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.?

**3.4 描述问题**  
- It is unclear to me whether the feedback neurons are used to optimize weights w from the pre-trained models.

**W是固定的**  
- Please describe how to obtain T\_{k}(z)?

**描述问题**  
- What is the difference between the first and second rows of Figure 5?  
  **描述问题**- 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  
 **加引用**- 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).

**描述问题**

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.

**文章对比问题**  
  
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.  
**容易的可视化实验**（**其实文章中已经有了**）  
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?

**笔误问题**  
  
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.   
**reviewer希望能做弱监督检测的实验**  
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 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.

**文章提到我们的方法有利于理解CNN，但是实验上没有体现出来，Reviewer提出了一个他认为更有意思的想法。**

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.

**3.2节的描述问题**  
  
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)?

**1.认为可视化的图片不够多**

**2.希望看到分割，分类，定位在有feedback和没有feedback的情况下的对比**  
  
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.  
1.reviewer误认为44.6%是oxford文章中用alexnet做的结果，实际上oxford使用alexnet的结果是46.4%，本文中的44.6%是用oxford的方法用googlenet跑出来的，这个文章可能表述的不够清楚，这条也比较简单能搞定

**2**. reviewer希望看到有和没有用**GraphCut**的效果对比试验

Other remarks:  
- 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?

1. 文章中的图片对比度不够好，不适合打印观看，策略：（1）用更好的可视化图片，当时写文章的时候没有换图片就后悔得不行；（2）看看有没有其他方式使得图片更加好看。  
   - 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 ... ?

文字错误，容易