

2024 Jiangxi Provincial Collegiate Programming Contest - TechGroup

座位: 无

队伍: 无

提交时间: 2024-05-25T16:15:15.962+08:00

```
1 @String(PAMI = {IEEE Trans. Pattern Anal. Mach. Intell.})
2 @String(IJCV = {Int. J. Comput. Vis.})
3 @String(CVPR= {IEEE Conf. Comput. Vis. Pattern Recog.})
4 @String(ICCV= {Int. Conf. Comput. Vis.})
5 @String(ECCV= {Eur. Conf. Comput. Vis.})
6 @String(NIPS= {Adv. Neural Inform. Process. Syst.})
7 @String(ICPR = {Int. Conf. Pattern Recog.})
8 @String(BMVC= {Brit. Mach. Vis. Conf.})
9 @String(TOG= {ACM Trans. Graph.})
10 @String(TIP = {IEEE Trans. Image Process.})
11 @String(TVCG = {IEEE Trans. Vis. Comput. Graph.})
12 @String(TMM = {IEEE Trans. Multimedia})
13 @String(ACMMM= {ACM Int. Conf. Multimedia})
14 @String(ICME = {Int. Conf. Multimedia and Expo})
15 @String(ICASSP= {ICASSP})
16 @String(ICIP = {IEEE Int. Conf. Image Process.})
17 @String(ACCV = {ACCV})
18 @String(ICLR = {Int. Conf. Learn. Represent.})
19 @String(IJCAI = {IJCAI})
20 @String(PR = {Pattern Recognition})
21 @String(AAAI = {AAAI})
22 @String(CVPRW= {IEEE Conf. Comput. Vis. Pattern Recog. Worksh.})
23 @String(CSVT = {IEEE Trans. Circuit Syst. Video Technol.})
24
25 @String(SPL = {IEEE Sign. Process. Letters})
26 @String(VR = {Vis. Res.})
27 @String(JOV = {J. Vis.})
28 @String(TVC = {The Vis. Comput.})
29 @String(JCST = {J. Comput. Sci. Tech.})
30 @String(CGF = {Comput. Graph. Forum})
31 @String(CVM = {Computational Visual Media})
32
33
34 @String(PAMI = {IEEE TPAMI})
35 @String(IJCV = {IJCV})
36 @String(CVPR = {CVPR})
37 @String(ICCV = {ICCV})
38 @String(ECCV = {ECCV})
39 @String(NIPS = {NeurIPS})
40 @String(ICPR = {ICPR})
41 @String(BMVC = {BMVC})
42 @String(TOG = {ACM TOG})
43 @String(TIP = {IEEE TIP})
44 @String(TVCG = {IEEE TVCG})
45 @String(TCSVT = {IEEE TCSVT})
46 @String(TMM = {IEEE TMM})
47 @String(ACMMM = {ACM MM})
48 @String(ICME = {ICME})
49 @String(ICASSP= {ICASSP})
50 @String(ICIP = {ICIP})
51 @String(ACCV = {ACCV})
```

```
52 @String(ICLR = {ICLR})
53 @String(IJCAI = {IJCAI})
54 @String(PR = {PR})
55 @String(AAAI = {AAAI})
56 @String(CVPRW= {CVPRW})
57 @String(CSVT = {IEEE TCSVT})
58
59 % PROMPTING
60
61 @inproceedings{jacovi-goldberg-2020-towards,
62     title = "Towards Faithfully Interpretable {NLP} Systems: How Should We Define and Evaluate Faithfulness?",
63     author = "Jacovi, Alon  and
64         Goldberg, Yoav",
65     booktitle = "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics",
66     month = jul,
67     year = "2020",
68     address = "Online",
69     publisher = "Association for Computational Linguistics",
70     url = "https://aclanthology.org/2020.acl-main.386",
71     doi = "10.18653/v1/2020.acl-main.386",
72     pages = "4198--4205",
73     abstract = "With the growing popularity of deep-learning based NLP models, comes a need for interpretable systems. But what is interpretability, and what constitutes a high-quality interpretation? In this opinion piece we reflect on the current state of interpretability evaluation research. We call for more clearly differentiating between different desired criteria an interpretation should satisfy, and focus on the faithfulness criteria. We survey the literature with respect to faithfulness evaluation, and arrange the current approaches around three assumptions, providing an explicit form to how faithfulness is {'`'}defined{''}`'} by the community. We provide concrete guidelines on how evaluation of interpretation methods should and should not be conducted. Finally, we claim that the current binary definition for faithfulness sets a potentially unrealistic bar for being considered faithful. We call for discarding the binary notion of faithfulness in favor of a more graded one, which we believe will be of greater practical utility.",
74 }
75
76 @inproceedings{rajani-etal-2019-explain,
77     title = "Explain Yourself! Leveraging Language Models for Commonsense Reasoning",
78     author = "Rajani, Nazneen Fatema  and
79         McCann, Bryan  and
80         Xiong, Caiming  and
81         Socher, Richard",
82     booktitle = "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics",
83     month = jul,
84     year = "2019",
85     address = "Florence, Italy",
86     publisher = "Association for Computational Linguistics",
87     url = "https://aclanthology.org/P19-1487",
88     doi = "10.18653/v1/P19-1487",
89     pages = "4932--4942",
```

```
90     abstract = "Deep learning models perform poorly on tasks that require commonsense reasoning,
which often necessitates some form of world-knowledge or reasoning over information not immediately
present in the input. We collect human explanations for commonsense reasoning in the form of natural
language sequences and highlighted annotations in a new dataset called Common Sense Explanations
(CoS-E). We use CoS-E to train language models to automatically generate explanations that can be
used during training and inference in a novel Commonsense Auto-Generated Explanation (CAGE)
framework. CAGE improves the state-of-the-art by 10{\%} on the challenging CommonsenseQA task. We
further study commonsense reasoning in DNNs using both human and auto-generated explanations
including transfer to out-of-domain tasks. Empirical results indicate that we can effectively
leverage language models for commonsense reasoning.",
91 }
92
93 @article{narang2020wt5,
94     title={Wt5?! training text-to-text models to explain their predictions},
95     author={Narang, Sharan and Raffel, Colin and Lee, Katherine and Roberts, Adam and Fiedel, Noah and
Malkan, Karishma},
96     journal={arXiv preprint arXiv:2004.14546},
97     year={2020}
98 }
99
100 @article{marasovic2021few,
101     title={Few-shot self-rationalization with natural language prompts},
102     author={Marasovi{\c{c}}, Ana and Beltagy, Iz and Downey, Doug and Peters, Matthew E},
103     journal={arXiv preprint arXiv:2111.08284},
104     year={2021}
105 }
106
107 @inproceedings{zhao2021calibrate,
108     title={Calibrate before use: Improving few-shot performance of language models},
109     author={Zhao, Zihao and Wallace, Eric and Feng, Shi and Klein, Dan and Singh, Sameer},
110     booktitle={International Conference on Machine Learning},
111     pages={12697--12706},
112     year={2021},
113     organization={PMLR}
114 }
115
116 @article{lu2021fantastically,
117     title={Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order
sensitivity},
118     author={Lu, Yao and Bartolo, Max and Moore, Alastair and Riedel, Sebastian and Stenetorp, Pontus},
119     journal={arXiv preprint arXiv:2104.08786},
120     year={2021}
121 }
122
123 @misc{agrawal2023reassessing,
124     title={Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-
Distribution Generalization},
125     author={Aishwarya Agrawal and Ivana Kajić and Emanuele Bugliarello and Elnaz Davoodi and Anita
Gergely and Phil Blunsom and Aida Nematzadeh},
126     year={2023},
127     eprint={2205.12191},
128     archivePrefix={arXiv},
129     primaryClass={cs.CL}
130 }
131
132 @article{liu2023pre,
```

```
133   title={Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language
134   processing},
135   author={Liu, Pengfei and Yuan, Weizhe and Fu, Jinlan and Jiang, Zhengbao and Hayashi, Hiroaki and
136   Neubig, Graham},
137   journal={ACM Computing Surveys},
138   volume={55},
139   number={9},
140   pages={1--35},
141   year={2023},
142   publisher={ACM New York, NY}
143 }
144
145 @article{liu2021makes,
146   title={What Makes Good In-Context Examples for GPT-3?},
147   author={Liu, Jiachang and Shen, Dinghan and Zhang, Yizhe and Dolan, Bill and Carin, Lawrence and
148   Chen, Weizhu},
149   journal={arXiv preprint arXiv:2101.06804},
150   year={2021}
151 }
152
153 @article{gupta2022visual,
154   title={Visual Programming: Compositional visual reasoning without training},
155   author={Gupta, Tanmay and Kembhavi, Aniruddha},
156   journal={arXiv preprint arXiv:2211.11559},
157   year={2022}
158 }
159
160 @article{suris2023vipergpt,
161   title={ViperGPT: Visual Inference via Python Execution for Reasoning},
162   author={Sur{\`i}s, D{\`a}c and Menon, Sachit and Vondrick, Carl},
163   journal={arXiv preprint arXiv:2303.08128},
164   year={2023}
165 }
166
167 @article{wei2022chain,
168   title={Chain of thought prompting elicits reasoning in large language models},
169   author={Wei, Jason and Wang, Xuezhi and Schuurmans, Dale and Bosma, Maarten and Chi, Ed and Le,
170   Quoc and Zhou, Denny},
171   journal={arXiv preprint arXiv:2201.11903},
172   year={2022}
173 }
174
175 @article{brown2020language,
176   title={Language models are few-shot learners},
177   author={Brown, Tom and Mann, Benjamin and Ryder, Nick and Subbiah, Melanie and Kaplan, Jared D and
178   Dhariwal, Prafulla and Neelakantan, Arvind and Shyam, Pranav and Sastry, Girish and Askell, Amanda
179   and others},
180   journal={Advances in neural information processing systems},
181   volume={33},
182   pages={1877--1901},
183   year={2020}
184 }
185
186 @article{zhou2022least,
187   title={Least-to-most prompting enables complex reasoning in large language models},
188   author={Zhou, Denny and Sch{\`a}rli, Nathanael and Hou, Le and Wei, Jason and Scales, Nathan and
189   Wang, Xuezhi and Schuurmans, Dale and Bousquet, Olivier and Le, Quoc and Chi, Ed},
```

```
182   journal={arXiv preprint arXiv:2205.10625},
183   year={2022}
184 }
185 @article{lester2021power,
186   title={The power of scale for parameter-efficient prompt tuning},
187   author={Lester, Brian and Al-Rfou, Rami and Constant, Noah},
188   journal={arXiv preprint arXiv:2104.08691},
189   year={2021}
190 }
191 @article{kojima2022large,
192   title={Large language models are zero-shot reasoners},
193   author={Kojima, Takeshi and Gu, Shixiang Shane and Reid, Machel and Matsuo, Yutaka and Iwasawa,
194     Yusuke},
195   journal={arXiv preprint arXiv:2205.11916},
196   year={2022}
197 }
198 @article{jin2021good,
199   title={A good prompt is worth millions of parameters? low-resource prompt-based learning for
200     vision-language models},
201   author={Jin, Woojeong and Cheng, Yu and Shen, Yelong and Chen, Weizhu and Ren, Xiang},
202   journal={arXiv preprint arXiv:2110.08484},
203   year={2021}
204 }
205 @inproceedings{yang2022empirical,
206   title={An empirical study of gpt-3 for few-shot knowledge-based vqa},
207   author={Yang, Zhengyuan and Gan, Zhe and Wang, Jianfeng and Hu, Xiaowei and Lu, Yumao and Liu,
208     Zicheng and Wang, Lijuan},
209   booktitle={Proceedings of the AAAI Conference on Artificial Intelligence},
210   volume={36},
211   number={3},
212   pages={3081--3089},
213   year={2022}
214 }
215 @article{hu2022promptcap,
216   title={PromptCap: Prompt-Guided Task-Aware Image Captioning},
217   author={Hu, Yushi and Hua, Hang and Yang, Zhengyuan and Shi, Weijia and Smith, Noah A and Luo,
218     Jiebo},
219   journal={arXiv preprint arXiv:2211.09699},
220   year={2022}
221 }
222 @article{lu2022learn,
223   title={Learn to explain: Multimodal reasoning via thought chains for science question answering},
224   author={Lu, Pan and Mishra, Swaroop and Xia, Tanglin and Qiu, Liang and Chang, Kai-Wei and Zhu,
225     Song-Chun and Tafjord, Oyvind and Clark, Peter and Kalyan, Ashwin},
226   journal={Advances in Neural Information Processing Systems},
227   volume={35},
228   pages={2507--2521},
229   year={2022}
230 }
231 @article{min2022rethinking,
232   title={Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?},
233   author={Min, Sewon and Lyu, Xinxi and Holtzman, Ari and Artetxe, Mikel and Lewis, Mike and
234     Hajishirzi, Hannaneh and Zettlemoyer, Luke},
```

```
232   journal={arXiv preprint arXiv:2202.12837},
233   year={2022}
234 }
235
236 @article{zhang2023multimodal,
237   title={Multimodal chain-of-thought reasoning in language models},
238   author={Zhang, Zhuosheng and Zhang, Aston and Li, Mu and Zhao, Hai and Karypis, George and Smola,
  Alex},
239   journal={arXiv preprint arXiv:2302.00923},
240   year={2023}
241 }
242 @article{eichenberg2021magma,
243   title={MAGMA--Multimodal Augmentation of Generative Models through Adapter-based Finetuning},
244   author={Eichenberg, Constantin and Black, Sidney and Weinbach, Samuel and Parcalabescu, Letitia and
  Frank, Anette},
245   journal={arXiv preprint arXiv:2112.05253},
246   year={2021}
247 }
248
249 @article{wang2022image,
250   title={Image as a foreign language: Beit pretraining for all vision and vision-language tasks},
251   author={Wang, Wenhui and Bao, Hangbo and Dong, Li and Bjorck, Johan and Peng, Zhiliang and Liu,
  Qiang and Aggarwal, Kriti and Mohammed, Owais Khan and Singhal, Saksham and Som, Subhojit and
  others},
252   journal={arXiv preprint arXiv:2208.10442},
253   year={2022}
254 }
255
256 @article{ouyang2022training,
257   title={Training language models to follow instructions with human feedback},
258   author={Ouyang, Long and Wu, Jeffrey and Jiang, Xu and Almeida, Diogo and Wainwright, Carroll and
  Mishkin, Pamela and Zhang, Chong and Agarwal, Sandhini and Slama, Katarina and Ray, Alex and others},
259   journal={Advances in Neural Information Processing Systems},
260   volume={35},
261   pages={27730--27744},
262   year={2022}
263 }
264
265 @article{nye2021show,
266   title={Show your work: Scratchpads for intermediate computation with language models},
267   author={Nye, Maxwell and Andreassen, Anders Johan and Gur-Ari, Guy and Michalewski, Henryk and
  Austin, Jacob and Bieber, David and Dohan, David and Lewkowycz, Aitor and Bosma, Maarten and Luan,
  David and others},
268   journal={arXiv preprint arXiv:2112.00114},
269   year={2021}
270 }
271
272 @article{wang2022self,
273   title={Self-consistency improves chain of thought reasoning in language models},
274   author={Wang, Xuezhi and Wei, Jason and Schuurmans, Dale and Le, Quoc and Chi, Ed and Zhou, Denny},
275   journal={arXiv preprint arXiv:2203.11171},
276   year={2022}
277 }
278
279 @article{zhang2022automatic,
280   title={Automatic chain of thought prompting in large language models},
```

```
281   author={Zhang, Zhuosheng and Zhang, Aston and Li, Mu and Smola, Alex},
282   journal={arXiv preprint arXiv:2210.03493},
283   year={2022}
284 }
285
286 @article{chen2021evaluating,
287   title={Evaluating large language models trained on code},
288   author={Chen, Mark and Tworek, Jerry and Jun, Heewoo and Yuan, Qiming and Pinto, Henrique Ponde de
Oliveira and Kaplan, Jared and Edwards, Harri and Burda, Yuri and Joseph, Nicholas and Brockman, Greg
and others},
289   journal={arXiv preprint arXiv:2107.03374},
290   year={2021}
291 }
292 % DATASET
293 @inproceedings{Schwenk2022A0KVQAAB,
294   title={A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge},
295   author={Dustin Schwenk and Apoorv Khandelwal and Christopher Clark and Kenneth Marino and Roozbeh
Mottaghi},
296   booktitle={European Conference on Computer Vision},
297   year={2022}
298 }
299 @inproceedings{hudson2019gqa,
300   title={Gqa: A new dataset for real-world visual reasoning and compositional question answering},
301   author={Hudson, Drew A and Manning, Christopher D},
302   booktitle={Proceedings of the IEEE/CVF conference on computer vision and pattern recognition},
303   pages={6700--6709},
304   year={2019}
305 }
306
307 @inproceedings{marino2019ok,
308   title={Ok-vqa: A visual question answering benchmark requiring external knowledge},
309   author={Marino, Kenneth and Rastegari, Mohammad and Farhadi, Ali and Mottaghi, Roozbeh},
310   booktitle={Proceedings of the IEEE/cvf conference on computer vision and pattern recognition},
311   pages={3195--3204},
312   year={2019}
313 }
314
315 @inproceedings{thrush2022winoground,
316   title={Winoground: Probing vision and language models for visio-linguistic compositionality},
317   author={Thrush, Tristan and Jiang, Ryan and Bartolo, Max and Singh, Amanpreet and Williams, Adina
and Kiela, Douwe and Ross, Candace},
318   booktitle={Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition},
319   pages={5238--5248},
320   year={2022}
321 }
322
323 @inproceedings{goyal2017making,
324   title={Making the v in vqa matter: Elevating the role of image understanding in visual question
answering},
325   author={Goyal, Yash and Khot, Tejas and Summers-Stay, Douglas and Batra, Dhruv and Parikh, Devi},
326   booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},
327   pages={6904--6913},
328   year={2017}
329 }
330 @inproceedings{zhu2016visual7w,
331   title={Visual7w: Grounded question answering in images},
```

```
332   author={Zhu, Yuke and Groth, Oliver and Bernstein, Michael and Fei-Fei, Li},
333   booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},
334   pages={4995--5004},
335   year={2016}
336 }
337 @inproceedings{johnson2017clevr,
338   title={Clevr: A diagnostic dataset for compositional language and elementary visual reasoning},
339   author={Johnson, Justin and Hariharan, Bharath and Van Der Maaten, Laurens and Fei-Fei, Li and
Lawrence Zitnick, C and Girshick, Ross},
340   booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},
341   pages={2901--2910},
342   year={2017}
343 }
344
345 % VL MODELS
346 @article{Zhang2022OPTOP,
347   title={OPT: Open Pre-trained Transformer Language Models},
348   author={Susan Zhang and Stephen Roller and Naman Goyal and Mikel Artetxe and Moya Chen and Shuohui
Chen and Christopher Dewan and Mona Diab and Xian Li and Xi Victoria Lin and Todor Mihaylov and Myle
Ott and Sam Shleifer and Kurt Shuster and Daniel Simig and Punit Singh Koura and Anjali Sridhar and
Tianlu Wang and Luke Zettlemoyer},
349   journal={ArXiv},
350   year={2022},
351   volume={abs/2205.01068}
352 }
353 @article{Dosovitskiy2020AnII,
354   title={An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale},
355   author={Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and
Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and
Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby},
356   journal={ArXiv},
357   year={2020},
358   volume={abs/2010.11929}
359 }
360 @article{tsimpoukelli2021multimodal,
361   title={Multimodal few-shot learning with frozen language models},
362   author={Tsimpoukelli, Maria and Menick, Jacob and Cabi, Serkan
and Eslami, SM and Vinyals, Oriol and Hill, Felix},
363   journal={Proc. Neural Information Processing Systems},
364   year={2021}
365 }
366 }
367
368 @article{chen2022pali,
369   title={Pali: A jointly-scaled multilingual language-image model},
370   author={Chen, Xi and Wang, Xiao and Changpinyo, Soravit and Piergiovanni, AJ and Padlewski, Piotr
and Salz, Daniel and Goodman, Sebastian and Grycner, Adam and Mustafa, Basil and Beyer, Lucas and
others},
371   journal={arXiv preprint arXiv:2209.06794},
372   year={2022}
373 }
374 @article{driess2023palm,
375   title={Palm-e: An embodied multimodal language model},
376   author={Driess, Danny and Xia, Fei and Sajjadi, Mehdi SM and Lynch, Corey and Chowdhery, Aakanksha
and Ichter, Brian and Wahid, Ayzaan and Tompson, Jonathan and Vuong, Quan and Yu, Tianhe and others},
377   journal={arXiv preprint arXiv:2303.03378},
378   year={2023}
```



```
379 }
380
381 @article{chung2022scaling,
382   title={Scaling instruction-finetuned language models},
383   author={Chung, Hyung Won and Hou, Le and Longpre, Shayne and Zoph, Barret and Tay, Yi and Fedus, William and Li, Eric and Wang, Xuezhi and Dehghani, Mostafa and Brahma, Siddhartha and others},
384   journal={arXiv preprint arXiv:2210.11416},
385   year={2022}
386 }
387
388 @article{tan2019lxmert,
389   title={Lxmert: Learning cross-modality encoder representations from transformers},
390   author={Tan, Hao and Bansal, Mohit},
391   journal={arXiv preprint arXiv:1908.07490},
392   year={2019}
393 }
394 @article{lu2019vilbert,
395   title={Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks},
396   author={Lu, Jiasen and Batra, Dhruv and Parikh, Devi and Lee, Stefan},
397   journal={Advances in neural information processing systems},
398   volume={32},
399   year={2019}
400 }
401
402 @article{manas2022mapl,
403   title={MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting},
404   author={Ma{\~n}as, Oscar and Rodriguez, Pau and Ahmadi, Saba and Nematzadeh, Aida and Goyal, Yash and Agrawal, Aishwarya},
405   journal={arXiv preprint arXiv:2210.07179},
406   year={2022}
407 }
408
409 @article{alayrac2022flamingo,
410   title={Flamingo: a visual language model for few-shot learning},
411   author={Alayrac, Jean-Baptiste and Donahue, Jeff and Luc, Pauline and Miech, Antoine and Barr, Iain and Hasson, Yana and Lenc, Karel and Mensch, Arthur and Millican, Katie and Reynolds, Malcolm and others},
412   journal={arXiv preprint arXiv:2204.14198},
413   year={2022}
414 }
415
416 @article{zeng2021multi,
417   title={Multi-grained vision language pre-training: Aligning texts with visual concepts},
418   author={Zeng, Yan and Zhang, Xinsong and Li, Hang},
419   journal={arXiv preprint arXiv:2111.08276},
420   year={2021}
421 }
422
423 @article{li2023blip,
424   title={Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models},
425   author={Li, Junnan and Li, Dongxu and Savarese, Silvio and Hoi, Steven},
426   journal={arXiv preprint arXiv:2301.12597},
427   year={2023}
```

```
428 }
429
430 @inproceedings{li2022blip,
431     title={BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language
432     Understanding and Generation},
433     author={Junnan Li and Dongxu Li and Caiming Xiong and Steven Hoi},
434     year={2022},
435     booktitle={ICML},
436 }
437
438 @inproceedings{herdade2019imagecap,
439     author = {Herdade, Simao and Kappeler, Armin and Boakye, Kofi and Soares, Joao},
440     booktitle = {Advances in Neural Information Processing Systems},
441     editor = {H. Wallach and H. Larochelle and A. Beygelzimer and F. d\textquotesingle Alch\'{e}-Buc and
442     E. Fox and R. Garnett},
443     pages = {},
444     publisher = {Curran Associates, Inc.},
445     title = {Image Captioning: Transforming Objects into Words},
446     url = {https://proceedings.neurips.cc/paper_files/paper/2019/file/680390c55bbd9ce416d1d69a9ab4760d-
447     Paper.pdf},
448     volume = {32},
449     year = {2019}
450 }
451
452 @inproceedings{koh2023grounding,
453     title={Grounding Language Models to Images for Multimodal Generation},
454     author={Koh, Jing Yu and Salakhutdinov, Ruslan and Fried, Daniel},
455     journal={arXiv:2301.13823},
456     year={2023}
457 }
458
459 @inproceedings{anderson2018bottom,
460     title={Bottom-up and top-down attention for image captioning and visual question answering},
461     author={Anderson, Peter and He, Xiaodong and Buehler, Chris and Teney, Damien and Johnson, Mark and
462     Gould, Stephen and Zhang, Lei},
463     booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},
464     pages={6077--6086},
465     year={2018}
466 }
467
468 @inproceedings{xu2015show,
469     title={Show, attend and tell: Neural image caption generation with visual attention},
470     author={Xu, Kelvin and Ba, Jimmy and Kiros, Ryan and Cho, Kyunghyun and Courville, Aaron and
471     Salakhudinov, Ruslan and Zemel, Rich and Bengio, Yoshua},
472     booktitle={International conference on machine learning},
473     pages={2048--2057},
474     year={2015},
475     organization={PMLR}
476 }
477
478 % ANALYSIS
479
480 @InProceedings{pmlr-v162-zhou22n,
481     title =    {{VLUe}: A Multi-Task Multi-Dimension Benchmark for Evaluating Vision-Language Pre-
482     training},
483     author =    {Zhou, Wangchunshu and Zeng, Yan and Diao, Shizhe and Zhang, Xinsong},
484     booktitle =    {Proceedings of the 39th International Conference on Machine Learning},
```

```
478   pages =    {27395--27411},
479   year =     {2022},
480   editor =   {Chaudhuri, Kamalika and Jegelka, Stefanie and Song, Le and Szepesvari, Csaba and Niu,
Gang and Sabato, Sivan},
481   volume =   {162},
482   series =   {Proceedings of Machine Learning Research},
483   month =    {17--23 Jul},
484   publisher = {PMLR},
485   pdf =      {https://proceedings.mlr.press/v162/zhou22n/zhou22n.pdf},
486   url =      {https://proceedings.mlr.press/v162/zhou22n.html},
487   abstract = {Recent advances in vision-language pre-training (VLP) have demonstrated impressive
performance in a range of vision-language (VL) tasks. However, there exist several challenges for
measuring the community's progress in building general multi-modal intelligence. First, most of the
downstream VL datasets are annotated using raw images that are already seen during pre-training,
which may result in an overestimation of current VLP models' generalization ability. Second, recent
VLP work mainly focuses on absolute performance but overlooks the efficiency-performance trade-off,
which is also an important indicator for measuring progress. To this end, we introduce the Vision-
Language Understanding Evaluation (VLUE) benchmark, a multi-task multi-dimension benchmark for
evaluating the generalization capabilities and the efficiency-performance trade-off ("Pareto SOTA")
of VLP models. We demonstrate that there is a sizable generalization gap for all VLP models when
testing on out-of-distribution test sets annotated on images from a more diverse distribution that
spreads across cultures. Moreover, we find that measuring the efficiency-performance trade-off of VLP
models leads to complementary insights for several design choices of VLP. We release the VLUE
benchmark to promote research on building vision-language models that generalize well to images
unseen during pre-training and are practical in terms of efficiency-performance trade-off.}
488 }
489
490 @article{yuksekgonul2022and,
491   title={When and why vision-language models behave like bag-of-words models, and what to do about
it?},
492   author={Yuksekgonul, Mert and Bianchi, Federico and Kalluri, Pratyusha and Jurafsky, Dan and Zou,
James},
493   journal={arXiv preprint arXiv:2210.01936},
494   year={2022}
495 }
496 @article{diwan2022winoground,
497   title={Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality},
498   author={Diwan, Anuj and Berry, Layne and Choi, Eunsol and Harwath, David and Mahowald, Kyle},
499   journal={arXiv preprint arXiv:2211.00768},
500   year={2022}
501 }
502
503 @inproceedings{selvaraju2017grad,
504   title={Grad-cam: Visual explanations from deep networks via gradient-based localization},
505   author={Selvaraju, Ramprasaath R and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna
and Parikh, Devi and Batra, Dhruv},
506   booktitle={Proceedings of the IEEE international conference on computer vision},
507   pages={618--626},
508   year={2017}
509 }
510
511 % TASK
512 @inproceedings{antol2015vqa,
513   title={Vqa: Visual question answering},
514   author={Antol, Stanislaw and Agrawal, Aishwarya and Lu, Jiasen and Mitchell, Margaret and Batra,
Dhruv and Zitnick, C Lawrence and Parikh, Devi},
515   booktitle={Proceedings of the IEEE international conference on computer vision},
```

```
516   pages={2425--2433},
517   year={2015}
518 }
519
520 @article{srivastava2022beyond,
521   title={Beyond the imitation game: Quantifying and extrapolating the capabilities of language
models},
522   author={Srivastava, Aarohi and Rastogi, Abhinav and Rao, Abhishek and Shueb, Abu Awal Md and Abid,
Abubakar and Fisch, Adam and Brown, Adam R and Santoro, Adam and Gupta, Aditya and Garriga-Alonso,
Adri{\`a} and others},
523   journal={arXiv preprint arXiv:2206.04615},
524   year={2022}
525 }
526
527 @article{zhang2019bertscore,
528   title={Bertscore: Evaluating text generation with bert},
529   author={Zhang, Tianyi and Kishore, Varsha and Wu, Felix and Weinberger, Kilian Q and Artzi, Yoav},
530   journal={arXiv preprint arXiv:1904.09675},
531   year={2019}
532 }
533
534 @inproceedings{radford2021learning,
535   title={Learning transferable visual models from natural language supervision},
536   author={Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and
Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and
others},
537   booktitle={International conference on machine learning},
538   pages={8748--8763},
539   year={2021},
540   organization={PMLR}
541 }
542
543
544 @inproceedings{shao2023prompting,
545   title={Prompting large language models with answer heuristics for knowledge-based visual question
answering},
546   author={Shao, Zhenwei and Yu, Zhou and Wang, Meng and Yu, Jun},
547   booktitle={Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition},
548   pages={14974--14983},
549   year={2023}
550 }
551
552 @article{lu2022unified,
553   title={Unified-io: A unified model for vision, language, and multi-modal tasks},
554   author={Lu, Jiasen and Clark, Christopher and Zellers, Rowan and Mottaghi, Roozbeh and Kembhavi,
Aniruddha},
555   journal={arXiv preprint arXiv:2206.08916},
556   year={2022}
557 }
558
559 @inproceedings{guo2023images,
560   title={From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large
Language Models},
561   author={Guo, Jiaxian and Li, Junnan and Li, Dongxu and Tiong, Anthony Meng Huat and Li, Boyang and
Tao, Dacheng and Hoi, Steven},
562   booktitle={Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition},
```

```
563   pages={10867--10877},
564   year={2023}
565 }
566
567 @article{peng2023kosmos,
568   title={Kosmos-2: Grounding Multimodal Large Language Models to the World},
569   author={Peng, Zhiliang and Wang, Wenhui and Dong, Li and Hao, Yaru and Huang, Shaohan and Ma,
570     Shuming and Wei, Furu},
571   journal={arXiv preprint arXiv:2306.14824},
572   year={2023}
573 }
574
575 @article{awadalla2023openflamingo,
576   title={Openflamingo: An open-source framework for training large autoregressive vision-language
577     models},
578   author={Awadalla, Anas and Gao, Irena and Gardner, Josh and Hessel, Jack and Hanafy, Yusuf and Zhu,
579     Wanrong and Marathe, Kalyani and Bitton, Yonatan and Gadre, Samir and Sagawa, Shiori and others},
580   journal={arXiv preprint arXiv:2308.01390},
581   year={2023}
582 }
583
584 @article{liu2023visual,
585   title={Visual instruction tuning},
586   author={Liu, Haotian and Li, Chunyuan and Wu, Qingyang and Lee, Yong Jae},
587   journal={arXiv preprint arXiv:2304.08485},
588   year={2023}
589 }
590
591 @inproceedings{agrawal-etal-2023-reassessing,
592   title = "Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-
593     Distribution Generalization",
594   author = "Agrawal, Aishwarya and
595     Kajic, Ivana and
596     Bugliarello, Emanuele and
597     Davoodi, Elnaz and
598     Gergely, Anita and
599     Blunsom, Phil and
600     Nematzadeh, Aida",
601   editor = "Vlachos, Andreas and
602     Augenstein, Isabelle",
603   booktitle = "Findings of the Association for Computational Linguistics: EACL 2023",
604   month = may,
605   year = "2023",
606   address = "Dubrovnik, Croatia",
607   publisher = "Association for Computational Linguistics",
608   url = "https://aclanthology.org/2023.findings-eacl.90",
609   doi = "10.18653/v1/2023.findings-eacl.90",
610   pages = "1201--1226",
```

```
607     abstract = "Vision-and-language (V\&L) models pretrained on large-scale multimodal data have
608     demonstrated strong performance on various tasks such as image captioning and visual question
609     answering (VQA). The quality of such models is commonly assessed by measuring their performance on
610     unseen data that typically comes from the same distribution as the training data. However, when
611     evaluated under out-of-distribution (out-of-dataset) settings for VQA, we observe that these models
612     exhibit poor generalization. We comprehensively evaluate two pretrained V\&L models under different
613     settings (i.e. classification and open-ended text generation) by conducting cross-dataset
614     evaluations. We find that these models tend to learn to solve the benchmark, rather than learning the
615     high-level skills required by the VQA task. We also find that in most cases generative models are
616     less susceptible to shifts in data distribution compared to discriminative ones, and that multimodal
617     pretraining is generally helpful for OOD generalization. Finally, we revisit assumptions underlying
618     the use of automatic VQA evaluation metrics, and empirically show that their stringent nature
619     repeatedly penalizes models for correct responses.",
620 }
621
622 @article{manas2023improving,
623     title={Improving Automatic VQA Evaluation Using Large Language Models},
624     author={Ma\~n}as, Oscar and Krojer, Benno and Agrawal, Aishwarya},
625     journal={arXiv preprint arXiv:2310.02567},
626     year={2023}
627 }
628
629 @article{zhu2023minigpt,
630     title={Minigpt-4: Enhancing vision-language understanding with advanced large language models},
631     author={Zhu, Deyao and Chen, Jun and Shen, Xiaoqian and Li, Xiang and Elhoseiny, Mohamed},
632     journal={arXiv preprint arXiv:2304.10592},
633     year={2023}
634 }
635
636 @article{guo2022images,
637     title={From images to textual prompts: Zero-shot vqa with frozen large language models},
638     author={Guo, Jiaxian and Li, Junnan and Li, Dongxu and Tiong, Anthony Meng Huat and Li, Boyang and
639     Tao, Dacheng and Hoi, Steven CH},
640     journal={arXiv preprint arXiv:2212.10846},
641     year={2022}
642 }
643
644 @article{qiao2022reasoning,
645     title={Reasoning with language model prompting: A survey},
646     author={Qiao, Shuofei and Ou, Yixin and Zhang, Ningyu and Chen, Xiang and Yao, Yunzhi and Deng,
647     Shumin and Tan, Chuanqi and Huang, Fei and Chen, Huajun},
648     journal={arXiv preprint arXiv:2212.09597},
649     year={2022}
650 }
651
652 @article{agrawal2022rethinking,
653     title={Rethinking evaluation practices in visual question answering: A case study on out-of-
654     distribution generalization},
655     author={Agrawal, Aishwarya and Kaji\c, Ivana and Bugliarello, Emanuele and Davoodi, Elnaz and
656     Gergely, Anita and Blunsom, Phil and Nematzadeh, Aida},
657     journal={arXiv preprint arXiv:2205.12191},
658     year={2022}
659 }
```

```
648
649 @article{touvron2023llama,
650     title={Llama 2: Open foundation and fine-tuned chat models},
651     author={Touvron, Hugo and Martin, Louis and Stone, Kevin and Albert, Peter and Almahairi, Amjad and
        Babaei, Yasmine and Bashlykov, Nikolay and Batra, Soumya and Bhargava, Prajjwal and Bhosale, Shruti
        and others},
652     journal={arXiv preprint arXiv:2307.09288},
653     year={2023}
654 }
```