* [Submissions](https://cmt3.research.microsoft.com/ACCV2018/Submission/Index)
* **Select Your Role :**
* [**Author**](https://cmt3.research.microsoft.com/ACCV2018/Submission/ReviewsSnapshot/702)
* [ACCV2018](https://cmt3.research.microsoft.com/ACCV2018/Submission/ReviewsSnapshot/702)
* [Oliver Schulte](https://cmt3.research.microsoft.com/ACCV2018/Submission/ReviewsSnapshot/702)

 Print

**View Reviews**

Paper ID

702

Paper Title

Multimodal Spatial Dynamic Memory Networks for Visual Question Answering

**Reviewer #1**

**Questions**

* **1. [Summary] Please provide a short summary of the paper and its contributions.** 
  + The paper describes an approach to Visual Question Answering based on spatially defined attention. In particular, the paper introduces a bidirectional Grid GRU and a spatial episodic memory. Captions are obtained by the Visual Genome dataset and experiments have been done on COCO-VQA and Visual 7W dataset.
* **2. [Paper Strengths] Please discuss the positive aspects of the paper. Be sure to comment on the paper's novelty, technical correctness, clarity and experimental evaluation. Notice that different papers may need different levels of evaluation: an algorithm paper may need fewer experiments, while an application paper may require thorough comparisons to existing methods. Also, please be sure to justify your comments in detail. For example, if you think the work is novel, not only say so, but also explain in detail why you think this is the case.** 
  + The paper provides very nice insights into all problems related to where and how captions should collect their semantics on an image, especially highlighting the need to account for spatial attention leading to a good focus on relevant regions in the image.   
    Accordingly, the authors introduce a grid GRU that can support spatial attention and spatial captioning on attention highlighted regions.  
    In particular, the authors underline the role of spatial regions to provide an adequate Dynamic memory network.  
    The paper is clear to understand and well organized. The state of the art misses some relevant recent papers, otherwise is well organized.  
    Experiments have been done on COCO-VQA and Visual7W comparing with a number of baseline methods. Experiments are well devised the only problem is that they miss recent important results.

The state of the art misses some relevant recent papers, otherwise is well organized. Really?

* **3. [Paper Weaknesses] Please discuss the negative aspects of the paper: lack of novelty or clarity, technical errors, insufficient experimental evaluation, etc. Justify your comments in detail; don't provide just generic critique. It is not reasonable to ask for comparisons with unpublished, non-peer-reviewed papers (eg. arXiv), or papers published after the ACCV 2018 deadline.** 
  + Despite a deep insight on all problems related to the spatial semantics of VQA the paper does not seem to effectively introduce new models in terms of Attention (Bahdanau, Luong), nor in terms of Episodic Memory.   
    The experiments do not evaluate caption quality, for example using standard automatic evaluation metrics, namely SPICE , CIDEr , METEOR , ROUGE-L and BLEU.  
    Another problem related to the experiments is the fact that the authors do not consider recently published results such as “ Beyond Bilinear: Generalized Multi-modal Factorized High-order Pooling for Visual Question Answering” and “Bottom-Up and Top-Down Attention for Image Captioning  
    and Visual Question Answering” providing comparable or better results.

The titles sound relevant. Can we say something like “these orthogonal methods can be added to our architecture”? It’s strange how they think there should be just one idea when clearly it takes several orthogonal ones.  
  
Minor problems:  
1. Line 326-328 what is the size of D? Why using here a one-shot encoding?  
2. Better explain line 332-335  
3. What is the range of confidence score?  
4. Line 337 textual features or textual fact?  
5. Lines 339-341 explain  
6. Line 383: where M\_1 and M\_2 come from? In what sense are trainable do you build a feedforward network to train them? Please give details.  
7. Optimization and loss function should be given.

* **4. [ Rebuttal Requests] Please note specific points that you would like the authors to address in their rebuttal.**
  + I think that comparison with recent results, as mentioned above, should be provided.  
    Likewise, a table with the caption accuracy should be provided.

Can we do this? We should also emphasize that we are *not* proposing a method to generate captions. He may have misunderstood this.

Also please answer the minor questions highlighted at point 3.

* **5. [Recommendation Justification / Detailed Comments ] Please explain to the AC, your fellow reviewers, and the authors your current recommendation for the paper. This explanation may include how you weigh the importance of the various strengths and weaknesses you described above.**
  + The paper provides interesting insights into the VQA models. Accordingly, it introduces the bidirectional grid GRU which is a novel model, despite bidirectional grid LSTM have been already introduced. There are some minor problems essentially due to the fact that the authors do not confront with recent results. Despite this, if the authors accept to review their paper in this direction, the paper is worth to be accepted.
* **6. [Overall Rating]**
  + Borderline

**Reviewer #2**

**Questions**

* **1. [Summary] Please provide a short summary of the paper and its contributions.** 
  + This paper proposes a model based on the memory network for the image question answering. It uses the grid LSTM to encode the spatial context both vertically and horizontally. Also, the model stores the captions of the detected boxes in memory. The proposed method is evaluated on three datasets.
* **2. [Paper Strengths] Please discuss the positive aspects of the paper. Be sure to comment on the paper's novelty, technical correctness, clarity and experimental evaluation. Notice that different papers may need different levels of evaluation: an algorithm paper may need fewer experiments, while an application paper may require thorough comparisons to existing methods. Also, please be sure to justify your comments in detail. For example, if you think the work is novel, not only say so, but also explain in detail why you think this is the case.** 
  + The paper is well written and easy to follow.
* **3. [Paper Weaknesses] Please discuss the negative aspects of the paper: lack of novelty or clarity, technical errors, insufficient experimental evaluation, etc. Justify your comments in detail; don't provide just generic critique. It is not reasonable to ask for comparisons with unpublished, non-peer-reviewed papers (eg. arXiv), or papers published after the ACCV 2018 deadline.** 
  + 1. The idea of using bounding box captions to help answer questions is investigated in the following paper before. Please explain the differences with your proposed method of using captioning in VQA problem?  
      
    Wu, Qi, et al. "Ask me anything: Free-form visual question answering based on knowledge from external sources." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

I hope this is easy to do.

2. Figure 3 is too complicated and not self-explanatory.  
  
3. The visualization in Figure 4 is in which dataset?  
  
4. The full model in Table 2 corresponds to which model? "MS-DMN+ResNet+MLB+sp"? or "MS-DMN-sp" ? Before I would think "MS-DMN-sp" is your full model.

I remember asking the same question. I thought you had answered it.

5. In SHAPE dataset, why the result "MS-DMN-sp" is not shown? Because it corresponds to your idea of using captions, this result will illustrate whether this idea is useful in this dataset.

Should be easy to add if we have the result

* **4. [ Rebuttal Requests] Please note specific points that you would like the authors to address in their rebuttal.**
  + See the paper weakness.
* **5. [Recommendation Justification / Detailed Comments ] Please explain to the AC, your fellow reviewers, and the authors your current recommendation for the paper. This explanation may include how you weigh the importance of the various strengths and weaknesses you described above.**
  + Weak accept.
* **6. [Overall Rating]**
  + Weak Accept

**Reviewer #3**

**Questions**

* **1. [Summary] Please provide a short summary of the paper and its contributions.** 
  + This submission addresses the visual question answering (VQA) problem by proposing a new framework. Compared to previous methods, the authors adopt bidirectional grid GRUs to encode 2D spatial context information. Besides, the authors leverage transfer learned region-grounded image captions to further boost performance. Experiments show that the proposed framework achieve good results on two datasets (COCO-VQA and Visual7W).
* **2. [Paper Strengths] Please discuss the positive aspects of the paper. Be sure to comment on the paper's novelty, technical correctness, clarity and experimental evaluation. Notice that different papers may need different levels of evaluation: an algorithm paper may need fewer experiments, while an application paper may require thorough comparisons to existing methods. Also, please be sure to justify your comments in detail. For example, if you think the work is novel, not only say so, but also explain in detail why you think this is the case.** 
  + (1) The authors adopt bidirectional grid GRUs to encode 2D spatial context information.  
      
    (2) The authors leverage transfer learned region-grounded image captions to further boost performance.
* **3. [Paper Weaknesses] Please discuss the negative aspects of the paper: lack of novelty or clarity, technical errors, insufficient experimental evaluation, etc. Justify your comments in detail; don't provide just generic critique. It is not reasonable to ask for comparisons with unpublished, non-peer-reviewed papers (eg. arXiv), or papers published after the ACCV 2018 deadline.** 
  + (1) Proofreading is required to further improve the writing quality. There are many typos in the submission which prevents readers to understand the details of the proposed framework. For example, the symbol is missing due to compiling errors in line 303-306, line 406.

oops

(2) Lack of novelty. The proposed bidirectional grid GRU unit is just based on previous grid GRU (from paper "grid long short-term memory") and modify it as bidirectional version. Besides, due to lack of ablation study, it is unknown whether bidirectional grid GPU performs better than grid GPU or not.

I thought we had an ablation study. Also it’s not our only idea, as he says we’ve also got the captions.

(3) Unfair comparison: For the second novelty (introducing transfer learned region-grounded image captions), it should be fair to compare with baselines which also incorporates with this external information. Without ablation study, it is unknown whether the proposed system performance increases more than other systems' performances when incorporate such information.

I thought we had an ablation study.

(4) Not state-of-the-art performance as claimed in the submission (line 29). In both tables 1, 2, the full model's performance is far lower than state-of-the-art performances. On COCO-VQA, even results in 2017 performance better than 67.4% in test set (http://www.visualqa.org/roe.html, higher than 70% for method "Visual Question Answering as a Meta Learning Task"). On Visual7W, in 2016, the performance has already reached 68.5% mAP in "Revisiting Visual Question Answering Baselines" (ECCV2016)).

I hope he’s misreading this

* **4. [ Rebuttal Requests] Please note specific points that you would like the authors to address in their rebuttal.**
  + Please check the "paper weaknesses" part above.
* **5. [Recommendation Justification / Detailed Comments ] Please explain to the AC, your fellow reviewers, and the authors your current recommendation for the paper. This explanation may include how you weigh the importance of the various strengths and weaknesses you described above.**
  + Considering the paper's weaknesses (proofreading, lack of novelty, unfair comparison, far behind state-of-the-art), I would suggest authors to modify the submission for future conferences.
* **6. [Overall Rating]**
  + Weak Reject

[Go Back](https://cmt3.research.microsoft.com/ACCV2018/Submission/Index)

© 2018 Microsoft Corporation [About CMT](https://cmt3.research.microsoft.com/Content/CMT.html" \t "_blank) | [Terms of Use](https://www.microsoft.com/en-us/servicesagreement/" \t "_blank) | [Privacy & Cookies](https://go.microsoft.com/fwlink/?LinkId=521839" \t "_blank) | [Request CMT Site](https://cmt3.research.microsoft.com/CMTSRM/" \t "_blank) [CMT Support](mailto:support@msr-cmt.org)