2024 Jiangxi Provincial Collegiate Programming Contest - TechGroup

座位: 无

队伍: 无

提交时间: 2024-05-25T16:17:09.648+08:00

```
@String(PAMI = {IEEE Trans. Pattern Anal. Mach. Intell.})
@String(IJCV = {Int. J. Comput. Vis.})
@String(CVPR= {IEEE Conf. Comput. Vis. Pattern Recog.})
@String(ICCV= {Int. Conf. Comput. Vis.})
@String(ECCV= {Eur. Conf. Comput. Vis.})
@String(NIPS= {Adv. Neural Inform. Process. Syst.})
@String(ICPR = {Int. Conf. Pattern Recog.})
@String(BMVC= {Brit. Mach. Vis. Conf.})
@String(TOG= {ACM Trans. Graph.})
@String(TIP  = {IEEE Trans. Image Process.})
@String(TVCG  = {IEEE Trans. Vis. Comput. Graph.})
@String(TMM  = {IEEE Trans. Multimedia})
@String(ACMMM= {ACM Int. Conf. Multimedia})
@String(ICME = {Int. Conf. Multimedia and Expo})
@String(ICASSP= {ICASSP})
@String(ICIP = {IEEE Int. Conf. Image Process.})
@String(ACCV  = {ACCV})
@String(ICLR = {Int. Conf. Learn. Represent.})
@String(IJCAI = {IJCAI})
@String(PR   = {Pattern Recognition})
@String(AAAI = {AAAI})
@String(CVPRW= {IEEE Conf. Comput. Vis. Pattern Recog. Worksh.})
@String(CSVT = {IEEE Trans. Circuit Syst. Video Technol.})

@String(SPL = {IEEE Sign. Process. Letters})
@String(VR   = {Vis. Res.})
@String(JOV  = {J. Vis.})
@String(TVC  = {The Vis. Comput.})
@String(JCST  = {J. Comput. Sci. Tech.})
@String(CGF  = {Comput. Graph. Forum})
@String(CVM = {Computational Visual Media})


@String(PAMI  = {IEEE TPAMI})
@String(IJCV  = {IJCV})
@String(CVPR  = {CVPR})
@String(ICCV  = {ICCV})
@String(ECCV  = {ECCV})
@String(NIPS  = {NeurIPS})
@String(ICPR  = {ICPR})
@String(BMVC  = {BMVC})
@String(TOG   = {ACM TOG})
@String(TIP   = {IEEE TIP})
@String(TVCG  = {IEEE TVCG})
@String(TCSVT = {IEEE TCSVT})
@String(TMM   = {IEEE TMM})
@String(ACMMM = {ACM MM})
@String(ICME  = {ICME})
@String(ICASSP= {ICASSP})
@String(ICIP  = {ICIP})
@String(ACCV  = {ACCV})
```

```
@String(ICLR  = {ICLR})
@String(IJCAI = {IJCAI})
@String(PR = {PR})
@String(AAAI = {AAAI})
@String(CVPRW= {CVPRW})
@String(CSVT = {IEEE TCSVT})

% PROMPTING

@inproceedings{jacovi-goldberg-2020-towards,
    title = "Towards Faithfully Interpretable {NLP} Systems: How Should We Define and Evaluate Faithfulness?",
    author = "Jacovi, Alon  and
      Goldberg, Yoav",
    booktitle = "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics",
    month = jul,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2020.acl-main.386",
    doi = "10.18653/v1/2020.acl-main.386",
    pages = "4198--4205",
    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.",
}

@inproceedings{rajani-etal-2019-explain,
    title = "Explain Yourself! Leveraging Language Models for Commonsense Reasoning",
    author = "Rajani, Nazneen Fatema  and
      McCann, Bryan  and
      Xiong, Caiming  and
      Socher, Richard",
    booktitle = "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics",
    month = jul,
    year = "2019",
    address = "Florence, Italy",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/P19-1487",
    doi = "10.18653/v1/P19-1487",
    pages = "4932--4942",
```

```
    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.",
}

@article{narang2020wt5,
  title={Wt5?! training text-to-text models to explain their predictions},
  author={Narang, Sharan and Raffel, Colin and Lee, Katherine and Roberts, Adam and Fiedel, Noah and Malkan, Karishma},
  journal={arXiv preprint arXiv:2004.14546},
  year={2020}
}

@article{marasovic2021few,
  title={Few-shot self-rationalization with natural language prompts},
  author={Marasovi{\'c}, Ana and Beltagy, Iz and Downey, Doug and Peters, Matthew E},
  journal={arXiv preprint arXiv:2111.08284},
  year={2021}
}

@inproceedings{zhao2021calibrate,
  title={Calibrate before use: Improving few-shot performance of language models},
  author={Zhao, Zihao and Wallace, Eric and Feng, Shi and Klein, Dan and Singh, Sameer},
  booktitle={International Conference on Machine Learning},
  pages={12697--12706},
  year={2021},
  organization={PMLR}
}

@article{lu2021fantastically,
  title={Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity},
  author={Lu, Yao and Bartolo, Max and Moore, Alastair and Riedel, Sebastian and Stenetorp, Pontus},
  journal={arXiv preprint arXiv:2104.08786},
  year={2021}
}

@misc{agrawal2023reassessing,
      title={Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization},
      author={Aishwarya Agrawal and Ivana Kajić and Emanuele Bugliarello and Elnaz Davoodi and Anita Gergely and Phil Blunsom and Aida Nematzadeh},
      year={2023},
      eprint={2205.12191},
      archivePrefix={arXiv},
      primaryClass={cs.CL}
}

@article{liu2023pre,
```

```
  title={Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing},
  author={Liu, Pengfei and Yuan, Weizhe and Fu, Jinlan and Jiang, Zhengbao and Hayashi, Hiroaki and Neubig, Graham},
  journal={ACM Computing Surveys},
  volume={55},
  number={9},
  pages={1--35},
  year={2023},
  publisher={ACM New York, NY}
}

@article{liu2021makes,
  title={What Makes Good In-Context Examples for GPT-$3 $?},
  author={Liu, Jiachang and Shen, Dinghan and Zhang, Yizhe and Dolan, Bill and Carin, Lawrence and Chen, Weizhu},
  journal={arXiv preprint arXiv:2101.06804},
  year={2021}
}

@article{gupta2022visual,
  title={Visual Programming: Compositional visual reasoning without training},
  author={Gupta, Tanmay and Kembhavi, Aniruddha},
  journal={arXiv preprint arXiv:2211.11559},
  year={2022}
}

@article{suris2023vipergpt,
  title={ViperGPT: Visual Inference via Python Execution for Reasoning},
  author={Sur{\'\i}s, D{\'\i}dac and Menon, Sachit and Vondrick, Carl},
  journal={arXiv preprint arXiv:2303.08128},
  year={2023}
}

@article{wei2022chain,
  title={Chain of thought prompting elicits reasoning in large language models},
  author={Wei, Jason and Wang, Xuezhi and Schuurmans, Dale and Bosma, Maarten and Chi, Ed and Le, Quoc and Zhou, Denny},
  journal={arXiv preprint arXiv:2201.11903},
  year={2022}
}
@article{brown2020language,
  title={Language models are few-shot learners},
  author={Brown, Tom and Mann, Benjamin and Ryder, Nick and Subbiah, Melanie and Kaplan, Jared D and Dhariwal, Prafulla and Neelakantan, Arvind and Shyam, Pranav and Sastry, Girish and Askell, Amanda and others},
  journal={Advances in neural information processing systems},
  volume={33},
  pages={1877--1901},
  year={2020}
}

@article{zhou2022least,
  title={Least-to-most prompting enables complex reasoning in large language models},
  author={Zhou, Denny and Sch{\"a}rli, Nathanael and Hou, Le and Wei, Jason and Scales, Nathan and Wang, Xuezhi and Schuurmans, Dale and Bousquet, Olivier and Le, Quoc and Chi, Ed},
```

```
  journal={arXiv preprint arXiv:2205.10625},
  year={2022}
}
@article{lester2021power,
  title={The power of scale for parameter-efficient prompt tuning},
  author={Lester, Brian and Al-Rfou, Rami and Constant, Noah},
  journal={arXiv preprint arXiv:2104.08691},
  year={2021}
}
@article{kojima2022large,
  title={Large language models are zero-shot reasoners},
  author={Kojima, Takeshi and Gu, Shixiang Shane and Reid, Machel and Matsuo, Yutaka and Iwasawa, Yusuke},
  journal={arXiv preprint arXiv:2205.11916},
  year={2022}
}
@article{jin2021good,
  title={A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models},
  author={Jin, Woojeong and Cheng, Yu and Shen, Yelong and Chen, Weizhu and Ren, Xiang},
  journal={arXiv preprint arXiv:2110.08484},
  year={2021}
}

@inproceedings{yang2022empirical,
  title={An empirical study of gpt-3 for few-shot knowledge-based vqa},
  author={Yang, Zhengyuan and Gan, Zhe and Wang, Jianfeng and Hu, Xiaowei and Lu, Yumao and Liu, Zicheng and Wang, Lijuan},
  booktitle={Proceedings of the AAAI Conference on Artificial Intelligence},
  volume={36},
  number={3},
  pages={3081--3089},
  year={2022}
}
@article{hu2022promptcap,
  title={PromptCap: Prompt-Guided Task-Aware Image Captioning},
  author={Hu, Yushi and Hua, Hang and Yang, Zhengyuan and Shi, Weijia and Smith, Noah A and Luo, Jiebo},
  journal={arXiv preprint arXiv:2211.09699},
  year={2022}
}

@article{lu2022learn,
  title={Learn to explain: Multimodal reasoning via thought chains for science question answering},
  author={Lu, Pan and Mishra, Swaroop and Xia, Tanglin and Qiu, Liang and Chang, Kai-Wei and Zhu, Song-Chun and Tafjord, Oyvind and Clark, Peter and Kalyan, Ashwin},
  journal={Advances in Neural Information Processing Systems},
  volume={35},
  pages={2507--2521},
  year={2022}
}

@article{min2022rethinking,
  title={Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?},
  author={Min, Sewon and Lyu, Xinxi and Holtzman, Ari and Artetxe, Mikel and Lewis, Mike and Hajishirzi, Hannaneh and Zettlemoyer, Luke},
```

```
  journal={arXiv preprint arXiv:2202.12837},
  year={2022}
}

@article{zhang2023multimodal,
  title={Multimodal chain-of-thought reasoning in language models},
  author={Zhang, Zhuosheng and Zhang, Aston and Li, Mu and Zhao, Hai and Karypis, George and Smola, Alex},
  journal={arXiv preprint arXiv:2302.00923},
  year={2023}
}
@article{eichenberg2021magma,
  title={MAGMA--Multimodal Augmentation of Generative Models through Adapter-based Finetuning},
  author={Eichenberg, Constantin and Black, Sidney and Weinbach, Samuel and Parcalabescu, Letitia and Frank, Anette},
  journal={arXiv preprint arXiv:2112.05253},
  year={2021}
}

@article{wang2022image,
  title={Image as a foreign language: Beit pretraining for all vision and vision-language tasks},
  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},
  journal={arXiv preprint arXiv:2208.10442},
  year={2022}
}

@article{ouyang2022training,
  title={Training language models to follow instructions with human feedback},
  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},
  journal={Advances in Neural Information Processing Systems},
  volume={35},
  pages={27730--27744},
  year={2022}
}

@article{nye2021show,
  title={Show your work: Scratchpads for intermediate computation with language models},
  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},
  journal={arXiv preprint arXiv:2112.00114},
  year={2021}
}

@article{wang2022self,
  title={Self-consistency improves chain of thought reasoning in language models},
  author={Wang, Xuezhi and Wei, Jason and Schuurmans, Dale and Le, Quoc and Chi, Ed and Zhou, Denny},
  journal={arXiv preprint arXiv:2203.11171},
  year={2022}
}

@article{zhang2022automatic,
  title={Automatic chain of thought prompting in large language models},
```

```
  author={Zhang, Zhuosheng and Zhang, Aston and Li, Mu and Smola, Alex},
  journal={arXiv preprint arXiv:2210.03493},
  year={2022}
}

@article{chen2021evaluating,
  title={Evaluating large language models trained on code},
  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},
  journal={arXiv preprint arXiv:2107.03374},
  year={2021}
}
% DATASET
@inproceedings{Schwenk2022AOKVQAAB,
  title={A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge},
  author={Dustin Schwenk and Apoorv Khandelwal and Christopher Clark and Kenneth Marino and Roozbeh Mottaghi},
  booktitle={European Conference on Computer Vision},
  year={2022}
}
@inproceedings{hudson2019gqa,
  title={Gqa: A new dataset for real-world visual reasoning and compositional question answering},
  author={Hudson, Drew A and Manning, Christopher D},
  booktitle={Proceedings of the IEEE/CVF conference on computer vision and pattern recognition},
  pages={6700--6709},
  year={2019}
}

@inproceedings{marino2019ok,
  title={Ok-vqa: A visual question answering benchmark requiring external knowledge},
  author={Marino, Kenneth and Rastegari, Mohammad and Farhadi, Ali and Mottaghi, Roozbeh},
  booktitle={Proceedings of the IEEE/cvf conference on computer vision and pattern recognition},
  pages={3195--3204},
  year={2019}
}

@inproceedings{thrush2022winoground,
  title={Winoground: Probing vision and language models for visio-linguistic compositionality},
  author={Thrush, Tristan and Jiang, Ryan and Bartolo, Max and Singh, Amanpreet and Williams, Adina and Kiela, Douwe and Ross, Candace},
  booktitle={Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition},
  pages={5238--5248},
  year={2022}
}

@inproceedings{goyal2017making,
  title={Making the v in vqa matter: Elevating the role of image understanding in visual question answering},
  author={Goyal, Yash and Khot, Tejas and Summers-Stay, Douglas and Batra, Dhruv and Parikh, Devi},
  booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},
  pages={6904--6913},
  year={2017}
}
@inproceedings{zhu2016visual7w,
  title={Visual7w: Grounded question answering in images},
```

```
  author={Zhu, Yuke and Groth, Oliver and Bernstein, Michael and Fei-Fei, Li},
  booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},
  pages={4995--5004},
  year={2016}
}
@inproceedings{johnson2017clevr,
  title={Clevr: A diagnostic dataset for compositional language and elementary visual reasoning},
  author={Johnson, Justin and Hariharan, Bharath and Van Der Maaten, Laurens and Fei-Fei, Li and Lawrence Zitnick, C and Girshick, Ross},
  booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},
  pages={2901--2910},
  year={2017}
}

% VL MODELS
@article{Zhang2022OPTOP,
  title={OPT: Open Pre-trained Transformer Language Models},
  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},
  journal={ArXiv},
  year={2022},
  volume={abs/2205.01068}
}
@article{Dosovitskiy2020AnII,
  title={An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale},
  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},
  journal={ArXiv},
  year={2020},
  volume={abs/2010.11929}
}
@article{tsimpoukelli2021multimodal,
  title={Multimodal few-shot learning with frozen language models},
  author={Tsimpoukelli, Maria and Menick, Jacob and Cabi, Serkan
          and Eslami, SM and Vinyals, Oriol and Hill, Felix},
  journal={Proc. Neural Information Processing Systems},
  year={2021}
}

@article{chen2022pali,
  title={Pali: A jointly-scaled multilingual language-image model},
  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},
  journal={arXiv preprint arXiv:2209.06794},
  year={2022}
}
@article{driess2023palm,
  title={Palm-e: An embodied multimodal language model},
  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},
  journal={arXiv preprint arXiv:2303.03378},
  year={2023}
```

```
}

@article{chung2022scaling,
  title={Scaling instruction-finetuned language models},
  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},
  journal={arXiv preprint arXiv:2210.11416},
  year={2022}
}

@article{tan2019lxmert,
  title={Lxmert: Learning cross-modality encoder representations from transformers},
  author={Tan, Hao and Bansal, Mohit},
  journal={arXiv preprint arXiv:1908.07490},
  year={2019}
}
@article{lu2019vilbert,
  title={Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks},
  author={Lu, Jiasen and Batra, Dhruv and Parikh, Devi and Lee, Stefan},
  journal={Advances in neural information processing systems},
  volume={32},
  year={2019}
}

@article{manas2022mapl,
  title={MAPL: Parameter-Efficient Adaptation of Unimodal Pre-Trained Models for Vision-Language Few-Shot Prompting},
  author={Ma{\~n}as, Oscar and Rodriguez, Pau and Ahmadi, Saba and Nematzadeh, Aida and Goyal, Yash and Agrawal, Aishwarya},
  journal={arXiv preprint arXiv:2210.07179},
  year={2022}
}

@article{alayrac2022flamingo,
  title={Flamingo: a visual language model for few-shot learning},
  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},
  journal={arXiv preprint arXiv:2204.14198},
  year={2022}
}

@article{zeng2021multi,
  title={Multi-grained vision language pre-training: Aligning texts with visual concepts},
  author={Zeng, Yan and Zhang, Xinsong and Li, Hang},
  journal={arXiv preprint arXiv:2111.08276},
  year={2021}
}

@article{li2023blip,
  title={Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models},
  author={Li, Junnan and Li, Dongxu and Savarese, Silvio and Hoi, Steven},
  journal={arXiv preprint arXiv:2301.12597},
  year={2023}
```

```
}

@inproceedings{li2022blip,
      title={BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation},
      author={Junnan Li and Dongxu Li and Caiming Xiong and Steven Hoi},
      year={2022},
      booktitle={ICML},
}

@inproceedings{herdade2019imagecap,
 author = {Herdade, Simao and Kappeler, Armin and Boakye, Kofi and Soares, Joao},
 booktitle = {Advances in Neural Information Processing Systems},
 editor = {H. Wallach and H. Larochelle and A. Beygelzimer and F. d\textquotesingle Alch\'{e}-Buc and E. Fox and R. Garnett},
 pages = {},
 publisher = {Curran Associates, Inc.},
 title = {Image Captioning: Transforming Objects into Words},
 url = {https://proceedings.neurips.cc/paper_files/paper/2019/file/680390c55bbd9ce416d1d69a9ab4760d-Paper.pdf},
 volume = {32},
 year = {2019}
}

@inproceedings{koh2023grounding,
  title={Grounding Language Models to Images for Multimodal Generation},
  author={Koh, Jing Yu and Salakhutdinov, Ruslan and Fried, Daniel},
  journal={arXiv:2301.13823},
  year={2023}
}

@inproceedings{anderson2018bottom,
  title={Bottom-up and top-down attention for image captioning and visual question answering},
  author={Anderson, Peter and He, Xiaodong and Buehler, Chris and Teney, Damien and Johnson, Mark and Gould, Stephen and Zhang, Lei},
  booktitle={Proceedings of the IEEE conference on computer vision and pattern recognition},
  pages={6077--6086},
  year={2018}
}

@inproceedings{xu2015show,
  title={Show, attend and tell: Neural image caption generation with visual attention},
  author={Xu, Kelvin and Ba, Jimmy and Kiros, Ryan and Cho, Kyunghyun and Courville, Aaron and Salakhudinov, Ruslan and Zemel, Rich and Bengio, Yoshua},
  booktitle={International conference on machine learning},
  pages={2048--2057},
  year={2015},
  organization={PMLR}
}
% ANALYSIS

@InProceedings{pmlr-v162-zhou22n,
  title =    {{VLUE}: A Multi-Task Multi-Dimension Benchmark for Evaluating Vision-Language Pre-training},
  author =       {Zhou, Wangchunshu and Zeng, Yan and Diao, Shizhe and Zhang, Xinsong},
  booktitle =    {Proceedings of the 39th International Conference on Machine Learning},
```

```
  pages =    {27395--27411},
  year =   {2022},
  editor =   {Chaudhuri, Kamalika and Jegelka, Stefanie and Song, Le and Szepesvari, Csaba and Niu, Gang and Sabato, Sivan},
  volume =   {162},
  series =   {Proceedings of Machine Learning Research},
  month =    {17--23 Jul},
  publisher =    {PMLR},
  pdf =    {https://proceedings.mlr.press/v162/zhou22n/zhou22n.pdf},
  url =    {https://proceedings.mlr.press/v162/zhou22n.html},
  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.}
}

@article{yuksekgonul2022and,
  title={When and why vision-language models behave like bag-of-words models, and what to do about it?},
  author={Yuksekgonul, Mert and Bianchi, Federico and Kalluri, Pratyusha and Jurafsky, Dan and Zou, James},
  journal={arXiv preprint arXiv:2210.01936},
  year={2022}
}
@article{diwan2022winoground,
  title={Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality},
  author={Diwan, Anuj and Berry, Layne and Choi, Eunsol and Harwath, David and Mahowald, Kyle},
  journal={arXiv preprint arXiv:2211.00768},
  year={2022}
}

@inproceedings{selvaraju2017grad,
  title={Grad-cam: Visual explanations from deep networks via gradient-based localization},
  author={Selvaraju, Ramprasaath R and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv},
  booktitle={Proceedings of the IEEE international conference on computer vision},
  pages={618--626},
  year={2017}
}

% TASK
@inproceedings{antol2015vqa,
  title={Vqa: Visual question answering},
  author={Antol, Stanislaw and Agrawal, Aishwarya and Lu, Jiasen and Mitchell, Margaret and Batra, Dhruv and Zitnick, C Lawrence and Parikh, Devi},
  booktitle={Proceedings of the IEEE international conference on computer vision},
```

```
  pages={2425--2433},
  year={2015}
}

@article{srivastava2022beyond,
  title={Beyond the imitation game: Quantifying and extrapolating the capabilities of language models},
  author={Srivastava, Aarohi and Rastogi, Abhinav and Rao, Abhishek and Shoeb, 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},
  journal={arXiv preprint arXiv:2206.04615},
  year={2022}
}

@article{zhang2019bertscore,
  title={Bertscore: Evaluating text generation with bert},
  author={Zhang, Tianyi and Kishore, Varsha and Wu, Felix and Weinberger, Kilian Q and Artzi, Yoav},
  journal={arXiv preprint arXiv:1904.09675},
  year={2019}
}

@inproceedings{radford2021learning,
  title={Learning transferable visual models from natural language supervision},
  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},
  booktitle={International conference on machine learning},
  pages={8748--8763},
  year={2021},
  organization={PMLR}
}


@inproceedings{shao2023prompting,
  title={Prompting large language models with answer heuristics for knowledge-based visual question answering},
  author={Shao, Zhenwei and Yu, Zhou and Wang, Meng and Yu, Jun},
  booktitle={Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition},
  pages={14974--14983},
  year={2023}
}

@article{lu2022unified,
  title={Unified-io: A unified model for vision, language, and multi-modal tasks},
  author={Lu, Jiasen and Clark, Christopher and Zellers, Rowan and Mottaghi, Roozbeh and Kembhavi, Aniruddha},
  journal={arXiv preprint arXiv:2206.08916},
  year={2022}
}

@inproceedings{guo2023images,
  title={From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models},
  author={Guo, Jiaxian and Li, Junnan and Li, Dongxu and Tiong, Anthony Meng Huat and Li, Boyang and Tao, Dacheng and Hoi, Steven},
  booktitle={Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition},
```

```
  pages={10867--10877},
  year={2023}
}

@article{peng2023kosmos,
  title={Kosmos-2: Grounding Multimodal Large Language Models to the World},
  author={Peng, Zhiliang and Wang, Wenhui and Dong, Li and Hao, Yaru and Huang, Shaohan and Ma, Shuming and Wei, Furu},
  journal={arXiv preprint arXiv:2306.14824},
  year={2023}
}

@article{awadalla2023openflamingo,
  title={Openflamingo: An open-source framework for training large autoregressive vision-language models},
  author={Awadalla, Anas and Gao, Irena and Gardner, Josh and Hessel, Jack and Hanafy, Yusuf and Zhu, Wanrong and Marathe, Kalyani and Bitton, Yonatan and Gadre, Samir and Sagawa, Shiori and others},
  journal={arXiv preprint arXiv:2308.01390},
  year={2023}
}

@article{liu2023visual,
  title={Visual instruction tuning},
  author={Liu, Haotian and Li, Chunyuan and Wu, Qingyang and Lee, Yong Jae},
  journal={arXiv preprint arXiv:2304.08485},
  year={2023}
}

@inproceedings{agrawal-etal-2023-reassessing,
    title = "Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization",
    author = "Agrawal, Aishwarya  and
      Kajic, Ivana  and
      Bugliarello, Emanuele  and
      Davoodi, Elnaz  and
      Gergely, Anita  and
      Blunsom, Phil  and
      Nematzadeh, Aida",
    editor = "Vlachos, Andreas  and
      Augenstein, Isabelle",
    booktitle = "Findings of the Association for Computational Linguistics: EACL 2023",
    month = may,
    year = "2023",
    address = "Dubrovnik, Croatia",
    publisher = "Association for Computational Linguistics",
    url = "https://aclanthology.org/2023.findings-eacl.90",
    doi = "10.18653/v1/2023.findings-eacl.90",
    pages = "1201--1226",
```

```
    abstract = "Vision-and-language (V{\&}L) models pretrained on large-scale multimodal data have demonstrated strong performance on various tasks such as image captioning and visual question answering (VQA). The quality of such models is commonly assessed by measuring their performance on unseen data that typically comes from the same distribution as the training data. However, when evaluated under out-of-distribution (out-of-dataset) settings for VQA, we observe that these models exhibit poor generalization. We comprehensively evaluate two pretrained V{\&}L models under different settings (i.e. classification and open-ended text generation) by conducting cross-dataset evaluations. We find that these models tend to learn to solve the benchmark, rather than learning the high-level skills required by the VQA task. We also find that in most cases generative models are less susceptible to shifts in data distribution compared to discriminative ones, and that multimodal pretraining is generally helpful for OOD generalization. Finally, we revisit assumptions underlying the use of automatic VQA evaluation metrics, and empirically show that their stringent nature repeatedly penalizes models for correct responses.",
}


@article{manas2023improving,
  title={Improving Automatic VQA Evaluation Using Large Language Models},
  author={Ma{\~n}as, Oscar and Krojer, Benno and Agrawal, Aishwarya},
  journal={arXiv preprint arXiv:2310.02567},
  year={2023}
}


@article{zhu2023minigpt,
  title={Minigpt-4: Enhancing vision-language understanding with advanced large language models},
  author={Zhu, Deyao and Chen, Jun and Shen, Xiaoqian and Li, Xiang and Elhoseiny, Mohamed},
  journal={arXiv preprint arXiv:2304.10592},
  year={2023}
}

@article{guo2022images,
  title={From images to textual prompts: Zero-shot vqa with frozen large language models},
  author={Guo, Jiaxian and Li, Junnan and Li, Dongxu and Tiong, Anthony Meng Huat and Li, Boyang and Tao, Dacheng and Hoi, Steven CH},
  journal={arXiv preprint arXiv:2212.10846},
  year={2022}
}

@article{qiao2022reasoning,
  title={Reasoning with language model prompting: A survey},
  author={Qiao, Shuofei and Ou, Yixin and Zhang, Ningyu and Chen, Xiang and Yao, Yunzhi and Deng, Shumin and Tan, Chuanqi and Huang, Fei and Chen, Huajun},
  journal={arXiv preprint arXiv:2212.09597},
  year={2022}
}



@article{agrawal2022rethinking,
  title={Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization},
  author={Agrawal, Aishwarya and Kaji{\'c}, Ivana and Bugliarello, Emanuele and Davoodi, Elnaz and Gergely, Anita and Blunsom, Phil and Nematzadeh, Aida},
  journal={arXiv preprint arXiv:2205.12191},
  year={2022}
}
```

```
@article{touvron2023llama,
  title={Llama 2: Open foundation and fine-tuned chat models},
  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},
  journal={arXiv preprint arXiv:2307.09288},
  year={2023}
}
```