On the (In)Effectiveness of Images for Text Classification

Chunpeng Ma[1], **Aili Shen**[2], **Hiyori Yoshikawa**[1],
Tomoya Iwakura[1], **Daniel Beck**[2], **Timothy Baldwin**[2]
[1] Fujitsu Laboratories LTD., Japan
[2] The University of Melbourne, Australia
{ma.chunpeng, y.hiyori, iwakura.tomoya}@fujitsu.com
ailis@student.unimelb.edu.au, {beck.d, tbaldwin}@unimelb.edu.au

Abstract

Images are core components of multi-modal learning in natural language processing (NLP), and results have varied substantially as to whether images improve NLP tasks or not. One confounding effect has been that previous NLP research has generally focused on sophisticated tasks (in varying settings), generally applied to English only. We focus on text classification, in the context of assigning named entity classes to a given Wikipedia page, where images generally complement the text and the Wikipedia page can be in one of a number of different languages. Our experiments across a range of languages show that images complement NLP models (including BERT) trained without external pre-training, but when combined with BERT models pre-trained on large-scale external data, images contribute nothing.

1 Introduction

Combining data from multiple modalities (e.g., text, images, categorical metadata, or user interaction features) has become commonplace in artificial intelligence. In NLP, examples include multi-modal machine translation (MMT) (Elliott et al., 2016; Elliott, 2018), visual question answering (VQA) (Goyal et al., 2017; Johnson et al., 2017), visual commonsense reasoning (VCR) (Zellers et al., 2019; Geva et al., 2019), and multi-modal pre-training (Lu et al., 2019; Chen et al., 2019).

While tasks such as VQA and VCR are multi-modal in nature, there has been research on traditionally text-based tasks such as text classification (Shen et al., 2020; Huang, 2018) and word embedding learning (Bruni et al., 2014) which has demonstrated that the addition of images boosts performance. At the same time, however, there is evidence of images providing no additional information, e.g. Caglayan et al. (2019) show that MMT models learn to ignore visual content when trained on a parallel corpus of image captions (Elliott et al., 2016). These mixed findings raise the question of *when* visual context is actually useful in NLP.

In this work, we take a first step towards answering this question, in focusing on the task of text classification, which has traditionally been addressed using textual data only. We identify two gaps in the literature on multi-modal NLP: (1) no results for pre-trained language models (LMs); and (2) no results for languages other than English. The first is important in terms of updating the research relative to state-of-the-art approaches, while the second relates to the question of how "language-independent" systems actually are (Bender, 2011). We fill these gaps via a text classification task over Wikipedia articles (Sekine et al., 2019). Our main findings are: (1) while images do help in a traditional supervised learning setting, *their utility disappears almost completely* when combined with a pre-trained LM; and (2) this phenomenon *is not restricted to English*, and generalises across a variety of languages from different families.

2 Task Description

This research is couched in the context of a shared-task dataset released by the SHINRA project (Sekine et al., 2019), aimed at classifying Wikipedia pages into fine-grained entity classes.[1] We chose this benchmark as many Wikipedia documents contain images, and data is provided for a total of 29 typologically-diverse languages.[2] The task is not trivial as it involves classifying Wikipedia documents into a set of 219 classes, with the possibility of multiple labels for a given document.[3]

[1] http://shinra-project.info/shinra2020ml/?lang=en

[2] Data is also provided for Greek but we do not include it in our experiments because there was no officially preprocessed data available for this language.

[3] See: http://ene-project.info/ene8/?lang=en

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 42–48
April 19 - 23, 2021.

hi	th	ar	da	bg	ro	he	tr	id	vi
30,546	59,790	73,053	86,237	89,016	92,001	96,433	111,591	115,642	116,279
hu	cs	no	ca	fi	uk	fa	sv	ko	nl
120,294	125,958	135,934	139,031	144,749	167,236	169,052	180,947	190,806	199,982
pt	pl	ru	es	zh	it	de	fr	en	
217,895	225,551	253,011	257,834	267,106	270,192	274,731	318,827	439,351	

Table 1: Statistics of annotated data for each language.

The number of annotated pages for each language in the SHINRA dataset is shown in Table 1 (sorted according to the number of pages). In addition to these annotated datasets — which form the basis of the experiments in this paper — there is a large amount of evaluation data for each language. In an evaluation campaign over these evaluation datasets, we achieved first place across 4 languages: English, Italian, Spanish and Catalan (Yoshikawa et al., 2020).

The SHINRA dataset contains only textual information from the original documents. In order to add images, we extract the image links from the English Wikipedia dump of June 2020[4] using the `zim` library.[5] The extracted images are then linked with image links in the source documents in the SHINRA dataset,[6] resulting in about 88% pages being augmented with images (noting that images are generally shared across Wikipedia pages for different languages other than English).

Out of the 30 languages in the original SHINRA dataset, we experiment primarily with Arabic ("ar"), English ("en"), Finnish ("fi"), Hindi ("hi"), and Mandarin Chinese ("zh"), selected to span five different language families and where the dataset size is relatively large. From the SHINRA data, we randomly sample 30k documents for each language, and construct a 80%/10%/10% fixed split for training/development/test in each language. We use a maximum of four images for each document.[7]

3 Baseline Experiments

Our first set of experiments is aimed at evaluating the empirical utility of images in the absence of pre-trained models. This is in line with previous work over similar text classification tasks (Shen et al., 2020; Huang, 2018).

Model and Features As our basic learner, we use a linear-kernel support vector machine (Cortes and Vapnik, 1995, SVM). For the textual inputs, we experiment with three representations: (1) a binary bag-of-words ("`BOW`"); (2) sent2vec ("`S2V`": Pagliardini et al. (2018)); and (3) `BERT` (Devlin et al., 2019). In this set of experiments, we train both `S2V` and `BERT` from scratch on the SHINRA training data only. We simply use the suggested configuration provided by developers, without any task-specific hyperparameter tuning. For `BERT`, we use the `[CLS]` token as the document representation. For each document, an image representation for each of the (up to) four images is generated. Specifically, following standard practice in the computer vision community, we firstly use the SIFT algorithm (Lowe, 1999) to extract hundreds of features, then use the K-means algorithm to cluster these features and generate frequency histograms, which are so-called visual bag-of-words (VBoW), and finally use an SVM to classify these histogram features. We also experiment with Faster R-CNN (Ren et al., 2015), pre-trained on Visual Genome (Krishna et al., 2017), following the settings of Anderson et al. (2018). We ensure the dimensionality of input features for the SVM and Faster R-CNN are the same (both are $1,024$), to remove this possible representational confound. Note that this is the externally pre-trained image model across all experiments, and that none of the text models in this first set of experiments involve pre-training on external resources (something we return to in Section 4).

Results and Analysis We report F_1 scores over the test set in Table 2. The main finding is that images improve performance in all settings, for all languages and both image representations. `S2V` and `BERT` both perform worse than the simple bag of words, because of the limited training data in each case. We would, of course, expect the models

[4] https://dumps.wikimedia.org/other/kiwix/zim/wikipedia/wikipedia_en_all_maxi_2020-06.zim

[5] https://github.com/openzim/libzim

[6] Because it is quite difficult to find correspondences between images and texts (Hessel et al., 2019), image links extracted are "document-level", instead of "sentence-level".

[7] When a document has less than 4 images, we pad the representation with blank images.

Text	Image	Language				
		ar	en	fi	hi	zh
BOW	—	65.1	72.3	72.1	67.1	74.7
BOW	SIFT+V	**68.2**	**74.1**	**73.2**	**69.0**	**76.0**
BOW	R-CNN	67.1	73.0	72.7	68.7	75.3
S2V	—	63.6	68.1	63.1	63.1	72.0
S2V	SIFT+V	66.0	70.2	66.3	66.9	72.9
S2V	R-CNN	65.4	69.0	65.0	65.2	72.3
BERT	—	59.1	65.3	51.9	60.5	68.6
BERT	SIFT+V	62.9	68.7	54.2	63.1	70.9
BERT	R-CNN	61.4	67.3	52.7	62.9	70.0

Table 2: F_1 score of the SVM models without external pre-training of the textual models, across the five languages. "SIFT+V" refers to the combination of SIFT and Visual Bag-of-Words features. "R-CNN" corresponds to features extracted from Faster R-CNN.

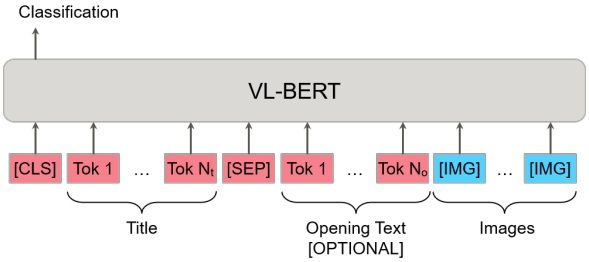


Figure 1: VL-BERT architecture applied to the SHINRA2020-ML task. The "opening text" segment are additional textual data obtained from the documents that are optional in our experimental setting.

to perform better with more extensive pre-training, as we return to explore in Section 4, but the focus here is on training of the textual models within the bounds of the training dataset.

Strikingly, the SIFT + Visual Bag-of-Words representation results in better performance than the pre-trained Faster R-CNN. A potential explanation is that Faster R-CNN is trained in a supervised way using Visual Genome (unlike the self-supervised setting of pre-trained BERT, for instance), over a set of labels that is not particularly well aligned with SHINRA (SHINRA includes many abstract classes such as RELIGION, NATIONALITY, and OFFENCE, whereas Visual Genome is focused on physical objects and attributes, and relations between objects; even among physical objects, SHINRA distinguishes between MEDICAL INSTITUTION, PUBLIC INSTITUTION, and RESEARCH INSTITUTE, most of which are represented simply as BUILDING in Visual Genome).

4 Adding a Pre-trained Textual Encoder

We next turn to a setting where we employ pre-trained textual models. This not only better reflects the state-of-the-art in text classification, but also allows us to investigate the effect of images under such conditions.

Model As the main backbone, we employ VL-BERT (Su et al., 2020), which uses a transformer to combine textual inputs and image embeddings within a BERT-style transformer, and has been shown to perform well on multimodal tasks. The visual embeddings are obtained from the combination of pre-trained Faster R-CNN and ResNet-101 (He et al., 2016), as illustrated in Figure 1. For the text modality, the input consists of two parts: the document title, and the opening text of the Wikipedia page in the form of the first 300 tokens. The token embeddings are obtained from a pre-trained BERT model, which is fine-tuned during training.[8] The full model is plugged into a one-layer feed-forward neural network (FFNN) with a 1,024d hidden layer, and training is performed by minimizing the cross-entropy over the SHINRA category labels. The model predict one label for each page. For the case of multi-label inputs, we choose one randomly as the "correct" label.

Results and Analysis Table 3 shows the performance of VL-BERT with different combinations of textual (document title = "T" and optionally the document body = "B") and image inputs, based on pre-trained BERT ("BERT_{pre}").

The first thing to notice is that the image-only model is well above the majority baseline, but well below the best multimodal model without an externally pre-trained text encoder from Table 2. This shows that images provide useful information for document classification, consistent with the earlier finding that images enhance the various text-only models. However, when combined with the externally pre-trained BERT_{pre} (over either the title only, or the title + document body), the utility of images is marginal at best. That is, the large-scale pre-training of BERT_{pre} both boosts overall performance, but much more surprisingly, removes any advantage from including images.

[8] We use bert-large-uncased for English, and bert-base-multilingual-uncased for the other languages, as obtained from https://huggingface.co/transformers/pretrained_models.html

Text	Image	ar	en	fi	hi	zh
—	✓	50.6	50.1	53.9	46.1	44.1
T	—	70.9	73.1	71.1	66.2	76.5
T	✓	70.8	73.2	71.2	66.7	76.7
T+B	—	**82.8**	88.7	87.7	**85.0**	**88.6**
T+B	✓	82.6	**88.8**	**88.0**	84.8	88.0
Best non pre-trained		68.2	74.1	73.2	69.0	76.0
Majority class		21.5	22.2	28.1	19.1	21.7

Table 3: F_1 scores for pre-trained VL-BERT. "T" = document title, and "T+B" = document title + body. We reproduce the best non-trained For comparison, we restate the result for the best non pre-trained model from Table 2, along with the majority class baseline.

Influence of the size of training data One hypothesis is that images are not useful due to the size of the training data (24k instances), and in lower-resource scenarios will improve performance. To test this, we perform additional experiments varying the training data size, ranging from 4k to 24k training instances, in steps of 4k.

Figure 2 plots the F_1 performance as the training set size increases. While we observe substantial improvements for the image-only approach (the bottom curve), the differences in the models with textual data are modest, and even in small-data settings, there is no real advantage in including images. We also separated the test data in terms of the number of images, and found no differences. See the Supplementary Material for details.

Results on the full SHINRA dataset In the previous experiments, we fixed the dataset size for all languages to control for training data volume. However, the SHINRA dataset includes many more documents for many of the languages. As a final experiment, we apply the VL-BERT models to the full dataset available for each language. The development and test data are also different in this configuration, so the results are not directly comparable with Tables 2 and 3.

In Table 4, we present results for BERT_{pre}, and mostly corroborate our earlier findings: while we do see improvements when including images in the case of the titles only, their utility decreases when we add the body of text for each document.

What caused the difference? Comparing the results from Sections 3 and 4, we see two main differences: the presence of external pre-training (BERT vs. BERT_{pre}), and the model architecture. To determine whether the model architecture is a cause of the performance difference, we train VL-BERT *from scratch*, using only text and images from the 24k training set used in Section 3.

Text	Image	ar	en	fi	hi	zh
—	✓	85.2	61.2	72.4	45.6	58.6
T	—	84.4	77.4	75.5	66.6	78.5
T	✓	86.9	79.1	75.8	67.3	80.7
T+B	—	94.7	90.3	91.7	85.8	89.8
T+B	✓	94.7	90.2	91.6	85.4	90.2

Table 4: Comparison of F_1 scores over the full SHINRA dataset for BERT_{pre}.

Text	Image	ar	en	fi	hi	zh
T+B	—	56.7	61.8	49.8	57.2	66.1
T+B	✓	58.6	63.2	52.4	60.1	68.7

Table 5: Comparison of F_1 scores for VL-BERT without external pre-training of BERT.

The results in Table 5 shows that even for VL-BERT, a neural-based model that is much more complex than the linear-kernel SVM, when BERT_{pre} is not used, images provide a gain in performance. Hence, having an externally pre-trained text encoder is the predominant determinant of whether visual content has utility in NLP tasks.

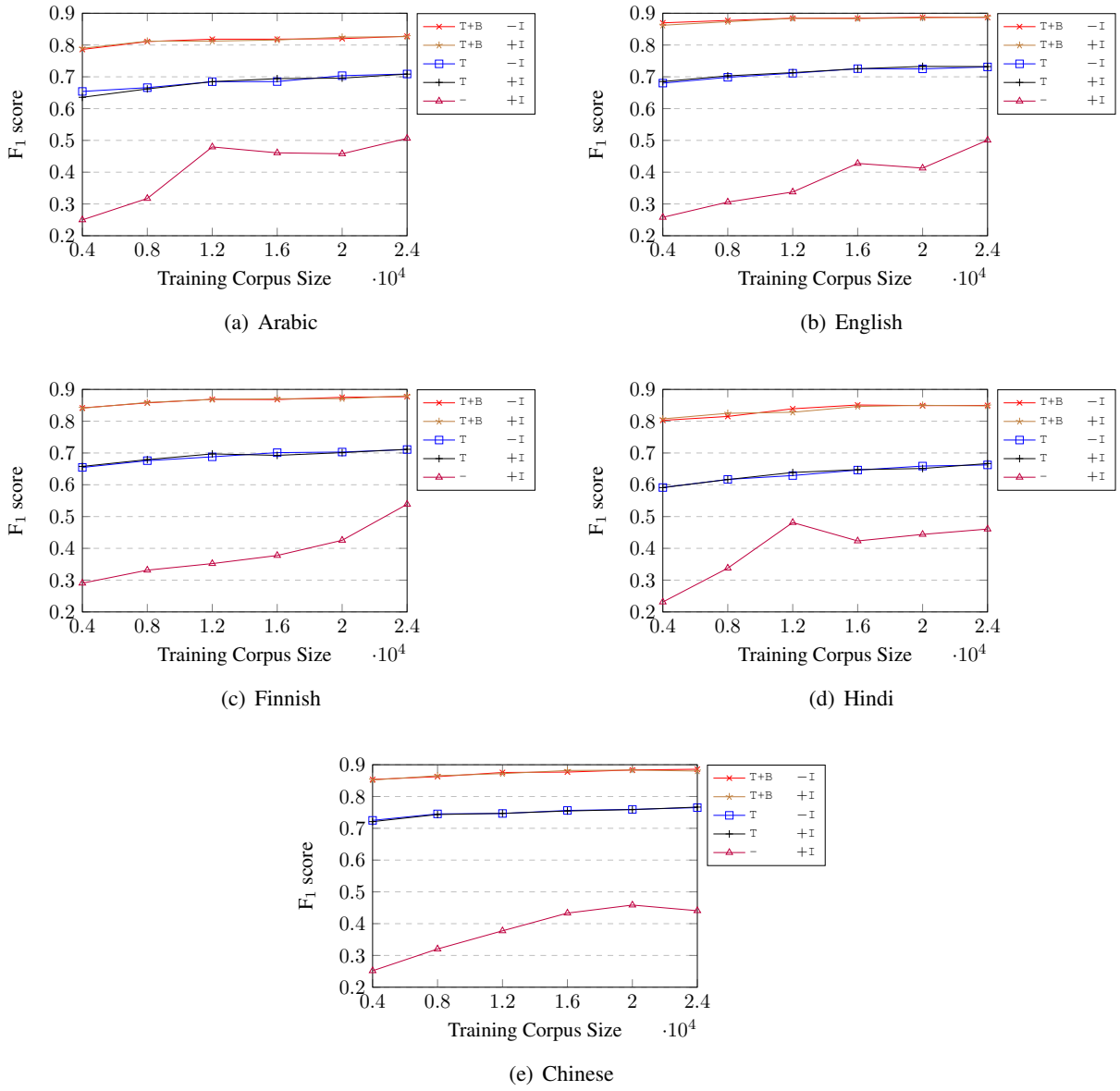
5 Discussion and Conclusion

We investigated the utility of images as a supplementary input for a text classification task, and found that although images have empirical utility in traditional supervised learning, when externally pre-trained language models are utilised, any advantage from the visual modality disappears. The results were remarkably consistent across different languages and different volumes of training data.

It is important to distinguish between "inherently multi-modal tasks" (e.g. VQA) and "potentially multi-modal tasks" (e.g. text classification) in drawing any broader conclusions about the (in)effectiveness of images. Here, a "potentially multi-modal task" in NLP means that the primary modality is text and the task is defined based on that single data modality, but there is potentially the option to include extra modalities such as images.

There remain a lot of open questions in more fully determining the (in)effectiveness of images for NLP tasks, even for text classification, such as:

- Due to the seeming redundancy between textual and visual representations of Wikipedia pages, is there any utility in multi-modal inputs for simple NLP tasks such as text clas-



(a) Arabic

(b) English

(c) Finnish

(d) Hindi

(e) Chinese

Figure 2: Model performance with different sizes of training corpus, with and without images (±I) and with and without text (in the form of the title ["T"] and optionally document body ["B"]).

sification in the era of large-scale pre-trained language models such as BERT and GPT-3 (Brown et al., 2020)?

- What performances do humans achieve in the single-modal setting and multi-modal setting? Can we get some insights by comparing the (potentially) different performances between humans and computers?
- Apart from images, what other modalities and forms of input (e.g. audio) could be effective in building better NLP models?
- Although pre-trained image models (e.g. Faster R-CNN) contribute a lot for vision tasks (e.g. object detection) and multi-modal tasks (e.g. VQA), for *"pure"* NLP tasks (e.g. text classification), they appear to work no better than traditional image representation feature extractors (e.g. SIFT). Why?
- In our experiments, we use at most 4 images for each page. Could instance selection enhance image utility?
- We focused on the text classification task, in classifying Wikipedia pages into different entities. Are our observations NLP task-independent?

References

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 IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. 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 4159–4170.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

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.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166.

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 IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

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.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. 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 2034–2045.

Jia-Bin Huang. 2018. Deep paper gestalt. *arXiv preprint arXiv:1812.08775*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Satoshi Sekine, Akio Kobayashi, and Kouta Nakayama. 2019. SHINRA: Structuring wikipedia by collaborative contribution. In *Automated Knowledge Base Construction (AKBC)*.

Aili Shen, Bahar Salehi, Jianzhong Qi, and Timothy Baldwin. 2020. A general approach to multimodal document quality assessment. *Journal of Artificial Intelligence Research*, 68:607–632.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hiyori Yoshikawa, Chunpeng Ma, Aili Shen, Qian Sun, Chenbang Huang, Guillaume Pelat, Akiva Miura, Daniel Beck, Timothy Baldwin, and Tomoya Iwakura. 2020. UOM-FJ at the NTCIR-15 SHINRA2020-ML task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, pages 201–207.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.