-license tool, Wikipedia2Vec (Yamada et al., 2020), for obtaining embeddings of words and entities from Wikipedia. This tool implements the conventional skip-gram model to learn the embeddings of words and its extension to learn the embeddings of entities (Yamada et al., 2016; Yamada & Shindo, 2019). These models enables us to learn embeddings of words and entities simultaneously, and places similar words and entities close to one another in a continuous vector space. In this paper, we use the embedding model based on the English language Wikipedia-20180420 set. To extract as much external knowledge as we can, we convert both P and Q as embedding vectors. For subset P, we directly use its corresponding entities in E, and each entity has an embedding vector. For subset Q, we use the original English words and each word has an embedding vector. The embeddings of entities in E and words in Q are combined together as the knowledge embeddings for the text T. Figure 2 in Appendix A shows the pipeline of knowledge extraction.

3.2. Input

KB-VLP represents each image-text pair as five parts (w, k^w, t, v, k^t), where w is the sequence of word embeddings for the text, t is the word embedding sequence for the image object tags, v is the set of region feature vectors for the image, k^w is the sequence of knowledge embedding vectors extracted from the text, and k^t is the sequence of knowledge embedding vectors extracted from the object tags.

For each image-text pair, most of the existing VLP models represent the input pair as a sequence of word embeddings for the text, and a set of region vectors for the image. Inspired by Li et al. (2020b) and Zhang et al. (2021), we adopt an additional input, a sequence of object tags, which are used as anchor points to ease the learning of image-text alignment. These object tags are the category names or semantically similar words of detected objects in the image. For generating v, we used a X152-C4 architecture as the object detection model (OD), which is initialized from an ImageNet-5K checkpoint (Deng et al., 2009). The OD model is pretrained on four vision datasets including Visual Genome (Krishna et al., 2016), COCO (Lin et al., 2014), Objects365 (Shao et al., 2019) and OpenImagesV5 (Kuznetsova et al., 2020). Given an image, the pretrained OD model generated the set of detected object names and the set of region features. Each region feature contains an vector of the image feature with 2048 dimension and a positional encoding of the region with 6 dimension. The image feature vector is concatenated with the positional encoding to construct the vectors in v, where each region vector in v has 2054 dimension. In pretraining, t uses the object tags in image captioning datasets and answer text in visual question answering datasets.

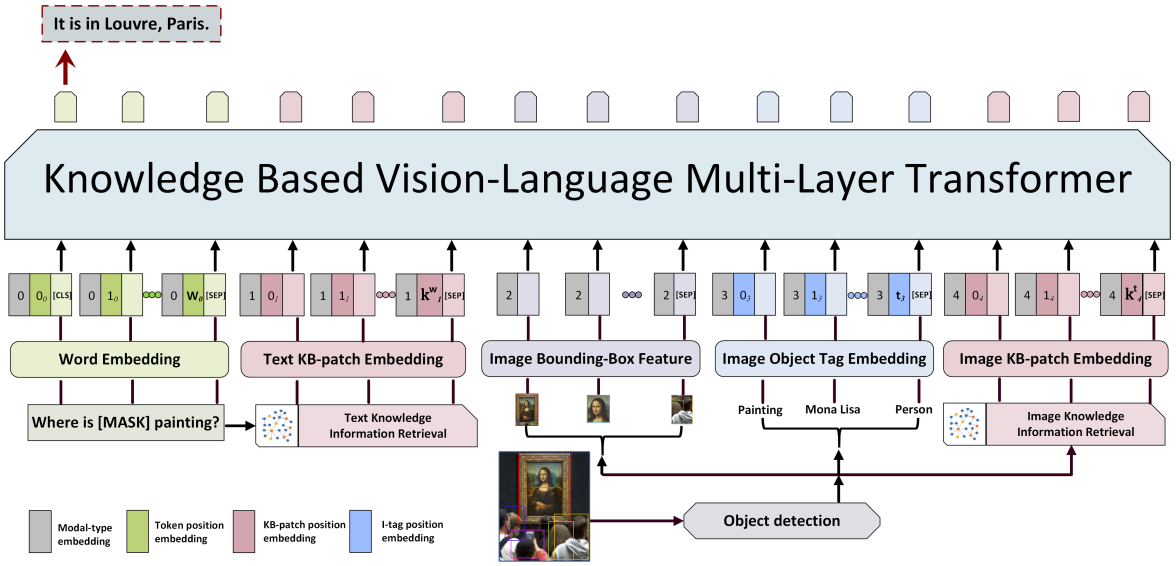

Figure 1. The KB-VLP model: Given an image-text pair, the input is represented as a tuple (w, k^w, v, t, k^t), where w is the text embedding sequence, k^w is the text-KB-patch embedding sequence for the related text entities in the knowledge base, v is the image region bonding-box feature embedding sequence, t is the object tag embedding sequence, and k^t is the image-KB-patch embedding sequence for the related image entities from the knowledge base.

For each text-image pair, we also extract the knowledge graph embeddings k^w and k^t from both the text and the image. The text in each pair is used for knowledge extraction and constructing k^w. We use the objects tags in each image for knowledge extraction and generating k^t. To speed up pretraining, we extract the knowledge embeddings for all the training pairs and load them in memory for training.

3.3. Pretraining Objective

KB-VLP is pretrained with two types of objectives: sequence-level and token-level. Sequence-level objective distinguishes the representations of the text, image and the external knowledge. Token-level objective distinguishes the semantic space of inputs. Thus, we propose the novel KB-VLP pretrainig loss $\mathcal{L}_{pretraining}$ as $\mathcal{L}_{pretraining} = \mathcal{L}_{SL} + \mathcal{L}_{TL}$, where $\mathcal{L}_{SL}$ is the loss from sequence-level pretraining and $\mathcal{L}_{TL}$ is the loss from token-level pretraining. Next, we introduce the details for each loss.

Sequence-Level Objective. The sequence-level loss $\mathcal{L}_{SL}$ is a four-way contrastive loss. Given the input tuples (w, k^w, v, t, k^t) from dataset D, we construct negative inputs by polluting the tuples to compute the loss. At each time, we keep the correct tuple or replace one of tuple elements including the text, tags or knowledge with a random element from another document, which results in three different types of polluted tuples: $(w_{neg}, k^w, v, t, k^t)$, $(w, k^w, v, t_{neg}, k^t)$ and $(w, k^w_{neg}, v, t, k^t_{neg})$. The correct tuple remains unchanged on 50% of occasions. In the remaining 50%, the three types of negative samples have equal probability of being generated. During pretraining, KB-VLP model aims to predict whether the tuple is correct or polluted. Following the tradition of Transformer-based pretraining, the encoding of the [CLS] token is used as the representation of the tuple input. We passed this encoding of [CLS] to a fully-connected layer $f(.)$ and predict four classes: the tuple is correct (c=0), w is unmatched (c=1), t is unmatched (c=2) or k^w, k^t are unmatched (c=3). Then the sequence-level loss is defined as $\mathcal{L}_{SL} = -\mathbb{E}_{(w,t,v,k^w,k^t;c)\sim D} \log p(c|f(w, t, v, k^w, k^t))$.

Token-Level Objective. The token-level loss $\mathcal{L}_{TL}$ has two parts. Firstly, we use the masked token loss $\mathcal{L}_{MTL}$ (Devlin et al., 2018) on the text elements (w and t). Secondly, on the knowledge embeddings k^w and k^t, we design a novel polluted knowledge loss $\mathcal{L}_{PKL}$. On 85% of occasions, the embeddings do not change. In the remaining 15%, the embedding vector can be polluted. It has 50% probability of being replaced with a random embedding vector of a word in the Wikipedia2Vec dictionary and 50% probability of staying as the original embedding. As the number of words and entities in Wikipedia are very large, predicting the original words or entities of polluted embeddings is inefficient for pretraining. Thus, we design two different objectives for the knowledge embeddings from the text and image. For the text knowledge embedding, the model only predicts whether the embedding is the original one ($c^w = 0$) or has been randomly replaced ($c^w = 1$). For the image knowledge embedding, the model predicts the original words or entities from a subset, which contains only the words or entities for object tags. Based on this design, the token-level loss is

Model	VQA		NLVR^2		GQA		OK-VQA			
	Dev	**Test-std**	**Dev**	**Test-P**	**Dev**	**Test-std**	**R@1**	**R@5**	**R@10**	**ACC-full**
NSM (Drew A. Hudson, 2019)	–	–	–	–	–	**63.17**	–	–	–	–
ViLBERT (Lu et al., 2019a)	70.63	70.92	–	–	–	–	–	–	–	–
VL-BERT (Su et al., 2020)	70.50	70.83	–	–	–	–	–	–	–	–
VisualBERT (Li et al., 2019)	70.80	71.00	67.40	67.00	–	–	–	–	–	–
LXMERT (Tan & Bansal, 2019)	72.42	72.54	74.90	74.50	60.00	60.33	–	–	–	–
12-in-1 (Lu et al., 2019b)	73.15	–	–	78.87	–	60.65	–	–	–	–
UNITER-B (Chen et al., 2019)	72.27	72.46	77.14	77.87	–	–	–	–	–	–
Oscar-B (Li et al., 2020b)	73.16	73.44	78.07	78.36	61.19	61.58	34.50	63.95	73.47	30.07
KB-VLP (ours)	**73.63**	**73.89**	**78.23**	**78.44**	**62.40**	62.57	**41.10**	**72.05**	**82.28**	**33.41**

Table 1. Results of KB-VLP on across VQA, GQA, NLVR^2 and OK-VQA show that our model can consistently outperform existing VLP baselines on most of the tasks.

defined as $\mathcal{L}_{TL} = \mathcal{L}_{MTL} + \mathcal{L}_{PKL}$.

3.4. Pretraining Corpus

We use the public corpus of Zhang et al. (2021) for pretraining. This corpus contains image-text pairs from several existing vision-language datasets, including COCO (Lin et al., 2014), Conceptual Captions (Sharma et al., 2018), SBU captions (Ordonez et al., 2011), Flickr30k (Young et al., 2014), GQA, VQA, VG-QAs, and a subset of OpenImages. As KB-VLP uses external knowledge during pretraining, we also add the training set from the OK-VQA (Marino et al., 2019) dataset to help the model learn how to align the external knowledge with the image-text pairs. The final corpus has about 5.65 million images, 2.5 million QA pairs, 4.68 million captions, and 1.67 million pseudo-captions. More implementation details are presented in appendix B.

4. Adapting to Vision-Language Tasks

After pretraining, we apply KB-VLP to several downstream vision-language understanding tasks including VQA, GQA, NLVR and OK-VQA. Each task poses different knowledge and reasoning challenges. The details of adapting KB-VLP on them are described in Appendix C and D.

5. Experiments

We conduct experiments on VQA, GQA, NLVR^2 and OK-VQA and compare our model against several baseline models. Table 1 presents the overall performance on the four tasks. Results show that KB-VLP outperforms the baseline models consistently. On VQA and NLVR^2, KB-VLP has better performance than all other baselines both on the Dev split and the Test split. The results on both datasets show that external knowledge can help text-image alignments and enhance the ability for visual understanding. On the GQA task, although our model does not outperform the Neural State Machine (NSM) (Drew A. Hudson, 2019), which is a neural network designed for structural learning and reasoning with strong structural priors, our model outperforms other VLP baselines. This provides evidence that adding access to additional knowledge at pretraining time can improve the structural learning and reasoning ability of the learned embeddings. In future, using stronger structural priors could be an interesting path to explore. On OK-VQA, we compare KB-VLP with the Oscar (Li et al., 2020b) pretraining model. The OK-VQA dataset requires models to use external knowledge to answer questions. As existing VLP models such as Oscar do not use such information during pretraining, KB-VLP provides a significant improvement on this dataset and task. We also perform qualitative analysis on the OK-VQA dataset in Appendix E and clustering analysis in Appendix F.

6. Conclusion

This paper proposes a new VLP method, KB-VLP, which in addition to text-image pairs, uses the text and image tags as queries to extract external knowledge from Wikipedia and takes that as additional input to the model. KB-VLP is pretrained with two novel pretraining tasks on a public corpus of $\sim$9M image-text pairs and finetuned on a set of vision-language downstream tasks. Experiments on four datasets demonstrate that KB-VLP has better performance compared to the baselines and in particular is successful at answering questions that require external knowledge.

References

Agarwal, O., Ge, H., Shakeri, S., and Al-Rfou, R. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *arXiv:2010.12688*, 2021.

Alberti, C., Ling, J., Collins, M., and Reitter, D. Fusion of detected objects in text for visual question answering. *Proceedings of EMNLP*, 2019.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *Proceedings of NeurIPS*, 2020.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. *Proceedings of ECCV*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*, 2018.

Drew A. Hudson, C. D. M. Learning by abstraction: The neural state machine. In *NeurIPS*, 2019.

He, B., Jiang, X., Xiao, J., and Liu, Q. Kgplm: Knowledge-guided language model pre-training via generative and discriminative learning. *arXiv:2012.03551*, 2020a.

He, B., Zhou, D., Xiao, J., Jiang, X., Liu, Q., Yuan, N. J., and Xu, T. Bert-mk: Integrating graph contextualized knowledge into pre-trained language models. *Proceedings of ACL*, 2020b.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv:2102.03334*, 2021.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv:1602.07332*, 2016.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. In *Proceedings of IJCV*, 2020.

Li, G., Duan, N., Fang, Y., Jiang, D., and Zhou, M. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of AAAI*, 2020a.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv:2004.06165*, 2020b.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context. *Proceedings of ECCV*, 2014.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Proceedings of NeurIPS*, 2019a.

Lu, J., Goswami, V., Rohrbach, M., Parikh, D., and Lee, S. 12-in-1: Multi-task vision and language representation learning. In *arXiv:1912.02315*, 2019b.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of NeurIPS*, 2011.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *Proceedings of NAACL*, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Rosset, C., Xiong, C., Phan, M., Song, X., Bennett, P., and Tiwary, S. Knowledge-aware language model pretraining. *arXiv:2007.00655*, 2021.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Li, J., Zhang, X., and Sun, J. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of ICCV*, 2019.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *Proceedings of ICLR*, 2020.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of ACL*, 2019.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *Proceedings of EMNLP*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.

Vrandecic, D. and Krotzsch, M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014.

Xiong, W., Du, J., Wang, W. Y., and Stoyanov, V. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *arXiv:1912.09637*, 2019.

Xu, S., Li, H., Yuan, P., Wang, Y., Wu, Y., He, X., Liu, Y., and Zhou, B. K-plug: Knowledge-injected pre-trained language model for natural language understanding and generation in e-commerce. *arXiv:2104.06960*, 2021.

Yamada, I. and Shindo, H. Neural attentive bag-of-entities model for text classification. In *Proceedings of The 23th SIGNLL Conference on Computational Natural Language Learning*, pp. 563–573. Association for Computational Linguistics, 2019.

Yamada, I., Shindo, H., Takeda, H., and Takefuji, Y. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 250–259. Association for Computational Linguistics, 2016.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., and Matsumoto, Y. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 23–30. Association for Computational Linguistics, 2020.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*, 2019.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

Yu, D., Zhu, C., Yang, Y., and Zeng, M. Jaket: Joint pre-training of knowledge graph and language understanding. *arXiv:2010.00796*, 2020.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. *arXiv:2101.00529*, 2021.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., and Gao, J. Unified vision-language pre-training for image captioning and vqa. *Proceedings of AAAI*, 2019.

Zhou, W., Lee, D.-H., Selvam, R. K., Lee, S., Lin, B. Y., and Ren, X. Pre-training text-to-text transformers for concept-centric common sense. *arXiv:2011.07956*, 2020.

Appendix:

A. Overview of the Knowledge Extraction Pipeline

Figure 2 presents the overview of our knowledge extraction pipeline. Details can be found in section 3.1.

B. Pretraining Implementation Details

KB-VLP uses the Transformer architecture from BERT-Base, initialized with parameters from BERT models. We use two linear projection matrices W_I and W_K to transform the image region features and knowledge embeddings the dimensionality of the BERT-Base model 768. As the knowledge embeddings contain both English words and knowledge graph entities. We add the positional embeddings to each knowledge embedding vectors so they follow the same positional order as the text from which they are derived. The AdamW optimizer is picked for model optimization and the learning rate is set to $5e^{-5}$. KB-VLP is trained for at least one million steps with a batch size of 768.

C. Adapting to Vision-Language Tasks

VQA. VQA is one of the most widely used visual question answering datasets. Following Antol et al. (2015), the model

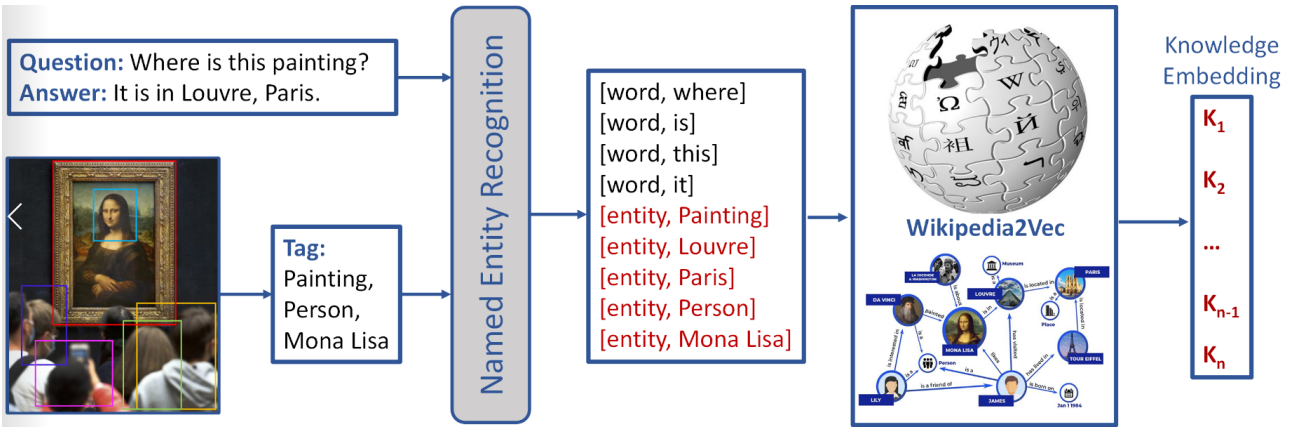

Figure 2. Overview of extracting knowledge on a text piece: given a text piece, we first perform named entity recognition on it and detect a set of entities and rest of words. Then, we use the Wikipedia2Vec tool to extract pretrained Wikipedia embedding vectors for each entity and word as our knowledge piece vectors for vision-language pretraining.

is required to answer natural language questions based on an image. Given an image and a question, the task is to select the correct answer from a multi-choice list. We use the VQA v2.0 dataset (Antol et al., 2015) for our experiments. VQA v2.0 is constructed based on the COCO image corpus and the dataset is split into a training set with 83k images and 444k questions, a validation set with 41k images and 214k questions and a test set with 81k images and 448k questions. The model picks the corresponding answer from a shared set of 3,129 answer candidates.

For VQA, the model takes one input sequence, which contains the concatenation of a question, object tags, region features, and extracted knowledge from the question and tags. Then KB-VLP [CLS] token is fed to a linear classifier for predicting the answer. Following Li et al. (2020b), we treat VQA as a multi-label classification problem. Each answer is assigned a soft target score based on its relevancy to the human answer responses, and we finetune the model by minimizing the cross-entropy loss against those soft target scores. At inference, we simply use a Softmax function for prediction.

GQA. GQA's evaluation is similar to VQA. The difference is that GQA dataset focuses on evaluating the reasoning capability of the model to answer a question. Our experiments are conducted on the public GQA dataset (Hudson & Manning, 2019). In the multi-choice setting, GQA requires model to choose an answer for each question from a shared answer set containing 1,852 candidate answers.

NLVR^2. The Natural Language Visual Reasoning for Real (NLVR^2) dataset (Suhr et al., 2019) asks a model to determine if a natural language statement is true or not of a pair of images. When fine-tuning, we construct two input sequences, each containing the concatenation of the text, an image and the extracted knowledge from text and the image. Then, the [CLS] tokens for the two sequences are concatenated as the joint input for a binary classifier to predict whether the statement is true.

OK-VQA. Outside Knowledge Visual Question Answering (OK-VQA) (Marino et al., 2019) is a new dataset that asks models to draw upon outside knowledge to answer questions. This dataset has 14,055 open-ended questions on COCO images and each question has 10 human annotated answers. We filter for questions with high-confidence answers in which 5 out of 10 annotated answers are the same (leaving 7,400 questions). As OK-VQA is designed to test how models use external knowledge, there are a substantial number of highly dissimilar answer candidates. This differentiates it from simpler multi-choice settings like VQA or GQA. Thus, we treat answer selection as an image-text retrieval task. During training, we formulate the task as a binary classification problem. Given an aligned tuple containing the image, question, answer, tags and extracted knowledge, we randomly select a different image, different knowledge or a different answer to construct a misaligned tuple. The [CLS] token is then used as the input to a binary classifier to predict whether the input is aligned or misaligned. During testing, we use the probability score to rank each answer for a given image-question pair and top-K retrieval results are used as the metric for evaluation. Also, we test our model on the whole testing set and calculate the accuracy based on the method described in Marino et al. (2019).

htb

Model	Oscar-B	KB-VLP	
Plants and Animals	0.32	0.43	+0.11
Science and Technology	0.25	0.35	+0.10
Sports and Recreation	0.38	0.43	+0.05
Geo, History, Lang, and Culture	0.34	0.48	+0.14
Brands, Companies, and Products	0.28	0.30	+0.02
Vehicles and Transportation	0.32	0.46	+0.13
Cooking and Food	0.34	0.44	+0.10
Weather and Climate	0.36	0.45	+0.09
People and Everyday	0.34	0.39	+0.05
Objects, Material and Clothing	0.39	0.35	-0.04
Other	0.36	0.43	+0.07

Table 2. Accuracy of each question type in OK-VQA.

D. Fine-tuning Settings

VQA. During training, we randomly sample a set of 2k images from the validation set as our validation set, the rest of images in training and validation sets are used in the VQA fine-tuning. We finetune the model for 30 epochs with a learning rate of $5e^{-5}$ and a batch size of 192.

GQA. The inputs for the GQA dataset are similar to the inputs for VQA, which is the concatenation of the question, object tags, region features, and extracted knowledge pieces. We finetune KB-VLP on an unbalanced "all-split" set for seven epochs with a learning rate $5e^{-5}$ and the batch size of 192.

NLVR^2. On the NLVR^2 dataset we follow the method described in C. We finetune the model for 35 epochs with a learning rate of $5e^{-5}$ and a batch size of 144.

OK-VQA. After filtering the question-answer pairs with high-confidence, the training set contains 4,690 questions and the testing set contains 2710 questions. We finetune KB-VLP 200 epochs on the filtered dataset with batch size 128. We use the learning rate $2e^{-5}$ and linearly decreases. We finetune the baseline model with the same parameter setting.

E. Qualitative Analysis on Experiments

Category Results on OK-VQA. Here we present qualitative analyses to illustrate how external knowledge influences the output of the pretraining model. We choose OK-VQA dataset for the qualitative analysis because the questions in this dataset most clearly benefit from external knowledge. Based on the types of knowledge required, questions in OK-VQA are categorized into 11 categories and the accuracy results of each category are reported in Table 2. In most categories, KB-VLP outperforms the Oscar model. This observation illustrates that the external knowledge used in KB-VLP includes many different aspects. Specifically, on categories "Plants and Animals", "Geo, History, Lang, and Culture" and "Vehicles and Transportation", KB-VLP provides the most significant improvements.

Correct Examples from KB-VLP. Existing VLP models are not able to learn much additional knowledge about these categories from general image captioning or visual question-answering datasets. The knowledge embeddings used in KB-VLP provide structural relations among entities that cannot be reflected from image-text pairs. Figure 3 has three examples comparing the answers generated by KB-VLP and Oscar. Comparing their generated answers, we find that the Oscar model is limited to visual detection and KB-VLP has stronger visual understanding and reasoning ability. For example, in the second example, the generated answer from Oscar is "Salty" instead of the correct answer "Rough". Presumably the Oscar model detects the sea and associates this with a frequently co-occurring word rather than correctly inferring the context of the question. Similarly, in the first example, Oscar thinks the city is "Chinatown" instead of "Tokyo". Without external knowledge, Oscar may not be able to learn the knowledge that Chinatown is not a city but Tokyo is a city.

Figure 4 presents more examples highlighting cases where KB-VLP generates the correct answer but Oscar does not. These examples are from several different categories and external knowledge is important to answer them. For example, the first question in the top row "What event is this?" requires the model to know that a concert usually has musicians and an audience. The second question in the bottom row "What item in this room is usually to wash hands?" requires knowledge about how the objects in the scene are used. The second question in the top row asks the model to find similar feline animals like the cat, which can only be provided using external knowledge about other types of cats.

Analysis on Failed Examples. Oscar only outperforms KB-VLP on questions about "Objects, Material and Clothing". One potential reason for this is that these questions require reliable object detection and rich knowledge about how they are used. This higher learning complexity may influence KB-VLP. It is possible that a larger knowledge base and more pretraining steps would help improve performance in this regard. Figure 5 presents three examples that KB-VLP fails to generate correct answers. These examples reflect that rich external knowledge vectors in KB-VLP may increase the complexity of visual understanding. For instance, in the first image, KB-VLP generates the answer "Electric Motor" instead of the correct answer "Key", because the model might learn high correlations between motorcycle and electric motor from external knowledge vectors

Figure 3. Three examples from OK-VQA that KB-VLP model generates correct answer but Oscar does not. Comparing the generated answers from KB-VLP and Oscar indicates that Oscar model is limited to visual detection and KB-VLP has stronger reasoning and understanding ability.

but fail to ground the knowledge to the question. Similarly, KB-VLP generates the "Playboy Bunny" instead of the correct answer "Mickey Mouse" and "Jewelry" instead of "Cloth". Although KB-VLP fails to generate correct answers on these examples, the analysis on these failed samples demonstrates that external knowledge can enhance the knowledge-awareness of existing VLP models.

F. Clustering Analysis on Image Representations from KB-VLP

To better understand the representations in KB-VLP, we randomly pick 50 image samples for each category in COCO dataset. We use t-SNE algorithm on the output representations of image samples and present the 2-D clustering visualization. Some interesting subset clustering results are provided in Figure 6, 7. In Figure 6, the samples of wine and beer are close with each other and samples of apple and vegetable are also close. They are rarely in one image but external knowledge embeddings provide latent relations. Apple is also relatively close with mobile phone because Apple is also a brand name, which has mobile phone products. In Figure 7, these four categories are all related to transportation. Boat and bus are closer than the other two because they share more similarities.

Broader Impacts

Multi-modal language and vision understanding has many applications. Examples include: information retrieval and tagging and designing accessible interfaces (i.e., image descriptions and closed captioning). However, we need to carefully understand the limitations and problems presented by the data that these methods are typically trained on. Datasets are often not representative of all people and demographic groups. A dataset crawled from the Internet is more likely to capture affluent western concepts and examples. While it is very challenging to create truly representative data, we can characterize datasets to help avoid models trained on them being applied in ways that are inappropriate. The fact that these datasets are not representative of all groups is one limitation of our work. Before a system such as the one presented here is deployed more work would need to be done to understand how such a model, in the context of an application, may disadvantage or advantage certain people. Training large models often consumes a lot of power and we must not neglect the environment impact of this process. During our experiments we made every effort to use the computational resources efficiently.

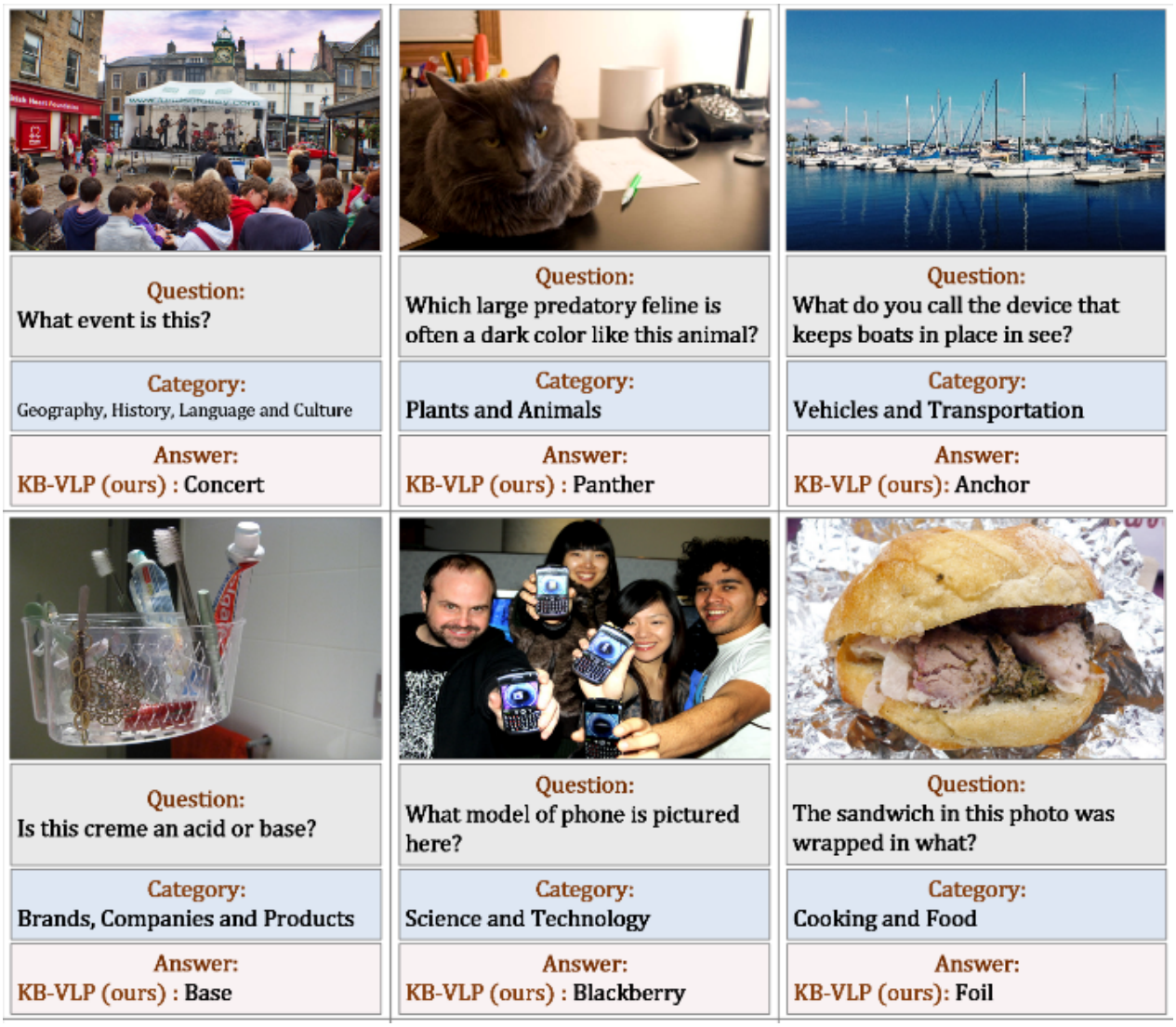

Figure 4. Examples from OK-VQA that KB-VLP model generates correct answer but Oscar does not.

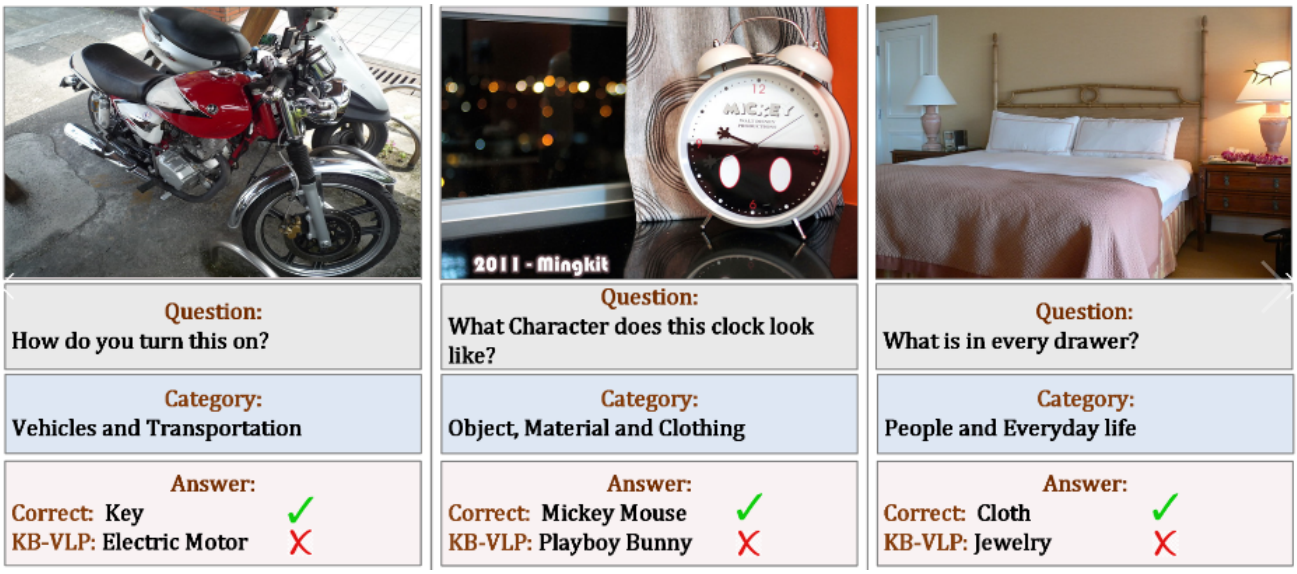

Figure 5. Three examples from OK-VQA that KB-VLP does not generate correct answers. In the first example, KB-VLP generates "Electric Motor" instead of the correct answer "Key". In the second example, KB-VLP generates "Playboy Bunny" instead of "Mickey Mouse" and in the third example, the model generate "Jewelry" instead of "Cloth".

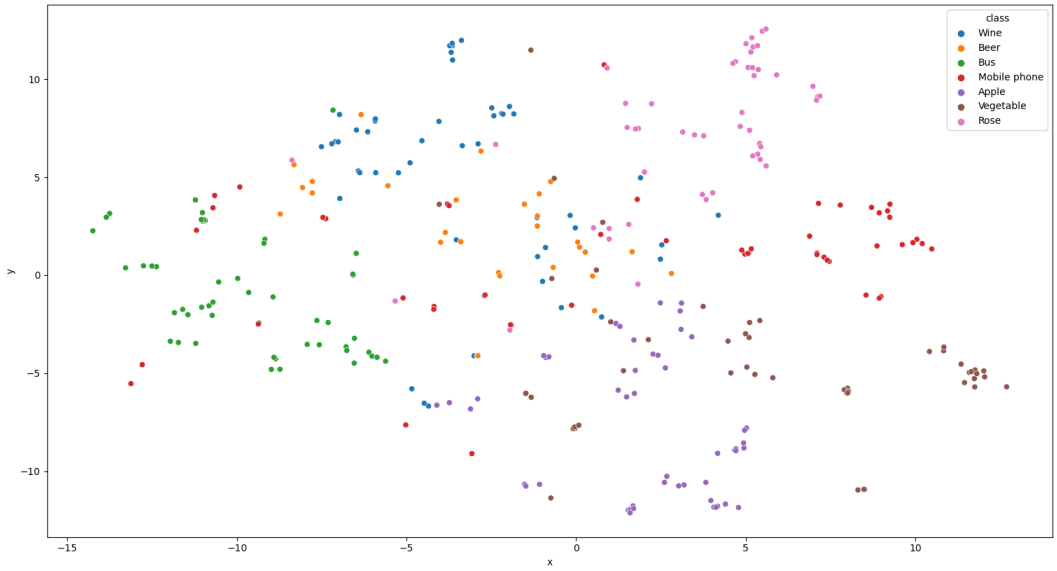

Figure 6. 2-D t-SNE visualization results on a subset of categories from COCO dataset. Wine and Beer are close with each other, Apple and Vegetable are close with each other, and Apple and Mobile Phone are relatively close. Although each pair of categories are rarely in one image, the external knowledge embeddings provide the latent relations for categories. Wine and beer are similar drinks. apple and vegetable share similar properties. Apple is also the brand and has high correlations with mobile phone.

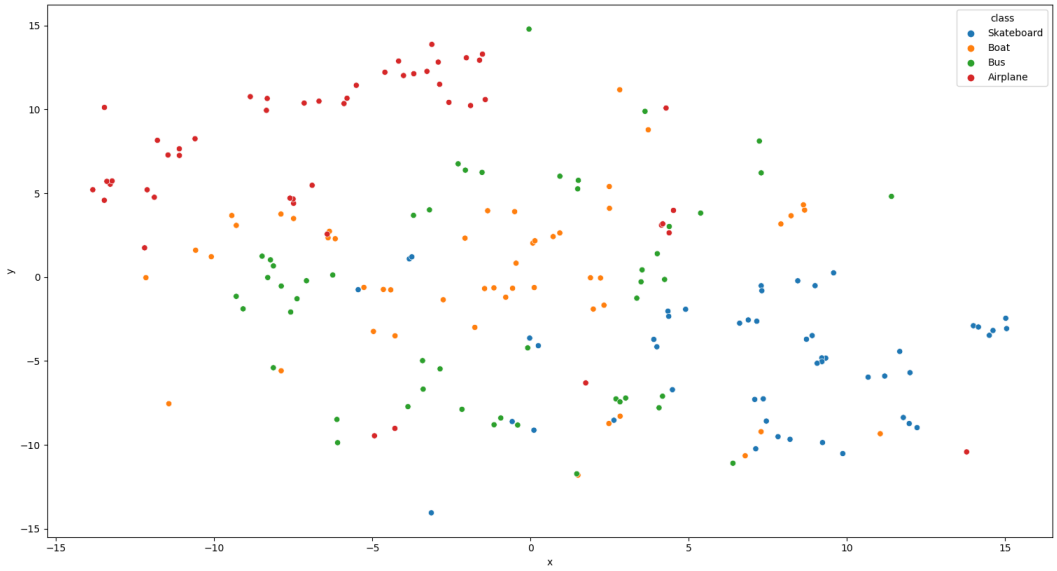

Figure 7. These four categories are all related to transportation. Boat and Bus are closer than the other two because they share more similarities.