**Reviewer #1**

**Questions**

**1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.**

* The paper proposes a method for transfer learning in visual and reasoning tasks, the specific tasks tackled in the paper are image and video question answering. A end-to-end recurrent SAMNet architecture is proposed comprising of memory unit, memory and visual retrieval units (i.e. attention mechanisms) and reasoning+update units for question-answering. The paper identifies different transfer learning settings of feature, temporal and reasoning transfer and demonstrates effective transfer learning in these setups with experiments on CLEVR and COG dataset.

**2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.**

* + The paper is well written, overview of the proposed architecture is easy to understand guided by an illustration in Fig. 2. The experiments section has important details on dataset used, their splits and different settings of ablation studies.   
    
  + Different transfer learning setups in the context of video and image/visual question answering are investigated, e.g. temporal transfer is pre-trained on short videos and then finetuned on longer videos, while reasoning transfer is pre-trained on one or subset of tasks and finetuned on others.   
    
  + Results in Fig. 4, 6-8 show good performance of the proposed SAMNet architecture in different setups on CLEVR and COG datasets  
    
  + Literature on transfer learning in context of vision language tasks is limited, therefore the method contributes an important benchmark and a novel setup for vqa transfer learning.

**3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).**

* - Some important ablation studies if added can help understand the effect of different modules of the architecture. For example, how does unrolling to different steps affect performance? How does change in the number of locations or size of memory unit affect performance?   
    
  - Though the overview of the architecture is clear, the details presented are not enough to reproduce the method and experiments in the future. The authors should consider making the code and their dataset splits available for future research.  
    
  - Performance of baseline method is demonstrated for temporal transfer, it would help if qa baselines are added for the reasoning and feature transfer setup as well.

**4. [Overall rating] Paper rating (pre-rebuttal)**

* Weak accept

**5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.**

* The proposed method studies different transfer learning setup in the context of VQA tasks and demonstrates good empirical performance. I'm positive about the technical contributions of this work and recommend accept.

**Reviewer #2**

**Questions**

**1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.**

* The submission works on a challenging visual recognition problem represented, here, by two datasets CLEVR and COG. Both are synthetic datasets that are believed the high-level reasoning is necessary to solve them. The submission introduces the Selective Attention Memory (SAM) Network to handle the problem on both datasets. The architecture appears novel (though it is not fully clear to me). The architecture is inspired by a broader work on using the memory in neural network architecture, e.g., by [Graves et al. "Neural Turing Machine"]. To handle the problem, it uses selective attention over the visual input together with memories to derive an answer. Finally, the submission shows interesting results on CLEVR and SOTA on COG.

**2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.**

* + SOTA on COG  
  + The architecture seems to be novel (though I am not fully sure). It is definitely interesting.  
  + Method is evaluated on visual reasoning tasks, which are arguably less common in the Computer Vision community, but arguably should not.

**3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).**

* The paper is not well written. The first part of the paper defines transfer learning and next moves to suitable refinements to capture more interesting aspects of visual reasoning. However, neither definition nor the refinements seem to be used in the other parts of the paper. That is, after reading the first 3 pages I would expect that the authors define an interesting (theoretical) framework for the transfer, and next use it. However, one can clearly see that even the language changes, e.g., later 'tasks' are used instead of the 3 transfer formulations. This space could be better used to introduce datasets, e.g., what kind of questions and examples are used in the COG dataset, and, perhaps more important, to better describe the method, or to explain results.  
    
  To me, the description of the method is very unclear. It should start high-level, how different modules interface each other, and what the motivation is for each such a module. Then it should define standard parts such as attention over questions, and introduce all the variables that appear from nowhere. I would also like to see a more clear Figure 2 depicting the architecture. There is very little correspondence between that figure and the equations used in the paper.  
    
  I would also like to see if the method is indeed retrieving objects from the frames correctly, and uses the memory in an expected way. This is something that I would expect on the paper that runs experiments on synthetic data where more aspects are under the control.

**4. [Overall rating] Paper rating (pre-rebuttal)**

* Weak reject

**5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.**

* I am inclining to reject the paper based on unclear and sloppy writing. However, I will give a chance to the authors for the rebuttal.

**6. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestions to make the submission stronger).**

* A few specific questions.  
    
  1. What are vo\_tt and mo\_t and how they differ (lines 400-403)? What is the difference between both objects?  
  2. How the temporal classifier \tau\_t is trained and used? Do you use its logits or classes directly (lines 389-397)?  
  3. What are va\_t and in general all other symbols. How they are computed.  
  4. What is pseudo-attention (line 453)? Why such a name?  
  5. What are reasoning operations (lines 389-391)?  
  6. How many 'reasoning operations' (what is k)?  
  7. Figure 4 shows results on CLEVR-CoGenT. CLEVR-CoGen uses a transfer between objects of type A and objects of type B (e.g. different combination of shapes and colors). Is this used here? Or this is standard CLEVR results? What are the results on the regular CLEVR, is it 95% from the supp. material?  
  8. Why the paper doesn't compare to other methoods on CLEVR?

**Reviewer #3**

**Questions**

**1. [Summary] In 3-5 sentences, describe the key ideas, experiments, and their significance.**

* The authors propose SAMNet, a video visual question answering technique that performs frame-by-frame temporal processing approach that can handle varying length videos and is meant to transfer better to new tasks. The authors describe various transfer learning settings and evaluate their approach in these settings on the COG video QA dataset and the CLEVR dataset.

**2. [Strengths] What are the strengths of the paper? Clearly explain why these aspects of the paper are valuable.**

* - the authors propose a model called SAMNet that tackles the visual question answer task   
    
  - the authors perform a number of experiments that demonstrate that SAMNet outperforms a set of baselines

**3. [Weaknesses] What are the weaknesses of the paper? Clearly explain why these aspects of the paper are weak. Please make the comments very concrete based on facts (e.g. list relevant citations if you feel the ideas are not novel).**

* - neither the abstract, introduction, nor the related work sections identify clear contributions; what makes SAMNet novel over other nets out there? What was missing in previous work that SAMNet has that causes it to have better performance? Why is it a good thing that a new taxonomy is proposed?  
    
  - the paper lacks motivation throughout. For example, what was the point of section 3? It seems like a set of definitions, but they are abstract and it's unclear what a reader is to take away from the section. In Section 4 the authors describe SAMNet, but the argument was never made what was missing in existing networks that requires a new network to be proposed. Is the problem with previous approaches that they have limited memory, and this is what motivates SAMNet? Is it that the number of addresses can be changed between training and testing (and why is that a good thing)?

**4. [Overall rating] Paper rating (pre-rebuttal)**

* Weak reject

**5. [Justification of rating] Please explain how the strengths and weaknesses aforementioned were weighed in for the rating. Please also mention what you expect to see from the rebuttal that may change your rating.**

* While the authors are tackling a challenging problem and seem to show reasonable quantitative experimental results, the paper is very difficult to read and requires much stronger motivation throughout. As it is, many details are described, but it's unclear which of the details are important and why various decisions were made. The authors do not clearly articulate what exactly is the main novel technical contribution in SAMNet that allows it to outperform the baselines in the experiments.

**6. [Detailed comments] Additional comments regarding the paper (e.g. typos, any suggestions to make the submission stronger).**

* The paper probably needs to be rewritten to clearly articulate the gaps in existing work and the main novel contributions of the current work that fills those gaps. As it is, the paper reads like a set of technical details, and it is up to the reader to guess why certain details might have been selected, and which details are the most important ones.