In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering

Peter Vickers[*] and **Nikolaos Aletras**[*] and **Emilio Monti**[+] and **Loïc Barrault**[*]

[*]Department of Computer Science
University of Sheffield
{pgjvickers1, n.aletras, l.barrault}
@sheffield.ac.uk

[+]Amazon United Kingdom
monti@amazon.co.uk

Abstract

Visual Question Answering (VQA) methods aim at leveraging visual input to answer questions that may require complex reasoning over entities. Current models are trained on labelled data that may be insufficient to learn complex knowledge representations. In this paper, we propose a new method to enhance the reasoning capabilities of a multi-modal pretrained model (Vision+Language BERT) by integrating facts extracted from an external knowledge base. Evaluation on the KVQA dataset benchmark demonstrates that our method outperforms competitive baselines by 19%, achieving new state-of-the-art results. We also perform an extensive analysis highlighting the limitations of our best performing model through an ablation study.

1 Introduction

Visual Question Answering (VQA) is a popular multi-modal task of answering a question about an image. It tracks both inter-modal interactions and reasoning capabilities of models (Wang et al., 2017; Marino et al., 2019). Recent studies have tested compositional reasoning (Johnson et al., 2016; Hudson and Manning, 2019) and the integration of external knowledge (Wang et al., 2017, 2016; Shah et al., 2019; Marino et al., 2019) for VQA. In this paper, we address Knowledge-aware VQA (KVQA) (Shah et al., 2019)[1] , defined as a VQA task where it is not reasonable to expect a model without access to a knowledge base to be able to answer the questions in the test set.

In a uni-modal textual context, both synthetic dataset (Kassner et al., 2020) and task-driven (Ding et al., 2020) studies of neural models have shown significant competence at symbolic reasoning. This is encouraging, as neural pretrained Language Models such as BERT (Devlin et al., 2019) achieve state-of-the-art results in a wide range of natural language inference tasks and benchmarks such as Natural Language Inference (Bowman et al., 2015). (Rajani et al., 2019) uses pretraining on a domain-specific dataset to improve CommonsenseQA by 10% absolute accuracy. Tamborrino et al. (2020) develop an improved training objective to improve COPA by 10% absolute accuracy.

Bouraoui et al. (2020) find that BERT is capable of relational induction, whilst Broscheit (2019); Petroni et al. (2020) find that BERT stores non-trivial world-knowledge.

Previous work has argued that restriction to a uni-modal context may itself impair reasoning performance (Barsalou, 2008; Li et al., 2020). In a bi-modal Vision + Language (V+L) context, datasets such as CLEVR and GQA allow for the evaluation of both model reasoning and language grounding. Within this setting, Ding et al. (2020) and Lu et al. (2020) show that appropriate neural models trained on large quantities of data can exhibit accurate reasoning.

In this paper, we propose a new method of applying a massively pretrained V+L BERT model (Chen et al., 2020) to the KVQA task (Shah et al., 2019). Our method is able to learn a set of reasoning types (confirming findings in Ding et al. (2020)) but can increase performance even more by incorporating external factual information. KVQA answers require attending to a knowledge base, allowing us to quantify the contribution of both explicit and implicit knowledge extracted from supervised training data. We also quantify the degree to which corpus bias makes certain question types harder, and outline how future datasets may be better balanced.

Our contributions are as follows:

- We perform factual integration into a V+L BERT-based model architecture VQA, leading to 19.1% accuracy improvement over previous baselines on KVQA.

[1]For data, examples, and licence information, please see https://malllabiisc.github.io/resources/kvqa/

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 468–475
August 1–6, 2021.

- We evaluate our model's reasoning capabilities through an ablation study, proposing explanations for poor performance on certain question types as well as highlighting our model's strong preference for text and facts over the image modality.

- We conduct a bias study of the KVQA dataset, revealing both strengths and potential improvements for future VQA datasets.

2 Related Work

VQA tasks explicitly encourage grounded reasoning (Antol et al., 2015), with emphasis on a variety of sub-domains, such as commonsense (Zellers et al., 2019), compositionality and grounding (Suhr et al., 2020), factual reasoning (Wang et al., 2017) or external knowledge reasoning (Wang et al., 2016; Marino et al., 2019; Shah et al., 2019).

State-of-the-art systems for external knowledge VQA are based on Memory networks (MemNet, (Weston et al., 2014)). In Shah et al. (2019), the facts are extracted from the Knowledge Graph (KG) by considering the visual (from image) and eventually textual (from Wikipedia caption) entities. They are then embedded using a Bi-LSTM encoder and fed into the memory. After the question is embedded in a similar way, the resulting representation is used to query the memory by soft attention. Several stacked memory layers are used to better model multi-hop facts.

Wang et al. (2016, 2017) introduce two datasets, KB-VQA and FVQA respectively, and address the task with systems that perform searches in a visual knowledge graph formed from the image and a KB. The question is first mapped to a query of the form ⟨visual object, relationship, answer source⟩, which is then used to extract the supporting facts from the KB. They report improved results when compared to systems using LSTM, SVM and hierarchical co-attention (Lu et al., 2016).

In Marino et al. (2019), the OK-VQA is presented with some baseline results obtained with MUTAN (Ben-younes et al., 2017), a multimodal tensor-based Tucker decomposition which models interactions between visual (from CNN) and textual (from RNN) representations. Those systems exhibit rather low performance compared to those obtained on standard VQA, demonstrating that the corpus requires external knowledge to be solved correctly.

Recent work has introduced methods to incorporate visual information to create Vision+Language BERT models through joint multimodal embeddings (Chen et al., 2020; Su et al., 2019; Lu et al., 2019). First, image and text are embedded into the same space, and then Transformer networks are applied as in the standard BERT model (Devlin et al., 2019).

Our work is most similar to that of Shah et al. (2019) since the same preprocessing pipeline is used. However, our system does not use a memory network, and instead relies on on a BERT-based model (UNITER, see section 3) to model the relationship between question, facts, and image with self-attention layers.

3 Methodology

To answer KVQA with Neural models, we first take the V+L BERT model UNITER (Chen et al., 2020) with the highest score on the commonsense VQA task, VCR (Zellers et al., 2019).

In order to allow UNITER to accept external KG facts, we cast these facts to a textual form 'Entity_1 Relation Entity_2'. To keep the input facts count small, we perform a *conditional search* of the KG. The KVQA task consists in finding a^*:

$$a^* = \underset{a \in A}{\operatorname{argmax}}\, p(a|q, i, K) \approx \underset{a \in A}{\operatorname{argmax}}\, p(a|q, i, k_{i,q}) \quad (1)$$

where a^* is the correct answer out of candidate set A; and q, i, and K are a question, image and knowledge base, respectively. As shown, we may reduce the KG through a conditional search to find the relevant subset of facts $k_{i,q}$.

To define the subset $k_{i,q}$, we follow Shah et al. (2019) in extracting all facts from the knowledge base that are up to two hops from any entities detected by the textual entity linking or the face detection.

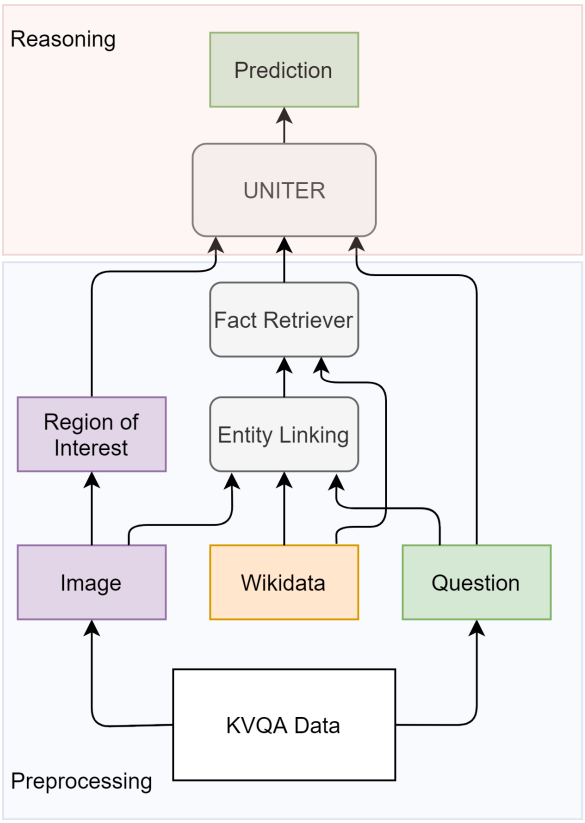

Figure 1: Our Model

Our model, as presented in section 2 consists of two stages: preprocessing, which implements relevant fact extraction, and reasoning, which selects an answer from the question, facts, and image features.

3.1 Preprocessing Stage

For preprocessing and fact acquisition, we broadly reproduce the fact and feature extraction process used in Shah et al. (2019). We perform object detection with the Faster R-CNN network (Ren et al., 2017). A seven-dimensional normalised size and location vector is concatenated with the Faster R-CNN features.

For person detection, we use MTCNN (Zhang et al., 2016) and Facenet (Schroff et al., 2015) models, pretrained on the MS-celeb-1M (Guo et al., 2016) dataset, to generate 128-dimensional embeddings. We predict names by nearest-neighbour comparison with the KVQA reference dataset. We treat the name identification as a multi-class classification problem, achieving a Micro-F1 of 0.539. Since this is lower than reported in Shah et al. (2019), we follow them in applying a textual entity linker (van Hulst et al., 2020) over supplied image descriptions. This setup achieves a per-image Micro-F1 of 0.686.

Normalised image location facts are generated from these detections, such as 'Barack Obama at 42 78', which would indicate that the centre bounding box for Barack Obama is at normalised (0-100) position x=42, y=78 of the image. We use the names of identified entities to query Shah et al.'s 2019 reduced Wikidata graph (Vrandečić and Krötzsch, 2014) up to two hops. The extracted facts are finally cast to the form 'subject relation object'.

3.2 Reasoning Stage

The neural model we use, UNITER, is pretrained on MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2016), Conceptual Captions (Sharma et al., 2018), and SBU Captions (Ordonez et al., 2011). It is a multi-task system that is trained on performing Masked Language Modeling, Image-Text Matching, and Masked Region Modeling (Chen et al., 2020).

4 Experimental Setup

We select the KVQA dataset for two reasons: to our knowledge, it is the largest external knowledge dataset (with 183k questions), and the questions are annotated with their reasoning types. We use accuracy as the evaluation metric and provide results over both the entire dataset and also for each question type as provided in the KVQA dataset.

The baseline systems for KVQA are those presented in (Shah et al., 2019) and discussed in section 2. The first baseline is a stacked BLSTM encoder, operating over question and facts. This system has an overall accuracy of 48.0% . The second is the MemNet architecture and has the previously highest performing baseline accuracy at 50.2%.

We use the UNITER_BASE pretrained model available at the ChenRocks GitHub repository[2] with custom classification layers (MLP +softmax output layer). For task training, we merge retrieved facts with the question, dividing each statement with the '[SEP]' token, following research that indicates that this token induces partitioning and pipelining of information across attention layers (Clark et al., 2019). The textual input stream is tokenised with the HuggingFace 'bert-base-uncased' tokeniser (Wolf et al., 2020). We set the maximum WordPiece sequences length to 412, the maximum visual objects count to 100, the learning rate to

[2] https://github.com/ChenRocks/UNITER

Question Type	Model		Entropy (Base 2)
	MemNet	**UNITER**	
1-Hop	61.0	**65.7**	7.8
1-Hop Counting	-	**78.0**	1.4
1-Hop Subtraction	-	**28.6**	4.3
Boolean	75.1	**94.6**	1.1
Comparison	50.5	**90.4**	2.1
Counting	49.5	**79.4**	2.3
Intersection	72.5	**79.4**	1.2
Multi-Entity	43.5	**77.1**	3.3
Multi-Hop	53.2	**87.9**	3.7
Multi-Relation	45.2	**75.2**	7.1
Spatial	**48.1**	21.2	11.5
Subtraction	**40.5**	34.4	6.0
Overall	50.2	69.3	7.6

Table 1: Results in terms of % accuracy of the considered systems break down into question types along with the question types distribution (last column).

Question Type	Q+F+I	Q+F	Q+I	F+I	Q	F	I
1-Hop	65.7	65.7	32.4	3.9	32.4	3.8	4.5
1-Hop Counting	78.0	78.0	30.3	0.0	30.3	0.0	0.0
1-Hop Subtraction	28.9	28.6	28.8	0.8	30.3	0.6	6.5
Boolean	94.6	94.6	55.2	1.3	55.2	1.0	10.5
Comparison	90.4	90.4	38.7	1.0	38.7	0.9	10.7
Counting	79.4	79.4	66.1	0.6	65.9	0.4	1.4
Intersection	79.4	79.4	61.0	0.4	60.6	0.3	0.0
Multi-Entity	77.1	77.1	41.3	0.8	41.2	0.7	6.4
Multi-Hop	87.9	87.9	29.0	0.8	28.9	0.8	0.0
Multi-Relation	75.2	75.2	25.1	3.0	25.0	3.0	2.5
Spatial	21.2	21.2	0.0	13.0	0.0	13.0	0.0
Subtraction	34.4	34.4	1.3	1.0	0.9	0.7	0.0
Overall	69.3	69.3	31.6	3.1	31.5	3.0	3.6

Table 2: Ablation Study of Information. Q=Question, I=Image, F=Facts. Image refers to the Image feature stream. Results are expressed as % accuracy by question type.

8×10^{-5} and use AdamW (Loshchilov and Hutter, 2017) as optimizer. Once preprocessing is completed, we train the UNITER model with the cross-entropy objective function for 80,000 iterations, which we empirically found to guarantee convergence.

5 Results

Table 1 shows the results of our system (UNITER), using a question label break-down similar to Shah et al. (2019). Overall, we observe that our system outperforms the previous baseline MemNet setting (see 'World+WikiCap+ORG' in Shah et al. (2019)) with an absolute improvement of 19%.

Our results show that UNITER is learning to perform reasoning more accurately than MemNet in all but two cases. In the question types involving multiple entities ('Multi-Entity', 'Multi-Hop', 'Multi-Relation'), the increase is the greatest, suggesting that UNITER is able to robustly learn these reasoning here. We speculate that stacked self-attention layers in BERT are able to better attend to the many involved entities than MemNet.

We now discuss the performance of our model on its weakest categories, namely 'Subtraction' and 'Spatial'. The poor performance on 'Subtraction' questions confirms previous results that BERT-like models require specialised pretraining for numerical reasoning tasks (Geva et al., 2020). In the case of our model specifically, we note the lack of numerical reasoning tasks in UNITER's pretraining regime. 'Spatial' is the model's least accurate question type (21.4%) and the biggest absolute decrease from MemNet (-26.7%). This question type requires two-hop reasoning where the second hop is a numerical operation of the form $\underset{y}{\operatorname{argmin}}(x_i - y_i)$. Both of these have been shown to be problematic for BERT (Kassner et al., 2020; Geva et al., 2020).

6 Analysis

UNITER performs well at the reasoning tasks in general, with the most surprising result being that it apparently does better at multi-hop reasoning than one-hop. We believe that this can be explained by the presence of unbalanced distribution of answer types in the dataset perturbing the results (see Table 1). We discuss this in Section 6.1.

In order to better understand the reasoning capability of our model and the impact of each input modality, we perform an inference time ablation study, presented in Table 2.

Ablation of Image features (column 'Q+F') does not change the performance, suggesting that the model is not attending to image features. To confirm this hypothesis, we performed an experiment with adversarial images, obtaining very similar results for each question type and the same overall score (69.30%). We explain this behaviour by the fact that the preprocessing pipeline extracts all the required information as explicit facts which the model prefers over the more ambiguous visual features. We leave a deeper analysis for further work.

An interesting case is the 'Spatial' questions, where facts alone are able to correctly answer 13% of the questions. This is likely the result of the answers to this question type being entities present in the facts. Again, we observe that the model is not able to learn this information from the visual features.

	Train Ablation		Adversarial Modality*	
Question Type	Q+I	Q	I	F
1-Hop	47.09	38.5	65.9	31.3
1-Hop Counting	66.1	61.5	75.2	50.5
1-Hop Subtraction	29.4	29.7	28.1	26.2
Boolean	83.9	67.3	94.1	57.5
Comparison	83.4	60.3	90.6	47.8
Counting	75.4	75.2	78.9	70.2
Intersection	67.6	67.9	76.8	61.2
Multi-Entity	69.4	57.2	76.4	47.6
Multi-Hop	56.5	50.2	87.9	38.4
Multi-Relation	47.3	38.9	75.2	28.3
Spatial	3.3	1.2	21.1	0.0
Subtraction	2.1	2.6	39.2	1.6
Overall	47.0	40.8	69.3	32.8

Table 3: Further Ablation and Adversarial Studies. *Adversarial Modality indicates that the sample from that modality was randomly assigned from the entire data split

6.1 Bias Studies

We briefly discuss the corpus bias, a well-known concern in VQA (Goyal et al., 2019). We consider question difficulty across three parameters: reasoning difficulty, task design, and corpus bias. Certain question types are inherently more complex, as discussed in Section 5. Additionally, the task may have different numbers of answer classes per task, effectively weakening any priors models might form (see Entropy column in Table 1). Finally, an unbalanced dataset may cause certain reasoning types to be underrepresented, making it harder for models to learn for them. 'Spatial' and 'Substraction' questions are among the least represented in the training dataset, which increase their difficulty for the model.

Unseen answer classes are also an issue. For 'Spatial' questions, only 54.2% of the test answers (output classes) are actually seen during training, placing an upper bound on accuracy. We find 98.4% of 'Spatial' questions the model answered correctly and 95.7% of 'Spatial' question the model answered incorrectly were supplied with adequate facts by the preprocessing pipeline.

Training time ablation and adversarial experiments To further probe the task, we perform a training time ablation with first facts, and then facts and images removed (see Table 3). In this we seek to exhibit the capability of our model to leverage the available modalities and to compensate for the missing ones.

Through comparing the training time and inference time ablations, we can better understand the importance of a modality to solving the task.

Through comparing train and inference ablation of facts ('Q+I' column of Table 3 and of Table 2) we observe that when facts are unavailable at train time, the model attends to images to obtain 47.0% accuracy, which is 15.4% more than the 31.6% obtained by the corresponding inference time ablation. This indicates that the visual modality can provide useful information for this task.

We observe a similar trend in the fact and image ablation setting ('Q' column of Table 3 and of Table 2) that the model is able to greater leverage questions to make accurate predictions when additional modalities are never available.

We also perform adversarial checks, where random images or facts from the data split are presented at inference time. These align closely with the ablation study, with adversarial images (Column 'I' of Table 3) performing within 0.1% of blanked images (Column 'Q+F' of Table 3) and adversarial facts (Column 'F' of Table 3) performing within 1% of blanked facts (Column 'Q+I' of Table 3). These results confirm the importance of factual data and the unimportance of raw image features to a model trained on the full data.

7 Conclusion and Future Work

We evaluated our model and found that it improves on the previous state of the art by a substantial margin (19.1%). An ablation study revealed the specific strengths and weaknesses of our model on certain question categories when evaluated on the KVQA dataset. We show that the UNITER model is not actually using the visual input.

In the future, we seek to create a large external knowledge dataset designed following KVQA with more entities besides persons to encourage grounded reasoning, and better calibration of answer types. We will also consider pretraining our model on closely related tasks. This will help to form a model capable of learning robust reasoning with a high degree of spatial specificity and entity discrimination.

Acknowledgements

Peter Vickers is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1.

Ethical Statement

This work is based on the open-source KVQA dataset, an English multimodal dataset, and the Wikidata knowledge base (also in English). No English-specific preprocessing was used for this research and the UNITER model is language agnostic, which tends to suggest that this could generalize to other languages. We will make our code publicly available to ensure the reproducibility of our experiments in the following repository

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 2425–2433.

Lawrence W Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:2631–2639.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing Relational Knowledge from BERT. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7456–7463.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics (ACL).

Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, pages 677–685. Association for Computational Linguistics.

Yen Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12375 LNCS:104–120.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186.

David Ding, Felix Hill, Adam Santoro, and Matt Botvinick. 2020. Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting Numerical Reasoning Skills into Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*, 127(4):398–414.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-celeb-1M: A dataset and benchmark for large-scale face recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9907 LNCS, pages 87–102.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 6693–6702.

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. REL: An Entity Linker Standing on the Shoulders of Giants. *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2197–2200.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:1988–1997.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are Pretrained Language Models Symbolic

Reasoners over Knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564, Online. Association for Computational Linguistics.

Ranjay Krishna, Justin Johnson, Yannis Kalantidis Yahoo, David Ayman Shamma, Yuke Zhu, Oliver Groth, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations Human trajectory forecasting View project hybrid intrusion detction systems View project Visual Genome Connecting Language and Vision Using Crowdsourced Dense Image A. *Article in International Journal of Computer Vision*, 123(1):32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What Does BERT with Vision Look At? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, PART 5, pages 740–755. Springer Verlag.

Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.0.

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*, volume 32, pages 13–23. Curran Associates, Inc.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-Task Vision and Language Representation Learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, volume 29, pages 289–297. Curran Associates, Inc.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:3190–3199.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, volume 24, pages 1143–1151. Curran Associates, Inc.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. Language models as knowledge bases? In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2463–2473. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 4932–4942.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 815–823.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-Aware Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8876–8884.

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

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *arXiv*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2020. A corpus for reasoning about natural language grounded in photographs. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 6418–6428. Association for Computational Linguistics (ACL).

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 3878–3887, Online. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Explicit knowledge-based reasoning for visual question answering. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 0, pages 1290–1296. International Joint Conferences on Artificial Intelligence.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2016. FVQA: Fact-based Visual Question Answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory Networks. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 6713–6724.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.