如何写好一个 rebuttal

陶仁帅
软件开发环境国家重点实验室
北京航空航天大学
2022.01

明确目标

面向审稿人：

澄清疑问、纠正误解、反驳偏见

改进工作本身

面向AC：

让他了解审稿人的担忧是否得到了解决

反对恶意审查

一、从积极评价写起

首先突出审稿人的积极评价，提醒AC 审稿人对这篇文章的肯定。rebuttal的其余部分则主要集中回应消极方面，认真仔细地回答审稿人的问题。

We thank the reviewers for their thoughtful feedback. We are encouraged they found our motivation and idea to be strong, clear (R4) and novel (R1), and our analysis insightful (R4). We are glad they found our approach to be intuitive (R4), evaluated with extensive/adequate/convincing experiments (R2, R4), and compared against appropriate baselines (R2) achieving significant improvements (R2, R4) while reusing existing data (R2). We are pleased R1 recognizes the importance of training VQA models to make more human-like decisions and that R4 recognizes the superiority of our gradient-based approach over model attention. We address reviewer comments below and will incorporate all feedback.

We thank reviewers for their insightful and positive feedback! We are encouraged that they find EmbodiedQA to be a novel task (R1, 2, 3), an important research problem (R1, 2), appropriately positioned w.r.t. prior work (R1, 3), the dataset thoughtfully created to avoid biases (R3) and of value to the community (R1, 2, 3), and the proposed methods reasonable (R3) and elegant (R2). One primary concern was insufficient discussion of results. We agree. We were constrained by space. We answer some specific questions below, but will incorporate all feedback in the final version.

Scores: All reviewers recommended weak accept.

We are encouraged that reviewers find our model interesting (R1), simple and highly effective (R2), and a strong case for the benefits of multi-task learning (R3). Moreover, R2 thinks our method opens doors for the creation of smaller, more focused, vision-language tasks. Reviewers found our experiments are carefully designed (R1), sound, thorough, and backed with sufficient ablations (R2).

We are pleased reviewers identified our contributions beyond just performance gains on many tasks. We design a clean multi-task V&L setting (R1) and our analysis of the overlaps and interactions between task groups adds insight (R2,R3). Our code release will unify many V&L tasks in a single framework – allowing future work to easily explore transfer and multi-task settings (R2).

二、直接回复审稿人的问题

简洁、完整地引用审稿人的问题或关注的核心。在表达自己的观点之前，先直接回应问题，再给出细节、描述背景或解释你的立场。可以深入解释，但需要先给出 Yes、No 或者 Not quite。

「是否在现实环境中进行评估？」

「我们不同意问题的前提。虽然这些环境是模拟的，但它们非常逼真。」

「为什么不与 GMAP 进行比较？」

「GMAP 在我们的环境中成本比较高。我们的环境有.....」

三、口语化

如下图所示，内容风格是相对口语化的，这样更容易理解，也不会被认为有明显的对立气场。

[R3] Given the use of Conceptual Captions (CC), are the comparisons to baselines fair? We believe these comparisons are fair. We agree that CC is a large, additional data source; however, being able to leverage this additional data for a diverse range of vision and language tasks is precisely our contribution! Existing approaches to vision and language tasks are simply not designed to do so – for instance, it is unclear how to train a standard VQA model like BAN with CC captioning data. Arguing from analogy, the widespread transfer of deep models pretrained on ImageNet also leveraged more data during pretraining; however, we do not find it unfair to pre-deep learning approaches that were not equipped to leverage that data. Finally, note that unlike ImageNet, CC is webly supervised, and did not involve expensive human annotation. We acknowledge that in caption-based image retrieval, CC data could have been used to pretrain existing work for a more direct architectural comparison – we will address.

四、回应问题的核心指向

不要纠结于讨论引用的关注点，主要解决审稿意见的指向问题。例如「你为什么不对 GLORP3 进行评估？」通常这个意思是质疑你的实验，回答问题，然后指出你已经对 X、Y 和 Z 进行了评估，这应该就足够了。请注意，提醒其他 Reviewer和AC 你进行了广泛的实验评估很有用，否则他们乍一看审稿人的评论，可能会留下错误的印象。

@R1 – Verification through human-study not scalable: Our evaluation is not entirely based on human studies. In fact, most of our evaluation is quantitative – see Section 5 and 6 where we quantitatively evaluate task performance and grounding, both of which show the effectiveness of HINT without requiring human studies. We conduct human studies just to evaluate whether HINTed models are more trustworthy to humans than base models.

五、不要害怕强调某一点

「表 4 中的第 2 行正好说明了这一点。」

「我们在测试时不需要 human-in-the-loop。」

请注意，类似这样的许多回复不仅直接，而且还能突出重点。

@R3, try simpler navigation (without ACT): R3's suggestion is exactly our LSTM+Q baseline reported in paper. As stated in L779-781, LSTM+Q vs. ACT+Q establishes benefit of our proposed model over a simple LSTM baseline. Both have identical inputs/outputs and are trained on shortest path navigation, so the performance improvement is solely from change in architecture. The distance to target reward shaping is only for ACT+Q-RL, not other models.

六、申明论文中本就包含所需细节

也就是说，如果审稿人想要的内容已经在论文中，请说出来，给出行 / 表 / 图的编号。

引用论文内容是为了向所有 Reviewer和AC 赢得信任：研究本身并不缺少重要细节。

（倒不一定要让 RAC 回去查看论文）

> *It would also be helpful if detailed experiment settings are detailed, e.g. GPU characteristics, DDPPO's hyperparameters, etc.*

GPU characteristics are detailed in both sections 5 and 6 -- we use Titan V100 GPUs and NCCL2.4.7 with Infiniband interconnect.

As described in Section 5, DD-PPO introduces a single additional hyperparameter, the preemption threshold. We study this hyper-parameter in section 5 -- figure 4 on page 5 shows its effect and figure 6 on page 13 provides a further breakdown. We find that under values of 60% and 80%, DD-PPO scales near-linearly under both heterogenous and homogenous workloads.

七、整理共同的关注点

如果多个审稿人提了相关的问题，可以一起回复，从而节约空间。

R1,R3 Can self-supervised pretraining be skipped given the large amount of data in the multi-task setting? This is an exciting experiment that we have not investigated! We began with pretrained ViLBERT in order to start from a near SOTA trunk model. It may be that training under multi-task supervision provides enough information that large-scale self-supervision is not needed. We will try this and report results in the camera ready as we cannot report it here.

八、承诺的都要做到

1. 不要光说「We will discuss Singh et al. in the paper.」，而是直接在 rebuttal 说明白；

2. 不要光说「We will explain what XXX stands for in the paper」，而是在 rebuttal 中解释它代表什么。补充一点，如果你表示将会把这些添加到论文中，Review和AC 会更容易相信你。

@R1, dialog-level evaluation: Thanks for the suggestion! Using Recall@5 to define round-level 'success', our best discriminative model MN-QIH-D gets 7.01 rounds out of 10 correct, while generative MN-QIH-G gets 5.37. Further, the mean first-failure-round (under $R@5$) for MN-QIH-D is 3.23, and 2.39 for MN-QIH-G. Fig. 1c and Fig. 1d show plots for all values of k in $R@k$. We will add this analysis.

九、解决审稿人的问题存在客观困难，如实告知

审稿人表示要进行额外的实验，但条件不允许？那就如实告知。他们询问有关趋势的判断，但你没有任何判断？那就直接说你思考过，但没有什么好的想法，之后会继续研究。没有足够的 GPU 算力来运行他们要求的实验？也可以坦白说出来。

R1,R3 Can self-supervised pretraining be skipped given the large amount of data in the multi-task setting? This is an exciting experiment that we have not investigated! We began with pretrained ViLBERT in order to start from a near SOTA trunk model. It may be that training under multi-task supervision provides enough information that large-scale self-supervision is not needed. We will try this and report results in the camera ready as we cannot report it here.

十、和审稿人争论

在某些情况下，审稿人可能没有认真对待自己的角色。确保其他 Reviewer和AC 意识到这一点，减少这些审稿人评论的影响程度。指出不合理或未经证实的评论，参考其他意见相左的审稿人评论，都会有所帮助，此外还可以包括向 AC 提交未公开的评论。

@R1 – "State of the art and beyond is moving away from human guided approaches.": Without any citations, we find this difficult to respond to. R1 states that localization is already done in wholly unsupervised ways, but it is unclear what exactly is being referred to here. Weakly supervised approaches for object localization do exist, but their performance is still significantly worse compared to fully supervised approaches. While approaches without human attention

R3 Why not remove the 'Verification (G4)' task group from the paper? We wholeheartedly disagree with R3's assumption that G4 is only included to "increase the number of tasks" and should be dropped due to its negative interactions. Part of our scientific contribution is studying interactions between different tasks. Beyond transparency (which itself warrants keeping negative results), these results also provided useful information to the community. Will include further ablations in the setting 'AT w/o G4' in camera ready.

十一、其他

1. 先写那些比较好回答的问题，再写不太明确和次要的问题。
2. 注意上图中彩色代码标注审稿人的小技巧，让对应的审稿人尽可能轻松地发现与其相关的回复——即使内容是合并的或者未按审稿人的顺序排列。
3. 与其和 Reviewer、AC 争论，不如向他们提供统计数据来支撑观点，或者用额外的实验结果来回应他们的担忧（如果条件允许）。每次发现自己与审稿人有不同的意见，问问自己是否可以用数据来证明这一点，因为这样可以提供直观的论据。
4. 如果审稿人给出了了建设性意见、错别字列表、相关研究指导、关于未来工作的思考，都要表示感谢。至少写一个简短的介绍，表达自己的谢意。

参考资料：https://deviparikh.medium.com/how-we-write-rebuttals-dc84742fece1、机器之心

软件开发环境国家重点实验室

State Key Laboratory of Software Development Environment

TNANKS

感谢聆听

陶仁帅

软件开发环境国家重点实验室

北京航空航天大学

联系邮箱： rstao@buaa.edu.cn

个人主页： https://rstao95.github.io