Is BERT Blind? Exploring the Effect of Vision-and-Language Pretraining on Visual Language Understanding —Supplementary Material—

Morris Alper*, Michael Fiman*, Hadar Averbuch-Elor
Tel Aviv University

Task	Metric	VisualBERT	LMXERT	BERT	CLIP
VLU					
Conc.	Pearson	0.400	0.421	0.233	0.513
	Spearman	0.412	0.370	0.238	0.495
	Kendall	0.281	0.249	0.159	0.339
NCD	Accuracy	0.467	0.400	0.267	0.823
CTD	Accuracy	0.314	0.431	0.353	0.800
NLU					
Cites	R@1	0.003	0.007	0.199	0.019
NLI	AUC	0.704	0.688	0.754	0.696

Table 1. Evaluating non-dual V&L encoders (VisualBERT and LMXERT) on several VLU and NLU tasks with BERT and CLIP added for reference.

Contents

1. Additional Results and Comparisons

1.1. Non-dual V&L encoder models

Although non-dual (fusion) encoder models are not directly comparable to purely textual encoders such as BERT or the text encoder component of CLIP which do not fuse modalities, we consider them here for completeness We evaluate the VisualBERT and LMXERT non-dual encoder models on several tasks from our task suite by only feeding them textual input. Results are shown in Table 1. As illustrated in the table, even though these models were trained using image features together with text tokens, the models outperform BERT on visual tasks, though the gap is not as significant as with dual encoder models.

1.2. Additional tasks using linear probing

We present two additional tasks for comparing V&L and unimodal models using linear probing, one VLU and one NLU task. For both tasks, we use a linear classifier on the pooled embedding output of a model for categorical prediction. Specifically, we use a logistic regression model using the scikit-learn `linear_model.LogisticRegression` implementation. For all tasks we use the default parameters except for `max_iter` which was changed according to task requirements to allow convergence. In particular, we use parameters `penalty='l2', C=1.0, solver='lbfgs'`.

1.2.1 Groundability classification

Task description. In paired text-image data, there is normally an implied mapping between referential expressions in the text and objects or regions in the accompanying image. The task of learning these mappings is known as *visual grounding* and is of general interest for visual semantic understanding [1, 9, 22]. In captions accompanying images, some expressions refer directly to regions in images while others give non-visual context; we refer to the former as *groundable* referents and the latter as *non-groundable*. A similar paradigm was recently proposed by Kim et al. [6] that separately considers “answerable” and “unanswerable” phrases.

We propose a *groundability classification* task, consisting of classifying referents in text as groundable or non-groundable. This is a text-only task as it uses text alone and the visual context is only implied. Since this task requires

[*] These authors contributed equally to this work

visual imagination to complete, we consider it to be a VLU task.

Experimental details. In line with previous works that consider person-centric visual grounding [1, 11], we construct a dataset of *person-centric groundability* sentences where a fixed human participant is either implied to be groundable (i.e., on-camera) or non-groundable (i.e., off-camera). The associated task consists of binary classification applied to these texts according to whether the given participant would be visible in a description of an event. Due to the lack of existing labelled data for this task, we created a synthetic dataset of sentences with a common format: *Alex* `[MASK]`*ing Riley's* `[MASK]`, where the first masked word is a randomly drawn verb, and the second masked word is a randomly drawn noun, and the task is to classify whether or not the second mentioned individual (i.e., Riley) is groundable. Groundability labels are estimated using zero-shot text classification with a pretrained natural language inference model. We created synthetic data for the groundability task by taking the prompt template "Alex __ing Riley's __", filling in various verb-noun pairs into the given slots, and filtering using a pretrained language model to select for natural-sounding samples. We then estimated ground-truth labels using zero-shot inference with a pretrained natural language inference (NLI) model.

To find verb-noun pairs, we listed all verbs and nouns in the Brown corpus of standard American English with part-of-speech labels [3]. We converted all text to lowercase and then selected the 5,000 most common verb lemmas and 1,000 most common noun lemmas in this corpus. Using all possible verb-noun combinations among these, inserted into the prompt template shown above, yielded 5M candidate phrases. From the given 5M candidates we sample randomly 200K phrases. We then calculate the total negative log-likelihood (NLL) for each candidate relative to the pretrained language model GPT2-large [13] and kept only those samples in the 20th percentile of NLL (i.e. the most likely samples), corresponding to 40,000 descriptions.

After generating these texts, we estimated labels using a pretrained NLI model. We used BART-large [7] fine-tuned on the MNLI dataset [23] (using the `facebook/bart-large-mnli` checkpoint from Hugging Face model hub[1]). This model takes pairs of texts as inputs (the "premise" and "hypothesis" texts) and outputs three probabilities per pair: p_c, p_n, p_e, corresponding to probabilities of a contradictory, neutral, or entailment relation between the texts respectively. As observed by Yin et al. [24], NLI can be used for zero-shot text classification by designing premise and hypothesis prompts for the task of interest. In our case, we use the following prompts:

Premise: "This is a picture of _______."

Hypothesis: "Riley can be seen in the picture."

For each of our 40,000 texts, we insert the text in the slot given in the premise and calculate p_e with the NLI model. If $p_e > 0.5$ we assign the sample label 1 (groundable), otherwise we assign it label 0 (non-groundable). Below are several example sentences from the synthetic dataset. Examples of Riley being groundable (sample label 1):

- Alex facing Riley's figure
- Alex viewing Riley's participation
- Alex seeing Riley's enjoyment

Examples of Riley being non-groundable (sample label 0):

- Alex hiding Riley's file
- Alex announcing Riley's absence
- Alex stealing Riley's evidence

For evaluation we created a test set, containing 200 sentences judged by human evaluators to be natural sounding, half labeled as groundable and the other half as non-groundable. To provide an example, sentences such as *Alex cutting Riley's hair* or *Alex blocking Riley's shot* were labeled as groundable, whereas sentences such as *Alex painting Riley's house* or *Alex counting Riley's vote* were labeled as non-groundable.

For this binary classification task, we apply linear probing to assess our models' understanding of groundability, and report ROC-AUC scores for each model. We also provide 95% confidence intervals, calculated using bootstrap resampling with 200 bootstraps, in order to analyze the robustness of these results.

Model	AUC (95% CI)
BERT	0.789 ± 0.0007
RoBERTa	0.799 ± 0.0005
ERNIE	0.766 ± 0.0006
CLIP	**0.822** ± **0.0007**

Table 2. **Groundability Classification Evaluation**. We report ROC-AUC with 95% bootstrap confidence intervals scores for a manually assembled test set comparing linear probing for text based encoders and V&L CLIP model. As shown above, CLIP significantly outperforms the unimodally trained models.

Results and discussion. Results for the groundability classification are provided in Table 2. As these results illustrate, CLIP significantly outperforms all unimodally trained text encoders on average. We observe that the score gaps are not as distinct as in the previous zero-shot tasks, as this task is a learnable task which requires training, allowing all models

[1] The model can be found here.

to learn this task to some extent. Nonetheless, CLIP's ability to surpass the unimodally trained encoders suggest that V&L trained text encoders have a better ability to grasp if an object is grounded or not due to additional perceptual information that is encoded during the pretraining phase. Furthermore, note that in comparison to the other VLU tasks, here the subject in question (i.e., Riley) is not directly connected to visual information and the prediction is based only on context relating to the performed action and the associated object. The improved performance on this task illustrates that V&L models can better encode higher-level perceptual reasoning.

1.2.2 Natural language inference

Task description. Natural language inference (NLI) refers to inferring the logical relation between pairs of statements, as well as more generally referring to logical inference based on text [19]. In particular, NLI commonly considers the following logical relations between sentences A and B:

- **Contradiction**: For example, A=*It is rainy outside.* is contradicted by B=*It is sunny outside.*, since they cannot be simultaneously true.
- **Neutral**: For example, A=*It is rainy outside.* is neutral with regards to B=*It is summer.*, since A neither contradicts nor entails B.
- **Entailment**: For example, A=*It is cold and rainy outside.* entails B=*It is cold outside.*, since if A is true then B must also be true.

Solving this task requires an understanding of the fine-grained semantics of language and logical reasoning. On the other hand, visual cues are not tightly related to this task and are even potentially misleading. For example, the sentences *This cup contains grape juice.* and *This cup contains wine.* are contradictory even though the scenes they describe are visually identical. Therefore, we consider this to be a non-visual NLU task.

Experimental details. For this task we use the MNLI dataset introduced by Adina et al. [23]. We remove sentence pairs with a neutral relation and treat this as a binary classification task to predict sentence pairs as contradictory or entailing. We perform 5-fold cross validation on a dataset of 261,775 pairs of sentences using 80% of samples for training and 20% for testing.

For each sentence pair, we concatenate the sentences' two pooled embeddings and apply linear probing. Note that some models such as BERT include a special `[SEP]` token for encoding sentence pairs as a single unit, but we encode sentences separately and concatenate their embeddings in order to have a fair comparison between all models. We report the ROC-AUC score on the MNLI test set.

Model	AUC ± std
BERT	0.754 ± 0.001
RoBERTa	0.777 ± 0.001
ERNIE	**0.787** ± 0.001
CLIP	0.696 ± 0.001

Table 3. **NLI Evaluation**. We report ROC-AUC scores for the NLI task using linear probing, comparing text based encoders to the V&L CLIP model. As depicted above, the V&L trained text encoder is inferior to all other text based encoders for this non-visual language understanding task.

Results and discussion. Results for NLI are provided in Table 3. As shown in the table, text-based models outperform CLIP by a large margin. Similar to our findings regarding linguistic acceptability classification, we see that V&L trained models are less effective in tasks that do not incorporate perceptive information, suggesting that for non-visual tasks, V&L pretraining is not necessarily beneficial.

1.3. Comparing usage of SP on text based models

In the main paper we presented results for text models using MLM probing, and for CLIP using Stroop probing (SP). To allow for a full comparison between both types of models, and to strengthen the choice of using MLM probing for text based models, we present additional results comparing SP and MLM probing for text based models. Table 4 presents results for comparing SP and MLM probing methods for BERT and RoBERTa. As illustrated, using SP with unimodally trained models results in lower performance than using MLM probing with these models. This result supports our choice of using MLM probing for text based models trained to perform MLM tasks as the preferred probing method.

1.4. Additional task results information

We provide additional detailed results for our suite tasks including the mean and standard deviation of the results over all used prompts in Table 5.

1.5. Qualitative analysis for V&L model misclassifications on color prediction

Our results for color association prediction show that V&L models outperform unimodally trained text encoders in the given setting. Additional qualitative analysis of the results show that even the reported misclassifications of V&L models such as CLIP may be explained by ambiguities in the dataset itself. For example, the noun "ash" has ground truth value "grey" in our dataset, while CLIP with

	Color		Shape	Knowledge		Proficiency			Sent.
Metric	acc_{CTD}	acc_{NCD}	acc	R@1	R@5	acc_{V}	acc_{N}	acc_{P}	acc
BERT-MLM	**0.353**	**0.400**	**0.559**	**0.198**	**0.522**	**0.898**	**0.753**	**0.893**	**0.618**
BERT-SP	0.137	0.067	0.412	0.000	0.005	0.048	0.038	0.013	0.596
RoBERTa-MLM	**0.431**	**0.333**	**0.431**	–	–	**0.877**	**0.718**	**0.881**	**0.666**
RoBERTa-SP	0.176	0.200	0.422	–	–	0.016	0.019	0.063	0.616

Table 4. **Comparing SP to MLM probing for text base models**. As the results show, using probing using MLM method for text based models outputs better results than using SP

	Concreteness			Color		Shape	Sent.
Metric	$\lvert\rho\rvert$	$\lvert r_s\rvert$	$\lvert\tau\rvert$	acc_{CTD}	acc_{NCD}	acc	acc
Unimodal							
BERT-base	0.27 ± 0.10	0.27 ± 0.09	0.18 ± 0.07	0.26 ± 0.13	0.25 ± 0.08	0.47 ± 0.08	0.56 ± 0.03
BERT-large	0.18 ± 0.13	0.26 ± 0.10	0.17 ± 0.06	0.28 ± 0.14	0.27 ± 0.15	0.51 ± 0.06	0.56 ± 0.03
DistilBERT	–	–	–	0.23 ± 0.08	0.31 ± 0.04	0.45 ± 0.09	0.56 ± 0.04
RoBERTa-base	0.30 ± 0.09	0.29 ± 0.10	0.19 ± 0.07	0.27 ± 0.10	0.27 ± 0.07	0.43 ± 0.00	0.61 ± 0.04
RoBERTa-large	0.21 ± 0.10	0.23 ± 0.11	0.16 ± 0.07	0.30 ± 0.12	0.26 ± 0.08	0.43 ± 0.00	$\mathbf{0.63} \pm \mathbf{0.06}$
DistilRoBERTa	–	–	–	0.24 ± 0.12	0.25 ± 0.10	0.43 ± 0.01	0.57 ± 0.02
ERNIE	0.23 ± 0.10	0.20 ± 0.12	0.13 ± 0.08	0.10 ± 0.04	0.13 ± 0.11	0.31 ± 0.08	0.53 ± 0.02
ERNIE-large	0.23 ± 0.08	0.22 ± 0.07	0.14 ± 0.05	0.12 ± 0.06	0.07 ± 0.09	0.30 ± 0.05	0.57 ± 0.05
SBERT	0.24 ± 0.09	0.25 ± 0.09	0.17 ± 0.06	0.13 ± 0.02	0.07 ± 0.01	0.43 ± 0.05	0.53 ± 0.02
V&L							
CLIP	$\mathbf{0.47} \pm \mathbf{0.09}$	0.49 ± 0.09	0.34 ± 0.07	0.67 ± 0.15	$\mathbf{0.70} \pm \mathbf{0.08}$	0.69 ± 0.08	0.52 ± 0.01
OpenCLIP	0.45 ± 0.12	0.47 ± 0.12	0.32 ± 0.09	$\mathbf{0.77} \pm \mathbf{0.12}$	0.66 ± 0.17	$\mathbf{0.79} \pm \mathbf{0.08}$	0.53 ± 0.01
FLAVA	0.46 ± 0.10	$\mathbf{0.52} \pm \mathbf{0.10}$	$\mathbf{0.36} \pm \mathbf{0.07}$	0.52 ± 0.30	0.47 ± 0.22	0.68 ± 0.10	0.50 ± 0.01

Table 5. **Mean and STD Results.** Additional details of mean and standard deviations calculated across prompts, for all tasks which use multiple prompts.

Word	Ground Truth	Predicted Color
apple	green	red
ash	grey	black
cauliflower	white	brown
cello	brown	black
chalk	white	grey
foam	white	grey
garlic	white	brown
lady finger	green	red
pear	green	yellow
sea	blue	grey
sky	blue	white

Table 6. **Qualitative results for CLIP misclassified objects from the CTD and NCD datasets.** As can be seen by analyzing the misclassified objects, most mistakes can be explained by ambiguity of the data.

SP predicts the color "black", which is arguably also correct. Table 6 presents all of the objects from both color datasets misclassified by CLIP, containing the ground truth and the predicted color. As seen there, most of these predictions may be interpreted as valid colors for the given objects.

1.6. Analysis of reporting bias in LAION

Prior works have noted that commonsense properties that can be inferred from text are less likely to be explicitly stated than incongruent properties, notably including color terms(e.g. *a (yellow) banana* vs. *a blue banana*) [4, 10, 17]. In particular, text in image captioning datasets such as the web-scale LAION dataset [16] (used to train OpenCLIP) might have a different incidence of reporting bias than the text used to train models such as BERT. To disentangle this from the effect of training on the visual modality, we provide an analysis of reporting bias in LAION for color associations.

We use the `laion-2B-en` subset of 2.33 billion English-language image-caption pairs in the LAION-5B dataset, and estimate reporting bias by searching for bigram pairs (c, w) where c is a basic color term[2] and w is a un-

[2] one of {red, orange, yellow, green, blue, black, white, grey, brown}

Word	Ground Truth	LAION
banana	yellow	green
cherry	red	black
orange	orange	red
soil	brown	red
swan	white	black
wood	brown	white

Table 7. **Reporting bias in the LAION dataset**, illustrated by unigram nouns from the CTD and NCD datasets, along with their ground truth colors and the most commonly preceding colors in LAION.

igram noun from our color association datasets (CTD and NCD). The empirical probability of color c immediately preceding w is $P(c|w) = n_{(c,w)}/n_w$, where n indicates the number of instances of the given ngram, and the associated color estimates are $\hat{c}_w = \arg\max_c P(c|w)$. For these estimates, the corresponding accuracy scores on the unigrams in our datasets are $acc_{\texttt{CTD}} = 0.549$ and $acc_{\texttt{NCD}} = 0.714$, significantly below the accuracies achieved by all of the multimodally trained models under consideration on these datasets for the color prediction task. We also provide qualitative examples in Table 7 showing the effect of reporting bias for various common nouns from these datasets. These results provide evidence that multimodally trained models' strong performance on VLU tasks cannot be explained away as stemming from a lack of reporting bias in the texts used to train them.

2. Additional Details

2.1. Models

Table 8 presents the different models and Hugging Face checkpoints used for comparing results on the presented tasks.

2.2. Prompts used per task

We present further implementation details elaborating the list of prompts used per task.

Concreteness Prediction As explained in the main paper, we use the following prompts to probe our models for the concreteness of words in context by using a cloze task paradigm with Stroop probing. For each word tested, we insert the masked prompt and the prompt with the tested word and calculate the cosine similarity between them.

- *Alice giving the* [$\star$] *to Bob*
- *Bob giving the* [$\star$] *to Alice*
- *I see the* [$\star$]
- *A photo of my* [$\star$]
- *A close-up photo of a* [$\star$]
- *A painting of the* [$\star$]
- *A photo of the* [$\star$]
- *A photo of a nice* [$\star$]
- *A drawing of the* [$\star$]

Color Association Prediction For the color association prediction, we use the following prompts. For each given object denoted as $\langle w \rangle$, we use all color options to probe for the correct color.

- *A picture of a* [$\star$] $\langle w \rangle$
- *A photo of a* [$\star$] $\langle w \rangle$
- *A photo of the* [$\star$] $\langle w \rangle$
- *A* [$\star$] $\langle w \rangle$
- [$\star$] $\langle w \rangle$
- *The normal color of a* $\langle w \rangle$ *is* [$\star$]
- $\langle w \rangle$ *usually has a* [$\star$] *color*
- $\langle w \rangle$ *s have a* [$\star$] *color*
- *What is the color of a* $\langle w \rangle$*?* [$\star$]
- *The natural color of a* $\langle w \rangle$ *is* [$\star$]

Shape Association Prediction For the shape association prediction, we use the following prompts. For each given object denoted as $\langle w \rangle$, we use the given shape o to probe for the correct object shape.

- *A photo of a* [$\star$] *shaped* $\langle w \rangle$
- *A photo of a* [$\star$] $\langle w \rangle$
- *A photo of the* [$\star$] $\langle w \rangle$
- *A* [$\star$] $\langle w \rangle$
- [$\star$] $\langle w \rangle$
- *An image of a* [$\star$] $\langle w \rangle$
- *A* $\langle w \rangle$ *usually has a* [$\star$] *shape*
- $\langle w \rangle$ *s commonly have a* [$\star$] *shape*
- *The basic shape of a* $\langle w \rangle$ *is* [$\star$]
- *What is the shape of a* $\langle w \rangle$*?* [$\star$]

Model family	Size	Pretraining	Params	MLM head?	Checkpoint
BERT [2]	base	text	110M	Y	`bert-base-uncased`
BERT [2]	large	text	340M	Y	`bert-large-uncased`
RoBERTa [8]	base	text	124M	Y	`roberta-base`
RoBERTa [8]	large	text	355M	Y	`roberta-large`
ERNIEv2 [20, 21]	base	text	109M	Y*	`ernie-2.0-base-en`
ERNIEv2 [20, 21]	large	text	335M	Y*	`ernie-2.0-large-en`
DistilBERT [15]	base	text	66M	Y	`distilbert-base-uncased`
DistilRoBERTa [15]	base	text	82M	Y	`distilroberta-base`
SBERT [14]	–	text	23M	N	`paraphrase-MiniLM-L6-v2`
FLAVA [18]	–	text & VLP	109M	Y	`facebook/flava-full`
CLIP [12]	–	VLP	63M	N	`openai/clip-vit-base-patch32`
OpenCLIP [5]	–	VLP	352M	N	`laion/CLIP-ViT-H-14-laion2B-s32B-b79K`

Table 8. **Models table**. Note that the number of parameters listed for CLIP, OpenCLIP and FLAVA refers to their text encoder components alone. * Note: ERNIE was trained with an MLM head, but because the public checkpoints provided do not include this, we do not evaluate it with MLM probing.

Sentiment Analysis For sentiment analysis, we concatenate the following prompts to the given reviews and use the different options for sentiment prediction.

- *Is this review positive?* `[*]`*; Yes, No*
- *Is this a good movie?* `[*]`*; Yes, No*
- *I conclude the movie was* `[*]`*; good, bad*
- *The film was* `[*]`*; good, bad*
- *I had a* `[*]` *time; good, bad*
- *The following movie review expresses what sentiment?* `[*]`*; Positive, Negative*
- *Sentiment expressed for the movie is* `[*]`*; Positive, Negative*
- *The overall review of the film is* `[*]`*; good, bad*
- *The movie was* `[*]`*; good, bad*
- *This movie is* `[*]`*; good, bad*

References

[1] Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384, 2021. 1, 2

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[3] W Nelson Francis and Henry Kucera. A standard corpus of present-day edited american english, for use with digital computers. *Brown University, Providence*, 1964. 2

[4] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30, 2013. 4

[5] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. 6

[6] Yongmin Kim, Chenhui Chu, and Sadao Kurohashi. Flexible visual grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 285–299, 2022. 1

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 2

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6

[9] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 1

[10] Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. The world of an octopus: How reporting bias influences a language model's perception of color. *arXiv preprint arXiv:2110.08182*, 2021. 4

[11] Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. Weakly supervised face naming with symmetry-enhanced contrastive loss. *arXiv preprint arXiv:2210.08957*, 2022. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language super-

vision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[14] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 6

[15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 6

[16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 4

[17] Vered Shwartz and Yejin Choi. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, 2020. 4

[18] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 6

[19] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019. 3

[20] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019. 6

[21] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pretraining framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975, 2020. 6

[22] Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, and Jia Deng. Structured matching for phrase localization. In *ECCV*, pages 696–711, 2016. 1

[23] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017. 2, 3

[24] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*, 2019. 2