VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge - Supplementary Material

Sahithya Ravi[1,2*] Aditya Chinchure[1,2*] Leonid Sigal[1,2] Renjie Liao[1] Vered Shwartz[1,2]
[1] University of British Columbia [2] Vector Institute for AI
{sahiravi, aditya10, lsigal, vshwartz}@cs.ubc.ca, rjliao@ece.ubc.ca

1. Implementation Details

In this section, we provide additional information about the implementation of each of the components of VLC-BERT.

1.1. Object Tags with YOLO

As described in Sec 3 of our paper, we utilize object tags to incorporate image context for generating commonsense inferences. In order to obtain the object tags, we use an off-the-shelf YOLO model for PyTorch, YOLOv5 *by Ultralytics* [2]. We use the pretrained *yolov5l* model to obtain object bounding boxes and the associated class name for each bounding box, on COCO 2014 and 2017 images for OK-VQA and A-OKVQA datasets respectively. We then use a confidence threshold of 0.5 to prune out objects that are unlikely to be useful. In addition, we prune out the *person* object name as well as the objects already present in the question phrase, to avoid unnecessary tags or repetitions. Finally, the two object names associated with the highest confidence bounding boxes are picked as the object tags for our model, O.

1.2. Knowledge Generation

As described in Sec 3.1.1 of our paper, we generate commonsense inferences from COMET [1] by inputting the question followed by "with" and two object tags into it. If S is the sentence consisting of the question and object tags and R is the relation type we want to generate from COMET, we provide it to COMET in the form $S^{(i)}$ $R^{(i)}$ `[GEN]` and let comet generate the commonsense inferences. Though COMET can support 50 relation types, we cherry-pick 30 relation types by removing duplicate relations and relations that are irrelevant to our work (*e.g*. HasPainIntensity). Table 3 (at the end of this document) provides the list of the 30 relations we used to generate commonsense expansions from COMET and the corresponding templates we used to convert COMET's output to natural language sentences. $\{0\}$ usually indicates the subject in the input sentence to COMET, and $\{1\}$ indicates the generated expansion.

1.3. Knowledge Selection

As described in Sec 3.1.2 of our paper, we augment S-BERT to perform semantic search and filter and rank the relevance of commonsense inferences. In order to perform semantic search, we utilize the sentence-transformers package for SBERT[1] [4]. We initialize our SBERT model from the pre-trained *msmarco-roberta-base-ance-firstp* model and train this model for 2 epochs on the training set of the corresponding task. To create the labels for this augmentation, we measure the overlap of the expansions to human annotated answers and assign a similarity score of *0.8* for overlapping expansions and a score of *0.2* for non-overlapping expansions. This augmented S-BERT model is then used to encode the question and commonsense sentences, before computing the sentence similarity between every commonsense sentence and the question, and picking the top k ($K = 5$) sentences to use in the input sequence of the VLC-BERT transformer.

1.4. VLC-BERT Transformer

Our implementation of the VLC-BERT transformer encoder is based on the publicly available implementation of VL-BERT[2] [5]. The hyperparameters we use for training VLC-BERT on the VQA 2.0 (only for pre-training), OK-VQA and A-OKVQA datasets are given in Table 1.

For generating sentence embeddings for commonsense inferences that are fed into the MHA block, we use the *all-mpnet-base-v2* pre-trained model from SBERT.

1.5. Implementation of Commonsense Subsets

In Sec 7.1, we describe the need for commonsense-specific subsets of OK-VQA and A-OKVQA, to show that our model improves on the baseline significantly. The lack

[*] Denotes equal contribution

[1] https://www.sbert.net

[2] https://github.com/jackroos/VL-BERT

Table 1: Hyperparameters of our model

Hyperparameter	**VQA P.T.**	**OK-VQA**	**A-OKVQA**
Batch Size	16	16	16
Gradient Accumulation	4	4	4
Epochs	5	20	20
Learning Rate	6.25e-7	6.25e-7	6.25e-7
Visual Size	768	768	768
Hidden Size	768	768	768
Warmup Method	linear	linear	linear
Warmup Steps	1000	1000	1000
MHA Heads	–	3	3
MHA Dropout	–	0.1	0.1

of any annotations for the type of reasoning required to answer the question led us to develop our own method to obtain the subsets. Below, we have the exact details required to re-create the subsets:

Named Entities. We use spaCy's entity recognizer[3]. If any word in the question or list of answers is recognized as an entity, we prune the question.

Numerical. We first attempt to check if a string is a number using Python's built-in function, `isdigit()`. If it is not a digit, then we use the `word2number` package[4] to attempt to convert words (*e.g.* ”twenty”) into numbers. If it is successful in doing so, we deem the word to be a number. If any word in the question or list of answers is recognized as a number, we prune the question.

Directional. We list commonly used directional words: `right, left, top, bottom, behind, under, inside, over, front, back, near, next`. If any word in the question is recognized as a directional word, we prune the question.

Symbol. We list commonly used symbol words: `logo, symbol, name, company, mascot, word, brand`. If any word in the question is recognized as a symbol word, we prune the question.

Color. We list commonly used color words: `blue, green, red, black, white, grey, purple, pink, yellow, orange`. If any word in the question or list of answers is recognized as a color word, we prune the question.

Time. Finally, we list commonly used time words: `century, year, time, month, day`. If any word in the question is recognized as a time word, we prune the question.

Table 2: Performance of our model on OK-VQA question categories.

Category	Base	w/ COMET
Vehicles and Transportation	40.1	**41.16**
Plants and Animals	**42.58**	41.65
People and Everyday Life	**40.09**	39.95
Sports and Recreation	51.53	**52.31**
Cooking and Food	42.36	**45.04**
Objects, Material and Clothing	39.86	**39.95**
Science and Technology	37.38	**38.57**
Weather and Climate	**50.7**	48.99
Brands, Companies and Products	33.6	**35.81**
Geog, Hist, Language and Culture	40.14	**43.4**
Other	41.23	**42.68**

As the task of recognizing the type of a question is challenging in itself, we tried to simplify it to a basic, reproducible method, in order to better evaluate on commonsense reasoning specific questions on the OK-VQA test set and the A-OKVQA validation set.

2. Evaluation

2.1. Main Evaluation

The standard deviation on our scores for the OK-VQA test set is 0.20 and for the A-OKVQA validation is 0.47.

2.2. Evaluation on OK-VQA Question Categories

The results provided in our paper only show the overall scores of our models on the OK-VQA [3] dataset. In Table 2, we share the results for each question category in the OK-VQA dataset. The OK-VQA dataset has questions divided into 11 different categories [3]. The results show that our model with external knowledge from COMET improves upon the baseline in all but three categories. Across all the models, we see that the 'Brands, Companies and Products' is the most challenging category, with low accuracy for both the baseline and the VLC-BERT with COMET models. This is expected, because the questions in this category often require the model to read text or symbols in the image, or identify company names and logos, which are challenging tasks outside the domain of our model.

2.3. Ablations

In this section, we present additional ablations to show the impact of different components of the VLC-BERT pipeline.

[3] https://spacy.io/api/entityrecognizer

[4] https://pypi.org/project/word2number/

Ablation on number of sentences. In order to test the impact of the number of commonsense inferences K, we report the performance with different K values. We ran our latest model with $K = 10$ and $K = 15$ sentences. On the A-OKVQA validation set, we obtain the following results: $\boxed{K = 5 : 44.95; K = 10 : 44.57; K = 15 : 43.93}$. We thus feed $K = 5$ commonsense inference sentences into VLC-BERT transformer, because we had observed that adding too many commonsense inferences also adds unnecessary noise in the model, which hurts performance.

Use of Object tags. In order to assess the importance of the number of object tags used in deriving commonsense inferences, we ran experiments with no (0) object tags, as well as all (>2) tags. For zero tags, we get $\boxed{44.42}$, and for all tags, we get $\boxed{44.62}$. These are slightly worse than the two tags version ($\boxed{44.95}$). This is in line with what we expected, since COMET is not designed to deal with complex sentences containing multiple entities, and 2 object tags stands as a good trade-off. Furthermore, in our qualitative results, we show examples of where object tags are useful in providing image context.

Impact of weak attn. supervision. Disabling weak attn. supervision, we obtain a result of $\boxed{44.89}$ which is slightly worse compared to $\boxed{44.95}$ with supervision. However, qualitative analysis shows that our model with supervision produces stronger attention weights for useful inferences compared to the model without.

3. Analysis

In this section, as mentioned in Section 7.2 of our paper we provide additional qualitative examples along with their attention scores, of where VLC-BERT improved as well as failed.

3.1. Error examples

We analyze the errors from the best version of VLC-BERT model. We randomly sample 50 erroneous examples from the validation set of A-OKVQA, analyze the errors, and classify them into five categories as shown in Figure 1. We provide an example of each category in Figure 2.

① **Visual**: The model is lacking deep scene understanding that either required to answer the question, or to generate relevant commonsense inferences. This includes cases where the object tags are insufficient for describing the scene. A majority of the errors we see in VLC-BERT fall in this category, in line with our conclusions and motivations for future work on commonsense models that involve deep scene understanding.

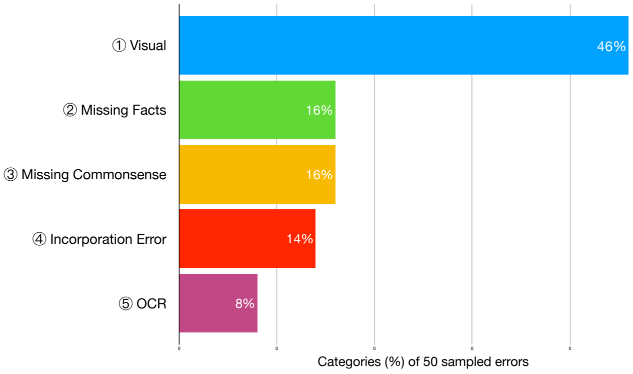


Figure 1: **Error analysis:** Percentage of error categories from AOKVQA

② **Missing Facts**: The model failed due to missing factual knowledge about named entities, types of entities and well-known facts.

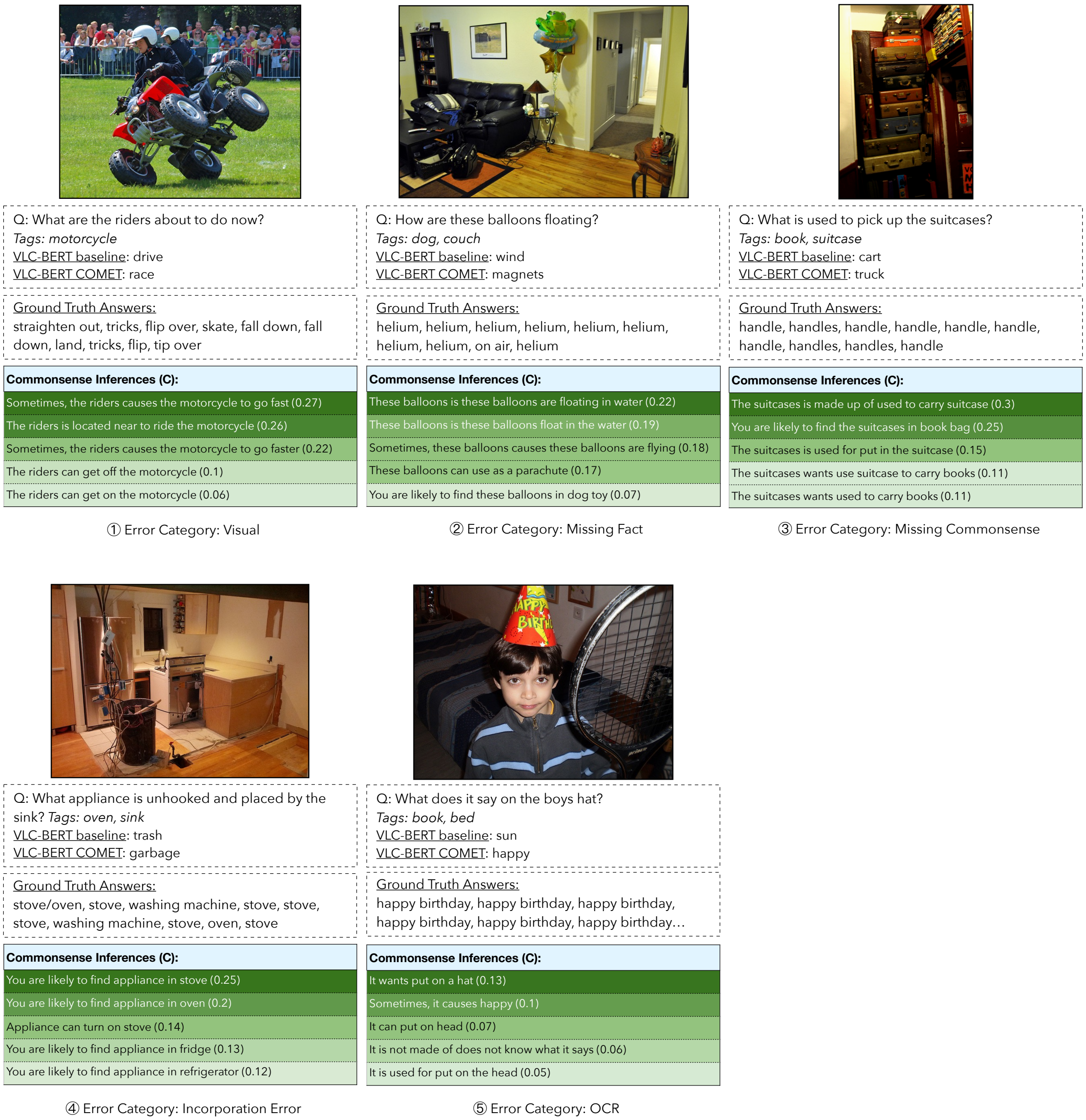
③ **Missing Commonsense**: The final commonsense inferences provided to VLC-BERT are missing the commonsense knowledge required to answer the question, either due to COMET not capturing this knowledge or semantic search not picking the right inferences.

④ **Incorporation Error**: Though the answer is provided in the commonsense inferences, and we attended highly to these inferences, it is still ignored by VLC-BERT, probably because the visual representation took priority. The commonsense inferences being much more condensed compared to other inputs of VLC-BERT could be one of the reasons for this.

⑤ **OCR**: The question involves reading text in the images and requires the VLC-BERT to support Optical Character Recognition (OCR).

3.2. Improvement examples

In Figure 3, we provide additional qualitative examples where commonsense from COMET helped in driving the model to make the right prediction.



Q: What are the riders about to do now?
Tags: motorcycle
VLC-BERT baseline: drive
VLC-BERT COMET: race

Ground Truth Answers:
straighten out, tricks, flip over, skate, fall down, fall down, land, tricks, flip, tip over

Commonsense Inferences (C):
Sometimes, the riders causes the motorcycle to go fast (0.27)
The riders is located near to ride the motorcycle (0.26)
Sometimes, the riders causes the motorcycle to go faster (0.22)
The riders can get off the motorcycle (0.1)
The riders can get on the motorcycle (0.06)

① Error Category: Visual

Q: How are these balloons floating?
Tags: dog, couch
VLC-BERT baseline: wind
VLC-BERT COMET: magnets

Ground Truth Answers:
helium, helium, helium, helium, helium, helium, helium, helium, on air, helium

Commonsense Inferences (C):
These balloons is these balloons are floating in water (0.22)
These balloons is these balloons float in the water (0.19)
Sometimes, these balloons causes these balloons are flying (0.18)
These balloons can use as a parachute (0.17)
You are likely to find these balloons in dog toy (0.07)

② Error Category: Missing Fact

Q: What is used to pick up the suitcases?
Tags: book, suitcase
VLC-BERT baseline: cart
VLC-BERT COMET: truck

Ground Truth Answers:
handle, handles, handle, handle, handle, handle, handle, handles, handles, handle

Commonsense Inferences (C):
The suitcases is made up of used to carry suitcase (0.3)
You are likely to find the suitcases in book bag (0.25)
The suitcases is used for put in the suitcase (0.15)
The suitcases wants use suitcase to carry books (0.11)
The suitcases wants used to carry books (0.11)

③ Error Category: Missing Commonsense

Q: What appliance is unhooked and placed by the sink? *Tags: oven, sink*
VLC-BERT baseline: trash
VLC-BERT COMET: garbage

Ground Truth Answers:
stove/oven, stove, washing machine, stove, stove, stove, washing machine, stove, oven, stove

Commonsense Inferences (C):
You are likely to find appliance in stove (0.25)
You are likely to find appliance in oven (0.2)
Appliance can turn on stove (0.14)
You are likely to find appliance in fridge (0.13)
You are likely to find appliance in refrigerator (0.12)

④ Error Category: Incorporation Error

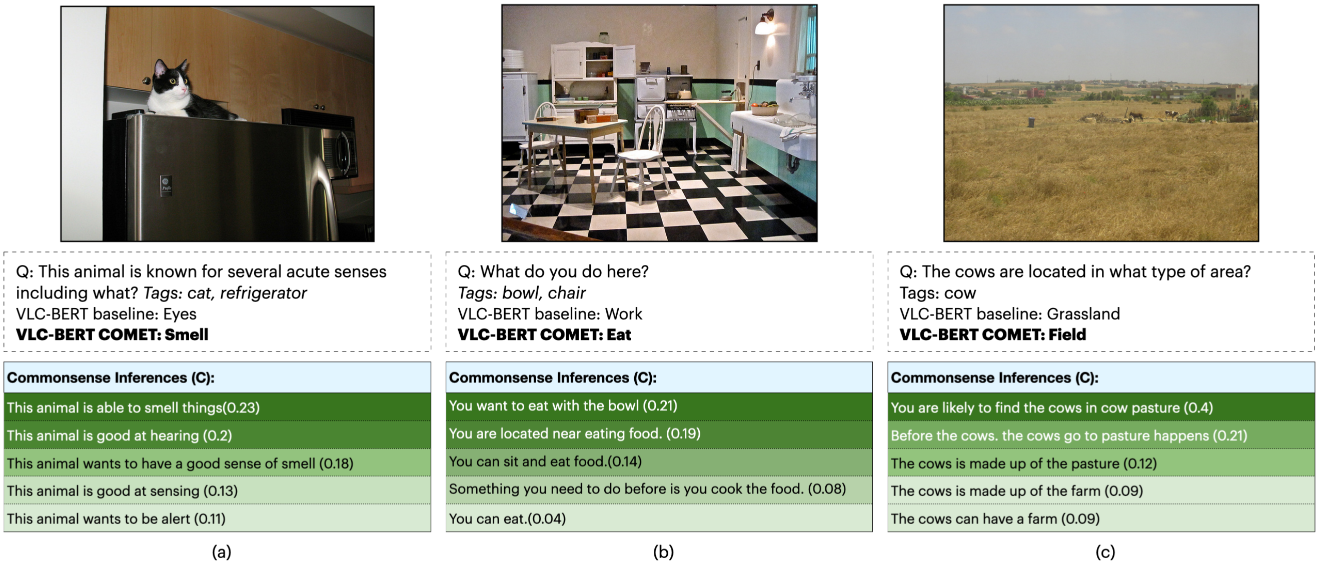
Q: What does it say on the boys hat?
Tags: book, bed
VLC-BERT baseline: sun
VLC-BERT COMET: happy

Ground Truth Answers:
happy birthday, happy birthday, happy birthday, happy birthday, happy birthday, happy birthday…

Commonsense Inferences (C):
It wants put on a hat (0.13)
Sometimes, it causes happy (0.1)
It can put on head (0.07)
It is not made of does not know what it says (0.06)
It is used for put on the head (0.05)

⑤ Error Category: OCR

Figure 2: **Error analysis:** We sample 50 erroneous examples from the A-OKVQA validation set, and categorize it into five categories.



Q: This animal is known for several acute senses including what? *Tags: cat, refrigerator*
VLC-BERT baseline: Eyes
VLC-BERT COMET: Smell

Commonsense Inferences (C):
This animal is able to smell things(0.23)
This animal is good at hearing (0.2)
This animal wants to have a good sense of smell (0.18)
This animal is good at sensing (0.13)
This animal wants to be alert (0.11)

(a)

Q: What do you do here?
Tags: bowl, chair
VLC-BERT baseline: Work
VLC-BERT COMET: Eat

Commonsense Inferences (C):
You want to eat with the bowl (0.21)
You are located near eating food. (0.19)
You can sit and eat food.(0.14)
Something you need to do before is you cook the food. (0.08)
You can eat.(0.04)

(b)

Q: The cows are located in what type of area?
Tags: cow
VLC-BERT baseline: Grassland
VLC-BERT COMET: Field

Commonsense Inferences (C):
You are likely to find the cows in cow pasture (0.4)
Before the cows. the cows go to pasture happens (0.21)
The cows is made up of the pasture (0.12)
The cows is made up of the farm (0.09)
The cows can have a farm (0.09)

(c)

Figure 3: **Qualitative examples:** (a) is from A-OKVQA, and (b) and (c) are from OK-VQA.

Table 3: Relations used for generating expansions from COMET and their corresponding sentence templates

#	Relation	Sentence template
1	AtLocation	You are likely to find {0} in {1}
2	CapableOf	{0} can {1}
3	Causes	Sometimes {0} causes {1}
4	CreatedBy	{1} is created by {0}
5	Desires	{0} wants {1}
6	HasA	{0} has {1}
7	HasFirstSubevent	The first thing you do when you {0} is {1}
8	HasProperty	{0} is {1}
9	HinderedBy	{0} is hindered by {1}
10	IsA	{0} is {1}
11	isAfter	{0} happens before {1}
12	isBefore	{1} happens before {0}
13	LocatedNear	{0} is located near {1}
14	MadeOf	{0} is made of {1}
15	MadeUpOf	{0} is made up of {1}
16	NotCapableOf	{0} is not capable of {1}
17	NotHasProperty	{0} does not have the property of {1}
18	NotIsA	{0} is not {1}
19	NotMadeOf	{0} is not made of {1}
20	ObjectUse	{0} is used for {1}
21	PartOf	{1} has {0}
22	SymbolOf	{0} is a symbol of {1}
23	UsedFor	{0} is used for {1}
24	xAttr	{0} is seen as {1}
25	xEffect	{0} then {1}
26	xIntent	Because {0} wanted {1}
27	xNeed	Before {0} needed {1}
28	xReact	As a result {0} feels {1}
29	xReason	{0} reasons {1}
30	xWant	As a result {0} wants {1}

References

[1] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021.

[2] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, Feb. 2022.

[3] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.

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