Fatality Killed the *Cat* or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts

Agostina Calabrese, Michele Bevilacqua and **Roberto Navigli**
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{calabrese.a,bevilacqua,navigli}@di.uniroma1.it

Abstract

Thanks to the wealth of high-quality annotated images available in popular repositories such as ImageNet, multimodal language-vision research is in full bloom. However, events, feelings and many other kinds of concepts which can be visually grounded are not well represented in current datasets. Nevertheless, we would expect a wide-coverage language understanding system to be able to classify images depicting RECESS and REMORSE, not just CATS, DOGS and BRIDGES. We fill this gap by presenting BabelPic, a hand-labeled dataset built by cleaning the image-synset association found within the BabelNet Lexical Knowledge Base (LKB). BabelPic explicitly targets non-concrete concepts, thus providing refreshing new data for the community. We also show that pre-trained language-vision systems can be used to further expand the resource by exploiting natural language knowledge available in the LKB. BabelPic is available for download at http://babelpic.org.

1 Introduction

There is growing research interest in developing effective systems capable of achieving some understanding of the content of an image. As in most fields of applied AI, this requires annotated data to train a supervised system on. While ImageNet[1] (Deng et al., 2009), one of the most influential projects in computer vision, was undeniably an important milestone towards image understanding, there is still a lot of ground to be covered. ImageNet's initial aim was to collect pictures for most WordNet synsets (Miller, 1995). Yet, at the time of writing, only some 21,841 nominal synsets are covered according to ImageNet's official website.

One issue with ImageNet and most other image repositories like COCO (Lin et al., 2014) and Flickr30kEntities (Plummer et al., 2015) is their focus on concepts denoting concrete, tangible things, such as CAT, TRAFFIC LIGHT and so on. Concepts whose denotation is not clearly identifiable with a set of objects having distinct boundaries, such as events (e.g., FATALITY, COMPETITION), emotions (e.g., SADNESS) and psychological features (e.g., SHARPNESS), have enjoyed less attention. For lack of a better term, we will henceforth refer to them as *non-concrete* (NC) concepts.

On one hand, the inclusion of NC concepts would be an important step towards wide-coverage image semantic understanding. On the other hand, it also goes in the same direction as recent multimodal language-vision approaches, e.g., mono- and cross-lingual Visual Sense Disambiguation (Barnard and Johnson, 2005; Loeff et al., 2006; Saenko and Darrell, 2008; Gella et al., 2016, 2019). Taking into account NC concepts could also be of crucial importance for fascinating language-focused applications, such as Multimodal Machine Translation. Last but not least, NC concepts would represent a significative benchmark for real-world multimodal applications. In fact, traditional computer vision approaches rely on the detection of objects within the image, but many NC concepts are not well described by a bag of objects. Consider, for instance, Figure 1. The two images illustrate different NC concepts (i.e., HIGH JUMP and POLE VAULT) which are different configurations of the same elementary objects (i.e., PERSON, ROD, BLEACHERS). Thus, NC concepts require complex image understanding, integrating a fair amount of common sense knowledge.

As a contribution towards this goal of expanding the scope of research, we introduce BabelPic, the first dataset for multimodal language-vision tasks with a focus on NC concepts and that is also linked to WordNet. BabelPic has been built by manually validating synset-image associations available in

[1] http://www.image-net.org

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4680–4686
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

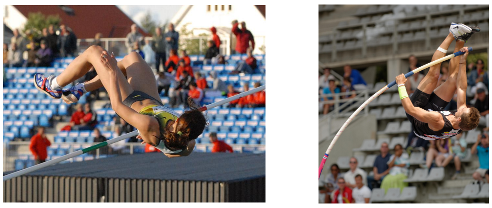

Figure 1: Two images described by the same bag of visual words but illustrating different NC concepts (i.e., high jump and pole vault).

BabelNet (Navigli and Ponzetto, 2012), a large multilingual resource linking WordNet to Wikipedia and other resources.

Furthermore, we provide a methodology to extend the BabelPic coverage to all the BabelNet synsets. To this end, we adapt the recently introduced Vision-Language Pre-training (VLP) model (Zhou et al., 2020). We define the verification of synset-image associations as a Visual Question Answering (VQA) task with two possible answers. The evaluation demonstrates that our methodology achieves high performances on zero-shot classification as well, thus enabling verification across the inventory. Thanks to the automatic production of a silver dataset, BabelPic constitutes a significant extension of ImageNet. A few examples from BabelPic (both gold and silver) are shown in Figure 2.

2 Related Work

To the best of our knowledge, no dataset of annotated images exists which has a focus on NC nominal and verbal concepts and is also linked to Lexical Knowledge Bases (LKB) such as WordNet and BabelNet. For example, the very popular ImageNet dataset, which includes images belonging to around 21,800 categories organized according to the WordNet nominal hierarchy, offers only sparse coverage of NC concepts. JFT (Hinton et al., 2015; Chollet, 2017; Sun et al., 2017) is an internal dataset at Google containing 300M images annotated with over 19,000 classes including objects, scenes (e.g., SUNSET), events (e.g., BIRTHDAY) and attributes (e.g., RED). JFT differs from our work in not being linked to an LKB and in not being publicly released. The Open Images dataset (Kuznetsova et al., 2018) contains 9M images annotated with 19,794 classes taken from JFT. While Open Images does contain NC labels, the classes are not linked to an LKB, thus limiting their usefulness. The Tencent ML-Images dataset (Wu et al., 2019) was created starting from a subset of ImageNet and Open Images and includes images annotated with 11,166 categories, which are then linked to WordNet synsets. The dataset differs from our work since any NC label has been explicitly discarded. Our work is in some sense similar to MultiSense (Gella et al., 2019) and VerSe (Gella et al., 2016), two datasets including images annotated with verbal senses. However, MultiSense is not directly linked to an LKB and neither of these two datasets deals with nominal synsets. Finally, we note that datasets including images annotated with object-level categories (Lin et al., 2014; Plummer et al., 2015) or videos (Loui et al., 2007; Dollár et al., 2009; Moneglia et al., 2014; Heilbron et al., 2015; Abu-El-Haija et al., 2016) are outside the scope of this work, since we are only interested in the main NC concepts depicted within images.

3 Gold Dataset

BabelPic is built by exploiting the link between WordNet (Miller, 1995) and Wikipedia within BabelNet[2] (Navigli and Ponzetto, 2012). Our approach is organised in a three-step process. First, we select a set of NC synsets from WordNet, on the basis of both their paradigmatic nature and relations in the knowledge base. Second, we gather all the corresponding images in BabelNet, which are themselves mostly taken from Wikipedia pages. Third, we manually validate the synset-images mapping. Note that, having defined the task as a validation of concept-image associations, we do allow images to be mapped to more than one concept and vice versa. For instance, both images in Figure 1 could be mapped to the concept COMPETITION as well. The result is a gold dataset containing 2,733 synsets and 14,931 images.

3.1 Synset selection

We decided to build our gold dataset starting from concepts related to events and emotions because these have been shown to be the most appealing NC concepts for the multimodal and vision communities (see Section 2). As a first step towards this goal, we select the nominal synsets belonging to the transitive closure of the hyponymy relation, rooted in the following set of WordNet synsets: {*feeling.n.01*,*event.n.01*}. To ensure that only NC concepts are selected, we filter out any synset connected by the hypernymy relation to at least one of the following synsets: *physical_entity.n.01*,

[2] https://babelnet.org

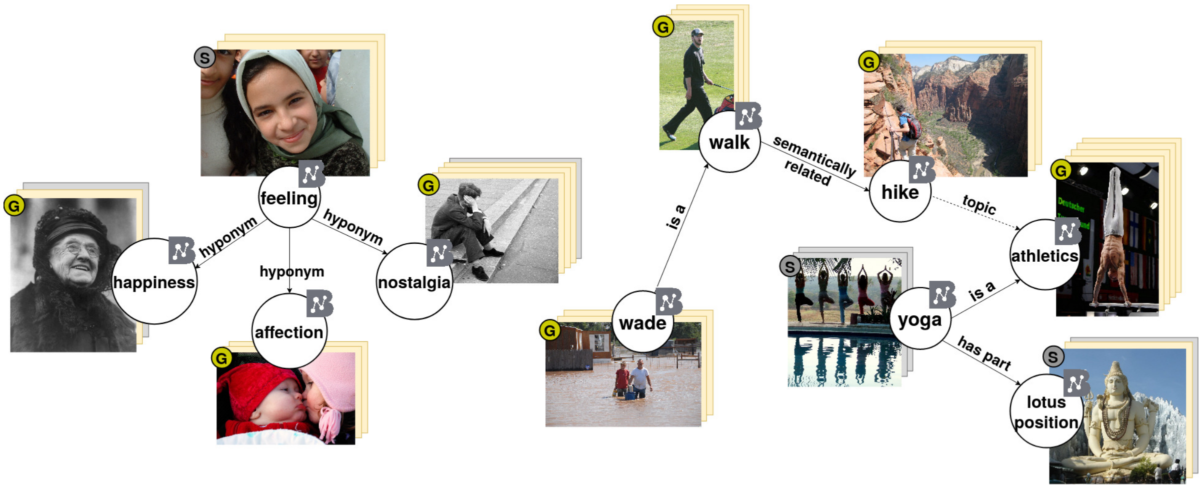

Figure 2: A few examples from BabelPic, both gold (G) and silver (S).

shape.n.02, *color.n.01*. This is done in order to discard concepts denoting tangible things that inherit from *abstraction.n.06* in WordNet (e.g., THUNDERBOLT). Furthermore, we select all the synsets belonging to the following WordNet lexicographer files: *verb.competition*, *verb.motion* and *verb.social*. This is done to create a dataset with an explicit focus on events, properties and verbs.

As a second step, we discard all the concepts belonging to either the mathematics or the physics domains since images are often not relevant (e.g., ROUNDING). Finally, we associate each selected synset with the first 15 corresponding images in BabelNet 4.0. Note that, in order to improve the quality of the dataset, we filter out images on the basis of simple heuristics. For example, we filter out all images where transparency is used and at least half of the pixels are white-coloured, as these are not likely be relevant. Most of the noise images from Wikipedia are removed as a result of this step.

3.2 Manual validation

The synset-image associations found are manually validated during phase 3. We have decided to use the services of two expert annotators who are familiar with the BabelNet resource, and the whole annotation process is performed through an *ad hoc* graphical interface. Annotators are shown tuples in the form $\langle s, l, g, i \rangle$, where s is the target synset, i is a candidate image for s, and l and g are, respectively, the main lemma and gloss (i.e., definition) for s. Annotators are asked to answer the question “is i pertinent to g?”. Possible answers are *yes* (i.e., i is an illustration of g), *no* (i.e., i is either not pertinent or in contradiction with g) and *discard* (i.e., i is a bad image). To maximize coverage, each annotator is assigned roughly half of the concept-image association candidates. However, in order to establish and agree on possible useful guidelines for the evaluation, annotators are asked to collaboratively perform the validation of a first sample of 500 instances. We also provide them with a few extra directions. For instance, we ask them to discard images in which the association cannot be verified without reading text depicted in the image. In addition to this collaboratively annotated sample, we select an intersection of 100 annotation instances which we then use to obtain an inter-annotator agreement figure. The level of agreement achieved is 80.39%, with a κ value of 0.6078 (moderate agreement). As for these shared examples, we include in our gold dataset only those instances that have been approved by both annotators. Our gold dataset is hence composed of all the validated synset-image associations.

4 Model

Since manual validation is time consuming, we are interested in developing a methodology for the automatic verification of synset-image associations. In the recent past there has been a great research effort to develop models for vision-language pretraining. Many such models (e.g., VLP (Zhou et al., 2020), VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019)) are built upon BERT (Devlin et al., 2019), a popular system for contextualized embeddings. BERT-based models achieve state-of-the-art scores on many language-vision tasks, hence they represent a promising resource for our task.

The system that we use to perform classification is the fine-tuned VLP model. Despite the fact that LXMERT (Tan and Bansal, 2019) achieves a slightly higher score on yes/no questions on the VQA 2.0 dataset (Goyal et al., 2017), our preference goes for the VLP system since it is pre-trained on a wider and more general dataset. More specifically, the VLP model is pre-trained on Conceptual Captions (CC) (Sharma et al., 2018), a dataset including more than 3M image-caption pairs, using two unsupervised vision-language tasks: bidirectional and sequence-to-sequence masked language prediction. The input images are preprocessed using Faster R-CNN (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017; Anderson et al., 2018), hence obtaining 100 object regions per image. The model input consists of both class-aware region embeddings and word embeddings, the former obtained by combining the corresponding region features with the probability of each object label and region geometric information. Furthermore, a Multi-Layer Perceptron (MLP) is trained during the fine-tuning phase in order to select the chosen answer starting from the hidden state of the encoder.

In order to adapt the VLP model to extend the BabelPic coverage to all the BabelNet synsets, we define the verification of synset-image associations as a VQA task with two possible answers. More specifically, we define a question template as in the following:

“Does the image depict l (g)?”

where l is the main lemma and g is the WordNet gloss of the target synset. We instantiate our template for each synset-image pair in the dataset, thus obtaining a textual question for each instance. We set the ground truth answers to either “yes” or “no”, hence reducing our classification task to VQA.

5 Experiments

To test the reliability of our approach for the automatic verification of concept-image associations we experiment in a zero-shot setting (see Section 5.3). As a first step toward this goal, we need to augment our dataset with negative instances (see Section 5.1) and select the most suitable VLP version (see Section 5.2). A deeper analysis of how the sampling of negative instances affects the performances of the system is described in Section 5.4.

5.1 Setting

In order to evaluate our methodology for the automatic verification of synset-image associations, we need to define a procedure for the generation of negative instances (i.e., irrelevant $\langle$synset, image$\rangle$ pairs). More specifically, we define a negative instance $\langle s, i\rangle$ by picking two different synsets s and s' and an image i associated with s' from our gold dataset. Negative instances can be distinguished on the basis of the relation connecting s to s':

Sibling: there exists a synset s'' in BabelNet s.t. both s and s' are connected to s'' by the hypernymy relation (e.g., FUN RUN and MARATHON).

Polysemy: both s and s' contain the same lemma (e.g., the synsets of *swim.v.01* and *swim.v.02*).

Unrelated: there exists no relation connecting s to s' in BabelNet (e.g., RACING and GLADFULNESS).

Exploiting the WordNet relations as mentioned above is also very effective in handling any potential issue due to images that are instances of multiple concepts. For instance, the images in Figure 1 could never be used as negative examples for COMPETITION because of the hyponymy relation connecting this concept to HIGH JUMP and POLE VAULT. Moreover, we manually validated a sample of the negative examples in order to ensure the reliability of our methodology.

The result is a dataset which is perfectly balanced between the two output classes. We split the dataset into training, validation and test sets following the 80%/10%/10% rule. Each class is proportionally distributed between the splits, as well as the relations used to define the negative instances. In order to test the system’s capability to handle previously unseen concepts, we force both the validation and test sets to contain also instances referring to synsets that are not present in the training set. We refer to the subset of the test set given by these instances as the zero-shot test. Statistics are reported in Table 1.

5.2 Pre-Trained vs. Fine-Tuned

In this work we refer to the VLP[3] model (Zhou et al., 2020) pre-trained on CC and fine-tuned for the VQA task on the VQA 2.0 dataset as, respectively, P-VLP and F-VLP. Note that both P-VLP

[3] https://github.com/LuoweiZhou/VLP

Split	N	C	I	S(%)	P(%)
Training	23,891	2,618	13,311	10.20	1.95
Validation	2,986	1,442	2,740	10.18	1.98
Test	2,987	1,416	2,715	10.21	1.94
Zero-Shot	502	43	490	11.55	2.19

Table 1: Overview of the BabelPic's splits: number of instances (N), concepts (C), images (I) and distribution of instances labelled as *sibling* (S) and *polysemy* (P).

Model	Validation		Test		Zero-Shot	
	P	F1	P	F1	P	F1
P-VLP	71.93	78.97	72.48	79.33	71.43	77.90
F-VLP	76.14	77.50	75.94	75.99	77.67	71.67

Table 2: Precision and F1 scores (as percentages) on the verification of synset-image associations.

and F-VLP are then further fine-tuned for the verification of concept-image associations on BabelPic's training split. Our experiments show that both systems are reliable on our task, achieving precision and F1 scores that are over 70% on all the splits (see Table 2). However, the F-VLP model proves to be the most stable for the task. In fact, in a common use case scenario it is more important to accept only correct synset-image associations than it is to detect all the correct pairs. More specifically, we value precision over recall, and thus prefer the fine-tuned VLP model.

5.3 Zero-Shot Classification

Our main interest is in developing a model capable of annotating images with synsets even when the target concept is new to the system (i.e., zero-shot). As shown in the last column of Table 2, both the P-VLP and F-VLP models are robust to zero-shot classification, achieving scores that are comparable to the performances registered on the other splits. The F-VLP system, in particular, is able to verify the associations between unseen synsets and images with precision 77.67%, hence enabling the automatic extension of BabelPic to any other synset.

5.4 Fine-Grained Analysis

Finally, we analyse the system performances on the different types of negative instances. The accuracy scores achieved by F-VLP are listed in Table 3. As one would expect, when the input synset-image pair is *unrelated*, the system is able to correctly classify most of the instances. When considering the instances labelled as *sibling*, the difficulty level increases and F-VLP achieves an accuracy score of 62.07%. This is not surprising when it is considered that discriminating between images representing sibling concepts (e.g., DISAPPOINTMENT and BOREDOM) can be tricky for humans as well. Finally, the instances labelled as *polysemy* prove to be the hardest ones, demonstrating that BabelPic can be an interesting benchmark for Visual Sense Disambiguation as well. The performances achieved by P-VLP follow the same trend.

Relation	Validation	Test	Zero-Shot
Unrelated	83.98	83.63	89.01
Sibling	51.64	53.11	62.07
Polysemy	30.51	44.83	45.45

Table 3: Accuracy scores (as percentages) achieved by F-VLP on all the different types of negative instances.

6 Conclusions

In this work we introduced BabelPic, a new resource for language-vision tasks, built by validating the existing image-to-synset associations in the BabelNet resource. BabelPic is innovative in being the first dataset with a focus on nominal and verbal non-concrete concepts linked to the WordNet and BabelNet Lexical Knowledge Bases. Furthermore, we presented a methodology to extend the resource by fine-tuning VLP, a state-of-the-art pre-trained language-vision architecture. In our approach, we automatically verify the synset-image associations by exploiting the natural language definitions in WordNet, showing strong results on zero-shot classification as well. We exploited our method for the automatic generation of a wide-coverage silver dataset containing around 10,013 synsets. We make BabelPic (both gold and silver data) available to the community for download at `http://babelpic.org`.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A large-scale video classification benchmark. *CoRR*, abs/1609.08675.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.

Kobus Barnard and Matthew Johnson. 2005. Word sense disambiguation with pictures. *Artif. Intell.*, 167(1-2):13–30.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2009. Pedestrian detection: A benchmark. In *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 304–311.

Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation. 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 1998–2004, Minneapolis, Minnesota. Association for Computational Linguistics.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–192, San Diego, California. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 961–970. IEEE Computer Society.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proceedings of Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 547–554, Sydney, Australia. Association for Computational Linguistics.

Alexander C. Loui, Jiebo Luo, Shih-Fu Chang, Dan Ellis, Wei Jiang, Lyndon S. Kennedy, Keansub Lee, and Akira Yanagawa. 2007. Kodak's consumer video benchmark data set: concept Definition and annotation. In *Proceedings of the 9th ACM SIGMM*

International Workshop on Multimedia Information Retrieval, MIR 2007, Augsburg, Bavaria, Germany, September 24-29, 2007, pages 245–254. ACM.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 13–23.

George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014. The IMAGACT visual ontology. An extendable multilingual infrastructure for the representation of lexical encoding of action. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3425–3432.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Kate Saenko and Trevor Darrell. 2008. Unsupervised learning of visual sense models for polysemous words. In *Proceedings of the Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 1393–1400.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the 2017 IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 843–852. IEEE Computer Society.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.

Baoyuan Wu, Weidong Chen, Yanbo Fan, Yong Zhang, Jinlong Hou, Jie Liu, and Tong Zhang. 2019. Tencent ML-Images: A large-scale multi-label image database for visual representation learning. *IEEE Access*, 7:172683–172693.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*.