ICCV2015
International Conference on Computer Vision 2015
December 11-18, Santiago, Chile

Reviews For Paper

Paper ID	1344
Title	Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks

Masked Reviewer ID: Assigned_Reviewer_13

Review:

Question	
Paper Summary. Please summarize in your own words what the paper is about.	The paper presents a visual attention mechanism of convNet by introducing binary feedback neurons followed by each convolutional layer. The binary feedback neurons are activated to indicate the effect of the neurons at each spatial location and channel to the class score. Following the motivation, a sparse set of binary feedback neurons that maximizes the class score are activated. The feedback neurons are defined for each hidden neuron of the convolutional layer and are obtained via gradient ascent. The paper presents a visualization of visual attention when conditioned on class label by computing the gradient of the score w.r.t. the original image. In addition, the model is evaluated on weakly supervised object localization tasks on imageNet 2014 localization database, and demonstrated improvement to the baseline CNN models.
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: a theoretical paper may need no experiments, while a paper presenting a new approach to a known problem may require thorough comparisons to existing methods. Also, please make sure to justify your comments in	I think the paper presents a novel method for analyzing the convNet by introducing the feedback neurons. The paper demonstrates strong empirical performance both in quantitatively and qualitatively. The visualization results look interesting as they can indicate the discriminative region in the image. In addition, the performance on weakly supervised object localization is significantly improved from the baseline method without feedback neurons, which is quite encouraging.

great detail. For example, if you think the paper is novel, not only say so, but also explain in detail why you think this is the case.	
Paper Weaknesses. Please discuss the negative aspects of the paper: lack of novelty or clarity, technical errors, insufficient experimental evaluation, etc. Please justify your comments in great detail. If you think the paper is not novel, explain why and give a reference to prior work. Do not ask the authors to cite your own work. If you think this is essential, write it in the confidential comments to the AC. If you think there is an error in the paper, explain in detail why it is an error. If you think the experimental evaluation is insufficient, remember that theoretical results/ideas are essential to ICCV and that a theoretical paper need not have experiments. It is not ok to reject a paper because it did not outperform other existing algorithms,	Overall, I think the paper requires some revision as in the current version it lacks some important descriptions of the model, which significantly limit the understanding of the paper. - The operation at the feedback layer, i.e., s_{k}(I,z), is not described. Do feedback neurons z change the hidden neuron activations of the higher layers (e.g., ReLU) during the optimization process described in Section 3.4.? - It is unclear to me whether the feedback neurons are used to optimize weights w from the pre-trained models. - Please describe how to obtain T_{k}(z)? - What is the difference between the first and second rows of Figure 5? - Missing citation: Sohn et al., Learning and Selecting Features Jointly with Point-wise Gated Boltzmann Machines, ICML 2013 - they introduced binary switch variables to gate the , which is similar - minor comments: line 403: ajusting -> adjusting line 481: how do we use class label as prior for deconv method? be consistent between s_{k} and s_{c} in section 3.4. be consistent in using GoogLeNet (Googlenet, GoogleNet). the naming convention for z should be consistent; currently it is used in many different names, such as activation variable (line 365), feedback neurons (line 403), hidden neuron activation (line 439).

especially if the theory is novel and interesting. It is also not ok to ask for comparisons with unpublished papers and papers published after the ICCV deadline. Last but not least, remember to be polite and constructive.	
Preliminary Rating. Please rate the paper according to the following choices. Oral: these are papers whose quality is in the top 10% of the papers at ICCV. Examples include a theoretical breakthrough with no experiments; an interesting solution to a new problem; a novel solution to an existing problem with solid experiments; or an incremental paper that leads to dramatic improvements in performance. Oral/Poster: these are very strong papers, which may have one weakness that makes you unsure as to whether they should be oral or poster. Poster: these are strong papers, which have more than one weakness. For example, a well-	Poster

written paper with solid experiments, but incremental; a paper on a well studied problem with solid theory, but weak experiments; or a novel paper with good experiments, but poorly written. Weak Reject: these are papers that have some promise, but they would be better off by being revised and resubmitted. Strong Reject: these are papers that have major flaws, or have been done before.	
Preliminary Evaluation. Please indicate to the AC, your fellow reviewers, and the authors your current opinion on the paper. Please summarize the key things you would like the authors to include in their rebuttals to facilitate your decision making. There is no need to summarize the paper.	The paper presents a novel method for visually analyzing the convNet using feedback neurons. The paper requires significant polishing in writing and currently is lacking some important information as described in 3. Nonetheless, the paper demonstrated its effectiveness in empirical evaluation both in qualitative and quantitative manners.
Confidence. Write "Very Confident" to stress that you are absolutely sure about your conclusions (e.g., you are an expert who works in the paper's area), "Confident" to stress that you are mostly sure about your	

conclusions (e.g., you are not an expert but can distinguish good work from bad work in that area), and "Not Confident" to stress that that you feel some doubt about your conclusions. In the latter case, please provide details as confidential comments to PC/AC chairs (point 7.).	Confident

Masked Reviewer ID: Assigned_Reviewer_14

Review:

Question	
Paper Summary. Please summarize in your own words what the paper is about.	The paper proposes to use top down feedback for selectively identifying the image/feature activation that maximally effect the the final output of the ConvNet. This method is then used for performing weakly supervised object localization.
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: a theoretical paper may need no experiments, while a paper presenting a new approach to a known problem may require thorough comparisons to existing methods. Also, please make	1. A simple method for determining the features that maximally effect the ConvNet output. The proposed method is superior to previous visualization methods in the sense that it produces a class specific saliency map instead of a generic object saliency map. 2. Good results on object localization. 3. Some intuition on features used by ConvNet for fine-grained recognition.

sure to justify your comments in great detail. For example, if you think the paper is novel, not only say so, but also explain in detail why you think this is the case.	
Paper Weaknesses. Please discuss the negative aspects of the paper: lack of novelty or clarity, technical errors, insufficient experimental evaluation, etc. Please justify your comments in great detail. If you think the paper is not novel, explain why and give a reference to prior work. Do not ask the authors to cite your own work. If you think this is essential, write it in the confidential comments to the AC. If you think there is an error in the paper, explain in detail why it is an error. If you think the experimental evaluation is insufficient, remember that theoretical results/ideas are essential to ICCV and that a theoretical paper need not have experiments. It is not ok to reject a paper because it did not outperform other	1.Authors cite a very related paper - "Deep Networks with Internal Selective Attention through Feedback Connections" (dasNet), but a discussion comparing their method with dasNet is missing. The paper under review and dasNet are similar in how they modulate attention, with the key difference that dasNet uses attention for modifying its output, whereas the current paper uses attention as a post-processing to obtain class specific saliency map. The dasNet formalism is quite general and the formalism presentd in this work is subsumed by the dasNet paper. The dasNet only had results on CIFAR - however, if it was trained on Imagenet - one would expect similar results to those presented in this paper. I would like to know if the authors attempted to train dasNet on Imagenet. If yes, then what happened and if not then why so. 2. If there are k classes in the image - authors present saliency visualization results for maximizing the scores for these specific k classes. How would the saliency results look if the score is maximized with respect to an object class not present in the image? For instance, if the image has no bicycles - and one tries to produce a saliency map wrt to a bicycle - would it lead to an almost zero saliency map or would it highlight some other unrelated object? It would be good to have some of these visualizations. 3. Line 685, authors mention that there are 20,000 images in the validation set. To best of my knowledge there are 50,000 images in the validation set. Can the authors please clarify this discrepancy? 4. I feel the vision community would take this paper more seriously, if the authors provided results on the detection benchmark in addition to the localization. There are numerous methods for weakly supervised detection (especially on PASCAL VOC 2007) and it would be great to see the performance of the proposed method under that setting. 5. The authors make a great deal about that their method can be used to understand representations in the ConvNet - but the visualization results only show that the proposed method of saliency is class specific. One more avenue that could be potentially interesting is to identify for example the group of

existing algorithms, especially if the theory is novel and interesting. It is also not ok to ask for comparisons with unpublished papers and papers published after the ICCV deadline. Last but not least, remember to be polite and constructive.	filters in a particular layer that cause the classification of the image into a specific object category. Then, one could use the same method to probe representations of these individual filters and in process get intuitions about what features are composed by the ConvNet to make a certain decision. I think such understanding would be of interest to many.
Preliminary Rating. Please rate the paper according to the following choices. Oral: these are papers whose quality is in the top 10% of the papers at ICCV. Examples include a theoretical breakthrough with no experiments; an interesting solution to a new problem; a novel solution to an existing problem with solid experiments; or an incremental paper that leads to dramatic improvements in performance. Oral/Poster: these are very strong papers, which may have one weakness that makes you unsure as to whether they should be oral or poster. Poster: these are strong papers, which have more than one	Poster

weakness. For example, a well-written paper with solid experiments, but incremental; a paper on a well studied problem with solid theory, but weak experiments; or a novel paper with good experiments, but poorly written. Weak Reject: these are papers that have some promise, but they would be better off by being revised and resubmitted. Strong Reject: these are papers that have major flaws, or have been done before.	
Preliminary Evaluation. Please indicate to the AC, your fellow reviewers, and the authors your current opinion on the paper. Please summarize the key things you would like the authors to include in their rebuttals to facilitate your decision making. There is no need to summarize the paper.	I like the idea proposed in the paper for using attention for modulating the saliency maps and the results for weakly supervised localization are good. However, I would like the authors to make the paper stronger by including results on the detection challenge and compare it with previous results on say PASCAL VOC 2007 - on which multiple results for weakly supervised detection have been reported.
Confidence. Write "Very Confident" to stress that you are absolutely sure about your conclusions (e.g., you are an expert who works in the paper's area), "Confident" to stress that you	

are mostly sure about your conclusions (e.g., you are not an expert but can distinguish good work from bad work in that area), and "Not Confident" to stress that that you feel some doubt about your conclusions. In the latter case, please provide details as confidential comments to PC/AC chairs (point 7.).	Very Confident

Masked Reviewer ID: Assigned_Reviewer_16

Review:

Question	
Paper Summary. Please summarize in your own words what the paper is about.	This paper introduces a feedback mechanism to convolutional networks, which helps better visualization and understanding of how/what CNNs learn during training. The authors introduces a new "feedback" layer, introduced in a pre-trained (on Imagenet classification task) CNN. The fine-tuning is achieved by readjusting the feedback neurons on each layer to optimize the class score of each training sample. These layers update their activation status to maximize the confidence output of the target top neuron. The authors show interesting qualitative results and a few quantitative results.
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: a theoretical paper may need no experiments, while a paper presenting a new approach to a	Feedback mechanisms are of extreme importance in human visual systems. This hints its importance to neural network based models as well. This paper proposes CNN architecture which fine tunes a pre-trained CNN taking into account top-down visual attention. This approach is very similar to Symonian et al. 2014. The feedback scheme allows per class neuron visualization on the original image space by analyzing its gradient. The paper is very well written

known problem may require thorough comparisons to existing methods. Also, please make sure to justify your comments in great detail. For example, if you think the paper is novel, not only say so, but also explain in detail why you think this is the case.	(apart from some small typos) and the ideas are well explained.
Paper Weaknesses. Please discuss the negative aspects of the paper: lack of novelty or clarity, technical errors, insufficient experimental evaluation, etc. Please justify your comments in great detail. If you think the paper is not novel, explain why and give a reference to prior work. Do not ask the authors to cite your own work. If you think this is essential, write it in the confidential comments to the AC. If you think there is an error in the paper, explain in detail why it is an error. If you think the experimental evaluation is insufficient, remember that theoretical results/ideas are essential to ICCV and that a theoretical paper	The high level idea of the paper (top down visual attention with a feedback layer) is interesting. Section 3.2 should be more clear. Interpreting the ReLU and max pooling layers as defined are not very well explained. Most of the model validation are shown on a few number of qualitative images (Figures 4, 5 and 6). This small subset of images is not enough to validate the model. The examples shown are relatively simple in the sense that there are not many objects on the scene. It would be interesting to see more qualitative results. For instance, how would the approach perform in semantic segmentation task (weakly supervised of not)? How does the feedback layer can improve accuracy in either classification, localization or segmentation (comparing the performance of similar architectures with and without it)? In the qualitative experiments, the authors state that a model using GoogleNet performs much better than using AlexNet (Table 2) and in Table 1, the results of the "Feedback" approach are shown using a GoogleNet agains the "Oxford" which does not use it. This way, it is not possible to disentangle what is the improvements of their model compared to the baseline. Also, it would be interesting to see how much does the GraphCut helps in achieving the results stated. It would be interesting to see how the model would perform with and without the GraphCut. Other remarks:

need not have experiments. It is not ok to reject a paper because it did not outperform other existing algorithms, especially if the theory is novel and interesting. It is also not ok to ask for comparisons with unpublished papers and papers published after the ICCV deadline. Last but not least, remember to be polite and constructive.	- Although the quantitative results can be seen in a digital version on the computer, it is almost impossible to see in a printed version. Maybe the contrast on the images could be increased to facilitate visualization? - line 213: would be better to say the name of the authors of reference [10] instead of their affiliation. - line 226: an "et al" after reference [21]? - equation 5: what is the "c" in s_c? - line 520: should be a comma instead of final point. - line 529: ...has THE potential to ... ?
Preliminary Rating. Please rate the paper according to the following choices. Oral: these are papers whose quality is in the top 10% of the papers at ICCV. Examples include a theoretical breakthrough with no experiments; an interesting solution to a new problem; a novel solution to an existing problem with solid experiments; or an incremental paper that leads to dramatic improvements in performance. Oral/Poster: these are very strong papers, which may have one weakness that makes you unsure as to whether	Weak Reject

they should be oral or poster. Poster: these are strong papers, which have more than one weakness. For example, a well-written paper with solid experiments, but incremental; a paper on a well studied problem with solid theory, but weak experiments; or a novel paper with good experiments, but poorly written. Weak Reject: these are papers that have some promise, but they would be better off by being revised and resubmitted. Strong Reject: these are papers that have major flaws, or have been done before.	
Preliminary Evaluation. Please indicate to the AC, your fellow reviewers, and the authors your current opinion on the paper. Please summarize the key things you would like the authors to include in their rebuttals to facilitate your decision making. There is no need to summarize the paper.	The paper tackles an interesting research direction on neural networks using feedback loops in a similar fashion as in Symonian et al. 2014. The experimental section could be vastly improved: most of model validation rely on a few sample images. These images are not enough to guarantee that the model captures top-down visual attention in a general setting. The quantitative results should be improved in different possible ways: (a) make a better comparison with model [25] (for example, use the same pre-trained network), (b) evaluate how the model performs in both weakly and fully supervised semantic segmentation, (c) verify how much of its performance accuracy comes from the use of GraphCut, (d) How much does the feedback layer help in different tasks (for example, compare the accuracy of the exact same model with and without the feedback trick), etc.
Confidence. Write "Very Confident" to stress that you are absolutely sure about your	

conclusions (e.g., you are an expert who works in the paper's area), “Confident” to stress that you are mostly sure about your conclusions (e.g., you are not an expert but can distinguish good work from bad work in that area), and “Not Confident” to stress that that you feel some doubt about your conclusions. In the latter case, please provide details as confidential comments to PC/AC chairs (point 7.).	Confident