Top-Down Neural Attention by Excitation Backprop

Jianming Zhang[1(✉)], Zhe Lin[2], Jonathan Brandt[2], Xiaohui Shen[2], and Stan Sclaroff[1]

[1] Boston University, Boston, USA
{jmzhang,sclaroff}@bu.edu
[2] Adobe Research, San Jose, USA
{zlin,jbrandt,xshen}@adobe.com

Abstract. We aim to model the top-down attention of a Convolutional Neural Network (CNN) classifier for generating task-specific attention maps. Inspired by a top-down human visual attention model, we propose a new backpropagation scheme, called Excitation Backprop, to pass along top-down signals downwards in the network hierarchy via a probabilistic Winner-Take-All process. Furthermore, we introduce the concept of contrastive attention to make the top-down attention maps more discriminative. In experiments, we demonstrate the accuracy and generalizability of our method in weakly supervised localization tasks on the MS COCO, PASCAL VOC07 and ImageNet datasets. The usefulness of our method is further validated in the text-to-region association task. On the Flickr30k Entities dataset, we achieve promising performance in phrase localization by leveraging the top-down attention of a CNN model that has been trained on weakly labeled web images.

1 Introduction

Top-down task-driven attention is an important mechanism for efficient visual search. Various top-down attention models have been proposed, *e.g.* [1–4]. Among them, the Selective Tuning attention model [3] provides a biologically plausible formulation. Assuming a pyramidal neural network for visual processing, the Selective Tuning model is composed of a bottom-up sweep of the network to process input stimuli, and a top-down Winner-Take-ALL (WTA) process to localize the most relevant neurons in the network for a given top-down signal.

Inspired by the Selective Tuning model, we propose a top-down attention formulation for modern CNN classifiers. Instead of the deterministic WTA process used in [3], which can only generate binary attention maps, we formulate the top-down attention of a CNN classifier as a *probabilistic* WTA process.

The probabilistic WTA formulation is realized by a novel backpropagation scheme, called *Excitation Backprop*, which integrates both top-down and

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46493-0_33) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part IV, LNCS 9908, pp. 543–559, 2016.
DOI: 10.1007/978-3-319-46493-0_33

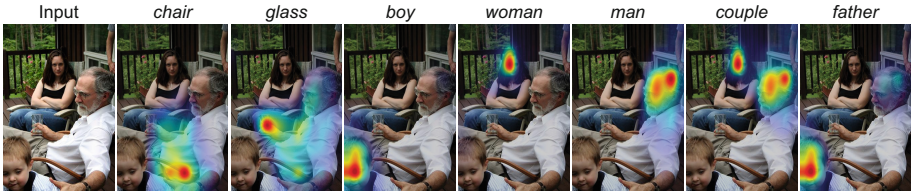

Fig. 1. A CNN classifier's top-down attention maps generated by our Excitation Backprop can localize common object categories, *e.g.* `chair` and `glass`, as well as fine-grained categories like `boy`, `man` and `woman` in this example image, which is resized to 224×224 for our method. The classifier used in this example is trained to predict $\sim$18 K tags using only weakly labeled web images. Visualizing the classifier's top-down attention can also help interpret what has been learned by the classifier. For `couple`, we can tell that our classifier uses the two adults in the image as the evidence, while for `father`, it mostly concentrates on the child. This indicates that the classifier's understanding of `father` may strongly relate to the presence of a child.

bottom-up information to compute the winning probability of each neuron efficiently. Interpretable attention maps can be generated by Excitation Backprop at intermediate convolutional layers, thus avoiding the need to perform a complete backward sweep. We further introduce the concept of contrastive top-down attention, which captures the differential effect between a pair of contrastive top-down signals. The contrastive top-down attention can significantly improve the discriminativeness of the generated attention maps.

In experiments, our method achieves superior weakly supervised localization performance *vs.* [5–9] on challenging datasets such as PASCAL VOC [10] and MS COCO [11]. We further explore the scalability of our method for localizing a large number of visual concepts. For this purpose, we train a CNN tag classifier to predict $\sim$18 K tags using 6M weakly labeled web images. By leveraging our top-down attention model, our image tag classifier can be used to localize a variety of visual concepts. Moreover, our method can also help to understand what has been learned by our tag classifier. Some examples are shown in Fig. 1.

The performance of our large-scale tag localization method is evaluated on the challenging Flickr30k Entities dataset [12]. Without using a language model or any localization supervision, our top-down attention based approach achieves competitive phrase-to-region performance *vs.* a fully-supervised baseline [12].

To summarize, the **main contributions** of this paper are:

- a top-down attention model for CNN based on a probabilistic Winner-Take-All process using a novel Excitation Backprop scheme;
- a contrastive top-down attention formulation for enhancing the discriminativeness of attention maps; and
- a large-scale empirical exploration of weakly supervised text-to-region association by leveraging the top-down neural attention model.

2 Related Work

There is a rich literature about modeling the top-down influences on selective attention in the human visual system (see [13] for a review). It is hypothesized that top-down factors like knowledge, expectations and behavioral goals can affect the feature and location expectancy in visual processing [1,4,14,15], and bias the competition among the neurons [3,15–18]. Our attention model is related to the Selective Tuning model of [3], which proposes a biologically inspired attention model using a top-down WTA inference process.

Various methods have been proposed for grounding a CNN classifier's prediction [5–9,19]. In [5,6,20], error backpropagation based methods are used for visualizing relevant regions for a predicted class or the activation of a hidden neuron. Recently, a layer-wise relevance backpropagation method is proposed by [9] to provide a pixel-level explanation of CNNs' classification decisions. Cao *et al.* [7] propose a feedback CNN architecture for capturing the top-down attention mechanism that can successfully identify task relevant regions. In [19], it is shown that replacing fully-connected layers with an average pooling layer can help generate coarse class activation maps that highlight task relevant regions. Unlike these previous methods, our top-down attention model is based on the WTA principle, and has an interpretable probabilistic formulation. Our method is also conceptually simpler than [7,19] as we do not require modifying a network's architecture or additional training. The ultimate goal of our method goes beyond visualization and explanation of a classifier's decision [6,9,20], as we aim to maneuver CNNs' top-down attention to generate highly discriminative attention maps for the benefits of localization.

Training CNN models for weakly supervised localization has been studied by [21–25]. In [21,24,25], a CNN model is transformed into a fully convolutional net to perform efficient sliding window inference, and then Multiple Instance Learning (MIL) is integrated in the training process through various pooling methods over the confidence score map. Due to the large receptive field and stride of the output layer, the resultant score maps only provide very coarse location information. To overcome this issue, a variety of strategies, *e.g.* image re-scaling and shifting, have been proposed to increase the granularity of the score maps [21,24,26]. Image and object priors are also leveraged to improve the object localization accuracy in [22–24]. Compared with weakly supervised localization, the problem setting of our task is essentially different. We assume a pre-trained deep CNN model is given, which may not use any dedicated training process or model architecture for the purpose of localization. Our focus, instead, is to model the top-down attention mechanism of *generic* CNN models to produce interpretable and useful task-relevant attention maps.

3 Method

3.1 Top-Down Neural Attention Based on Probabilistic WTA

We consider a generic feedforward neural network model. The goal of a top-down attention model is to identify the task-relevant neurons in the network.

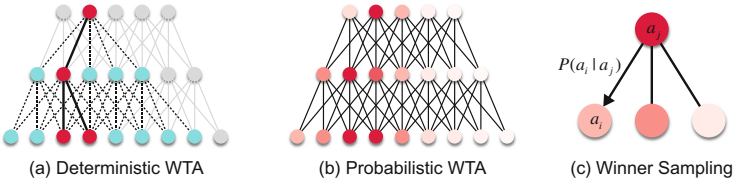

Fig. 2. Deterministic WTA [3] *vs.* our probabilistic WTA for modeling top-down attention. (a) Given a selected output unit, the red dots denote the winners identified by the top-down layer-wise deterministic WTA scheme in the processing cone, and the cyan ones are inhibited. (b) In our probabilistic WTA scheme, winner neurons are generated by a stochastic sampling process (shown in (c)). The top-down signal is specified by a probability distribution over the output units. The shading of a dot in (b) indicates the its relative likelihood of winning against the other ones in the same layer.

Given a selected output unit, a deterministic top-down WTA scheme is used in the biologically inspired Selective Tuning model [3] to localize the most relevant neurons in the processing cone (see Fig. 2 (a)) and generate a binary attention map. Inspired by the deterministic WTA, we propose a *probabilistic* WTA formulation to model a neural network's top-down attention (Fig. 2(b) and (c)), which leverages more information in the network and generates soft attention maps that can capture subtle differences between top-down signals. This is critical to our contrastive attention formulation in Sect. 3.3.

In our formulation, the top-down signal is specified by a prior distribution $P(A_0)$ over the output units, which can model the uncertainty in the top-down control process. Then the winner neurons are recursively sampled in a top-down fashion based on a conditional winning probability $P(A_t|A_{t-1})$, where $A_t, A_{t-1} \in \mathcal{N}$ denote the selected winner neuron at the current and the previous step respectively, and $\mathcal{N}$ is the overall neuron set. We formulate the top-down relevance of each neuron as its probability of being selected as a winner in this process. Formally, given a neuron $a_i \in \mathcal{N}$ (note that a_i denotes a specific neuron and A_t denotes a variable over the neurons), we would like to compute its *Marginal Winning Probability* (MWP) $P(a_i)$. The MWP $P(a_i)$ can be factorized as

$$P(a_i) = \sum_{a_j \in \mathcal{P}_i} P(a_i|a_j)P(a_j), \tag{1}$$

where $\mathcal{P}_i$ is the parent node set of a_i (in top-down order). As Eq. 1 indicates, given $P(a_i|a_j)$, $P(a_i)$ is a function of the marginal winning probability of the parent nodes in the preceding layers. It follows that $P(a_i)$ can be computed in a top-down layer-wise fashion.

Our formulation is equivalent to an absorbing Markov chain process [27] with $p_{ij} := P(a_i|a_j)$ as the transition probability and neurons at the network bottom as the absorbing nodes. $P(a_i)$ can then be interpreted as the expected number of visits when a walker randomly starts from a node in the output layer according to $P(A_0)$. This expected number of visits can be computed by a simple matrix

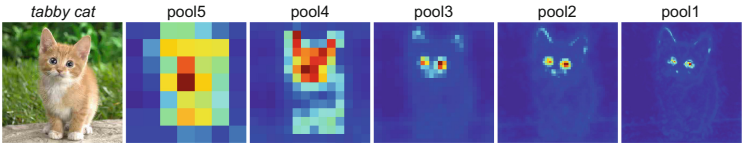

Fig. 3. Example Marginal Winning Probability (MWP) maps computed via Excitation Backprop from different layers of the public `VGG16` model [29] trained on ImageNet. The input image is shown on the right. The MWP maps are generated for the category `tabby cat`. Neurons at higher-level layers have larger receptive fields and strides. Thus, they can capture larger areas but with lower spatial accuracy. Neurons at lower layers tend to more precisely localize features at smaller scale.

multiplication using the fundamental matrix of the absorbing Markov chain [27]. (Detailed explanation can be found in the supplementary material.) In this light, the MWP $P(a_i)$ is a linear function of the the top-down signal $P(A_0)$, which will be shown to be convenient later (see Sect. 3.3).

3.2 Excitation Backprop

In this section, we propose the Excitation Backprop method to realize the probabilistic WTA formulation for modern CNN models.

A modern CNN model [28–30] is mostly composed of a basic type of neuron a_j, whose response is computed by $\widehat{a}_j = \varphi(\sum_i w_{ij}\widehat{a}_i + b_i)$. Here w_{ij} is the weight, $\widehat{a}_i$ is the input, b_i is the bias and φ is the nonlinear activation function. We call this type of neuron an *Activation Neuron*. We have the following assumptions about the activation neurons.

A1. The response of the activation neuron is non-negative.
A2. An activation neuron is tuned to detect certain visual features. Its response is positively correlated to its confidence of the detection.

A1 holds for a majority of the modern CNN models, as they adopt the Rectified Linear Unit (ReLU) as the activation function[1]. **A2** has been empirically verified by many recent works [6,19,31,32]. It is observed that neurons at lower layers detect simple features like edge and color, while neurons at higher layers can detect complex features like objects and body parts.

Between activation neurons, we define a connection to be *excitatory* if its weight is non-negative, and *inhibitory* otherwise. Our Excitation Backprop passes top-down signals through excitatory connections between activation neurons. Formally, let $\mathcal{C}_j$ denote the child node set of a_j (in the top-down order). For each $a_i \in \mathcal{C}_j$, the conditional winning probability $P(a_i|a_j)$ is defined as

$$P(a_i|a_j) = \begin{cases} Z_j \widehat{a}_i w_{ij} & \text{if } w_{ij} \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

[1] We discuss some exceptions and the remedies in the supplementary material.

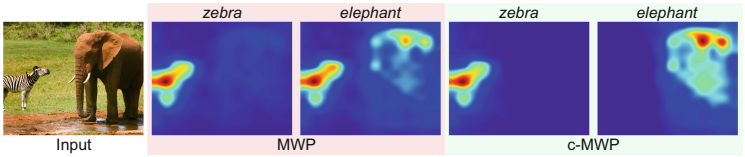

Fig. 4. Marginal Winning Probability (MWP) *vs.* contrastive MWP (c-MWP). The input image is resized to 224×224, and we use `GoogleNet` pretrained on ImageNet to generate the MWP maps and c-MWP maps for "zebra" and "elephant". The MWP map for "elephant" does not successfully suppress the zebra. In contrast, by cancelling out common winner neurons for "elephant" and "non-elephant", the c-MWP map more effectively highlights the elephant.

$Z_j = 1/\sum_{i:w_{ij}\geq 0} \widehat{a}_i w_{ij}$ is a normalization factor so that $\sum_{a_i \in \mathcal{C}_j} P(a_i|a_j) = 1$. In the special case when $\sum_{i:w_{ij}\geq 0} \widehat{a}_i w_{ij} = 0$, we define Z_i to be 0. Note that the formulation of $P(a_i|a_j)$ is valid due to **A1**, since $\widehat{a}_i$ is always non-negative.

Equation 2 assumes that if a_j is a winner neuron, the next winner neuron will be sampled among its child node set $\mathcal{C}_j$ based on the connection weight w_{ij} and the input neuron's response $\widehat{a}_i$. The weight w_{ij} captures the top-down feature expectancy, while $\widehat{a}_i$ represents the bottom-up feature strength, as assumed in **A2**. Due to **A1**, child neurons of a_j with negative connection weights always have an inhibitory effect on a_j, and thus are excluded from the competition.

Equation 2 recursively propagates the top-down signal layer by layer, and we can compute attention maps from any intermediate convolutional layer. For our method, we simply take the sum across channels to generate a marginal winning probability (MWP) map as our attention map, which is a 2D probability histogram. Figure 3 shows some example MWP maps generated using the pre-trained `VGG16` model [29]. Neurons at higher-level layers have larger receptive fields and strides. Thus, they can capture larger areas but with lower spatial accuracy. Neurons at lower layers tend to more precisely localize features at smaller scales.

3.3 Contrastive Top-Down Attention

Since the MWP is a linear function of the top-down signal (see Sect. 3.1), we can compute any linear combination of MWP maps for an image by a single backward pass. All we need to do is linearly combine the top-down signal vectors at the top layer before performing the Excitation Backprop. In this section, we take advantage of this property to generate highly discriminative top-down attention maps by passing down pairs of contrastive signals.

For each output unit o_i, we virtually construct a dual unit $\bar{o}_i$, whose input weights are the negation of those of o_i. For example, if an output unit corresponds to an `elephant` classifier, then its dual unit will correspond to a `non-elephant` classifier. Subtracting the MWP map for `non-elephant` from the one for `elephant` will cancel out common winner neurons and amplify the

discriminative neurons for `elephant`. We call the resulting map a *contrastive* MWP map, which can be computed by a single backward pass. More details can be found in our supplementary material. In practice we weight the target unit and its dual equally, and truncate the contrastive MWP map at zero so that only positive parts are kept. Our probabilistic formulation ensures that there are always some positive parts on the contrastive MWP map, unless the MWP map and its dual are identical. Figure 4 shows some examples.

4 Experiments

We implement Excitation Backprop in Caffe [33] (available at our project website[2]). Implementation details are included in our supplementary material.

4.1 The Pointing Game

The goal of this section is to evaluate the *discriminativeness* of different top-down attention maps for localizing target objects in crowded visual scenes.

Evaluation setting. Given a pre-trained CNN classifier, we test different methods in generating a top-down attention map for a target object category present in an image. Ground truth object labels are used to cue the method. We extract the maximum point on the top-down attention map. A hit is counted if the maximum point lies on one of the annotated instances of the cued object category, otherwise a miss is counted. We measure the localization accuracy by $Acc = \frac{\#Hits}{\#Hits+\#Misses}$ for each object category. The overall performance is measured by the mean accuracy across different categories.

We call this the *Pointing Game*, as it asks the CNN model to point at an object of designated category in the image. The pointing game does not require highlighting the full extent of an object, and it does not account for the CNN model's classification accuracy. Therefore, it purely compares the *spatial selectiveness* of the top-down attention maps. Moreover, the pointing game only involves minimum post-processing of the attention maps, so it can evaluate different types of attention maps more fairly.

Datasets. We use the test split of the PASCAL VOC07 dataset [10] (4952 images) and the validation split of the MS COCO dataset [11] (40137 images). In particular, COCO contains 80 object categories, and many of its images have multiple object categories, making even the simple Pointing Game rather challenging. To evaluate success in the Pointing Game, we use the groundtruth bounding boxes for VOC07 and the provided segmentation masks for COCO.

CNN classifiers. We consider three popular CNN architectures: `CNN-S` [34] (an improved version of AlexNet [28]), `VGG16` [29], and `GoogleNet` [30]. These models vary a lot in depth and structure. We download these models from the Caffe Model Zoo website [35]. These models are pre-trained on ImageNet [36].

[2] http://www.cs.bu.edu/groups/ivc/excitation-backprop.

For both VOC07 and COCO, we use the training split to fine-tune each model. We follow the basic training procedure for image classification, and thus no multi-scale training is used. Only the output layer is fine-tuned using the multi-label cross-entropy loss for simplicity, since the classification accuracy is not our focus. More details are included in our supplementary material.

Test methods. We compare Excitation Backprop (MWP and c-MWP) with the following methods: (Grad) the error backprogation method [5], (Deconv) the deconvolution method originally designed for internal neuron visualization [6], (LRP) layer-wise relevance propagation [9], and (CAM) the class activation map method [8]. We implement Grad, Deconv and CAM in Caffe. For Deconv, we use an improved version proposed in [20], which generates better maps than the original version [6]. For Grad and Deconv, we follow [5] to use the maximum absolute value across color channels to generate the final attention map. Taking the mean instead of maximum will degrade their performance. For LRP, we use the software provided by the authors, which only supports CPU computation. For `VGG16`, this software can take 30 s to generate an attention map on an Intel Xeon 2.90GHz×6 machine[3]. Due to limited computational resources, we only evaluate LRP for `CNN-S` and `GoogleNet`.

Note that CAM is only applicable to certain architectures like `GoogleNet`, which do not have fully connected layers. At test time, it acts like a fully convolutional model to perform dense sliding window evaluation [21,37]. Therefore, the comparison with CAM encompasses the comparison with the dense evaluation approach for weakly supervised localization [21].

To generate the full attention maps for images of arbitrary aspect ratios, we convert each testing CNN classifier to a fully convolutional architecture as in [21]. All the compared methods can be easily extended to fully convolutional models. In particular, for Excitation Backprop, Grad and Deconv, the output confidence map of the target category is used as the top-down signal to capture the spatial weighting. However, all input images are resized to 224 in the smaller dimension, and no multi-scale processing is used.

For different CNN classifiers, we empirically select different layers to compute our attention maps based on a held-out set. We use the conv5 layer for `CNN-S`, pool4 for `VGG16` and pool2 for `GoogleNet`. We use bicubic interpolation to upsample the generated attention maps. The effect of the layer selection will be analysed below. For Grad, Deconv and LRP we blur their maps by a Gaussian kernel with $\sigma = 0.02 \cdot \max\{W, H\}$, which slightly improves their performance since their maps tend to be sparse and noisy at the pixel level. In the evaluation, we expand the groundtruth region by a tolerance margin of 15 pixels, so that the attention maps produced by CAM, which are only 7 pixels in the shortest dimension, can be more fairly compared.

Results. The results are reported in Table 1. As the pointing game is trivial for images with large dominant objects, we also report the performance on a difficult

[3] On COCO, we need to compute about 116 K attention maps, which leads to over 950 h of computation on a single machine for LRP using `VGG16`.

Table 1. Mean accuracy (%) in the pointing game. For each method, we report two scores for the overall test set and a difficult subset respectively. `Center` is the baseline that points at image center. The second best score of each column is underlined.

	VOC07 Test (All/Diff.)			COCO Val. (All/Diff.)		
	CNN-S	VGG16	GoogleNet	CNN-S	VGG16	GoogleNet
Center	69.5/42.6	69.5/42.6	69.5/42.6	27.7/19.4	27.7/19.4	27.7/19.4
Grad [5]	78.6/59.8	76.0/56.8	79.3/61.4	38.7/30.1	37.1/30.7	42.6/36.3
Deconv [6]	73.1/45.9	75.5/52.8	74.3/49.4	36.4/28.4	38.6/30.8	35.7/27.9
LRP [9]	68.1/41.3	-	72.8/50.2	32.5/24.0	-	40.2/32.7
CAM [8]	-	-	80.8/61.9	-	-	41.6/35.0
MWP	73.7/52.9	76.9/55.1	79.3/60.4	35.0/27.7	39.5/32.5	43.6/37.1
c-MWP	**78.7/61.7**	**80.0/66.8**	**85.1/72.3**	**43.0/37.0**	**49.6/44.2**	**53.8/48.3**

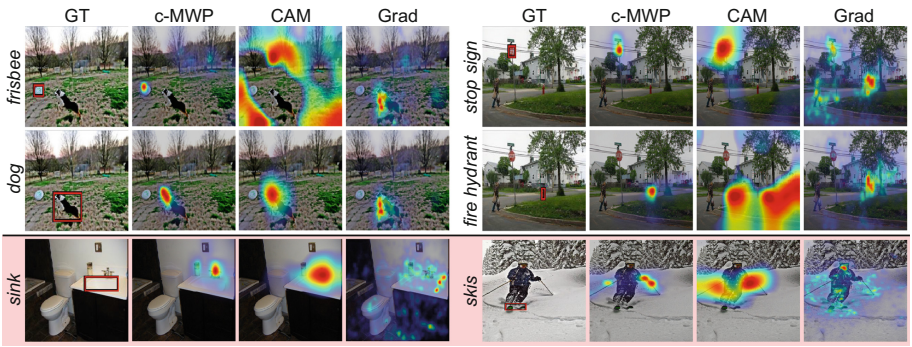

Fig. 5. Example attention maps using `GoogleNet`. For visualization, the maps are superimposed on the images after some postprocessing (slight blur for Grad and thresholding for CAM). (*Top two rows*) Our c-MWP is very discriminative and can often localize challengingly small objects like `frisbee`, `stop sign` and `fire hydrant`. (*Bottom row*) Two typical failure cases of top-down neural attention are shown. Since `faucet` often co-occurs with `sink`, the CNN's attention falsely focuses on the faucet in the image. It is the same case for `ski poles` and `skis`.

subset of images for each category. The difficult set includes images that meet two criteria: (1) the total area of bounding boxes (or segments in COCO) of the testing category is smaller than 1/4 the size of the image and (2) there is at least one other distracter category in the image.

Our c-MWP consistently outperforms the other methods on both VOC07 and COCO across different CNN models. c-MWP is also substantially better than MWP, which validates the idea of contrastive attention. `GoogleNet` provides the best localization performance for different methods, which is also observed by [7,8]. Using `GoogleNet`, our c-MWP outperforms the second best method by about 10 % points on the difficult sets of VOC07 and COCO. In particular, our c-

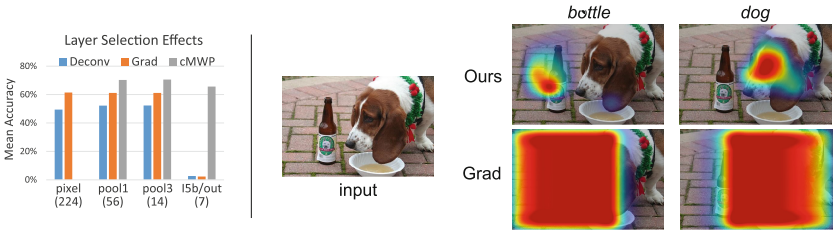

Fig. 6. Effects of layer selection on VOC07 *difficult* set. (*Left*) For Grad, Deconv and our c-MWP, we compare their attention maps from three different layers in the `GoogleNet`. At I5b/out, Grad and Deconv fail to generate meaningful attention maps, while our method can still achieve reasonable accuracy. (*Right*) We show example attention maps by our c-MWP and Grad from the I5b/out layer.

Table 2. Analysis of contrastive attention on VOC07 *difficult* set using `GoogleNet`. We evaluate two variants of Excitation Backprop for the contrastive attention map computation compared with our full model. We also test the contrastive attention idea for Grad, Deconv and CAM and their original scores are shown in brackets. See text for details.

	Excitation backprop			Other methods		
	full	post-norm	w/o norm	c-Grad	c-Deconv	c-CAM
Mean Acc. (%)	**70.6**	58.1	41.6	N.A	67.7 (49.4)	61.9 (61.9)

MWP gives the best performance in 69/80 object categories of COCO, especially for small objects like `remote`, `tie` and `baseball bat` (see our supplementary material).

Example attention maps are shown in Fig. 5. As we can see, our c-MWP maps can accurately localize the cued objects in rather challenging scenes. More examples are included in our supplementary material.

Layer selection effects. We use `GoogleNet` to analyze the effects of layer selection. For a comparison, we also report the performance of Grad and Deconv by taking the maximum gradient magnitude across feature map channels in the intermediate layers. Results are reported in Fig. 6. We choose three intermediate layers in `GoogleNet`: pool1, pool3 and Inception_5b/output (I5b/out), whose spatial resolutions are 56, 14 and 7 in the shortest dimension respectively. Performance does not vary much across all methods at the chosen layers except I5b/out. Our c-MWP only gets a slight decrease in accuracy (mainly due to the map's low spatial resolution), while Grad and Deconv do not generate meaningful attention maps (see Fig. 6). This is because the attention maps of Grad and Deconv at I5b/out are not conditioned on the activation values of I5b/out, and thus fail to leverage the spatial information captured by I5b/out.

Analysis of contrastive top-down attention. The proposed contrastive attention is conceptually simple, which basically subtracts one attention map

Table 3. Bounding box localization error on ImagNet Val. using `GoogleNet`. *The score of Feedback is from the original paper.

	Grad [5]	Deconv [6]	LRP [9]	CAM [8]	Feedback* [7]	c-MWP	MWP
Opt. α	5.0	4.5	1.0	1.0	-	0.0	1.5
Loc. Error (%)	41.6	41.6	57.8	48.1	38.8	57.0	**38.7**

from its dual using the virtual contrastive output unit. We test this idea for Grad, Deconv and CAM and the performance is reported in Table 2. For Grad, the gradient magnitude map is identical to its dual since the gradients of the dual map are just the negation of the reference map. As a result, the subtraction gives a zero map. For CAM, the performance remains the same because the dual map is again a negation of the reference attention map and the maximum point will not be changed by the subtraction. However, the proposed contrastive attention works for Deconv, when the attention map and its dual are L1-normalized before subtraction. Deconv shares a similar spirit of our method as it discards negative/inhibitatory signals by thresholding at ReLU layers, but it also introduces non-linearity in the propagation process. Therefore, it requires two backward passes and proper normalization, while our method can directly propagate the contrastive signal via a single pass and achieves better performance.

Our probabilistic WTA formulation produces well-normalized attention maps that enable direct subtraction. We report the performance of two variants of our method in Table 2. We remove the normalization factor Z_i in Eq. 2 and pass down the contrastive signal. This leads to a significant degradation in performance (w/o norm). Then we compute the attention map and its dual separately and do the subtraction after L1-normalization (post-norm). The performance is improved but still substantially lower than our full method. This analysis further confirms the importance of our probabilistic formulation.

4.2 Localizing Dominant Objects

We now turn to a different evaluation setting [7]. The goal of this setting is bounding box (bbox) localization of dominant objects in the image.

Dataset and evaluation. We follow the protocol of Feedback Net [7] for a fair comparison. The test is performed on the ImageNet Val. set (∼50 K images), where each image has a label representing the category of dominant objects in it. The label is given, so the evaluation is based on the localization error rate with an IOU threshold at 0.5. Images are resized to 224×224.

As in [7], simple thresholding is used to extract a bbox from an attention map. We set the threshold $\tau = \alpha\mu_I$, where μ_I is the mean value of the map. Then the tightest bbox covering the white pixels is extracted. The parameter α is optimized in the range $[0 : 0.5 : 10]$ for each method on a held out set.

Results. Table 3 reports the results based on the same `GoogleNet` model obtained from Caffe Model Zoo [35] as in [7]. We find that c-MWP performs

poorly, but our MWP obtains competitive results against Feedback and other methods. Compared with Feedback, our method is conceptually much simpler. Feedback requires modification of a CNN's architecture and needs 10–50 iterations of forward-backward passes for computing an attention map.

Note that this task favors attention maps that fully cover the *dominant* object in an image. Thus, it is very different from the Pointing Game, which favors discriminativeness instead. Our c-MWP usually only highlights the most discriminative part of an object due to the competition between the contrastive pair of top-down signals. This experiment highlights the versatility of our method, and the value of the non-contrastive version (MWP) for dominant object localization.

4.3 Text-to-Region Association

Text-to-region association in unconstrained images [12] is very challenging compared to the object detection task, due to the lack of fully-annotated datasets and the large number of words/phrases used in the natural language. Moreover, an image region can be referred to by potentially many different words/phrases, which further increases the complexity of the fully-supervised approach.

By leveraging the top-down attention of a CNN image tag classifier, we propose a highly scalable approach to weakly supervised word-to-region association. We train an image tag classifier using ∼6M weakly labeled thumbnail images collected from a commercial stock image website[4] (Stock6M). Each image is 200-pixels in the longest dimension and comes with about 30–50 user tags. These tags cover a wide range of concepts, including objects, scenes, body parts, attributes, activities, and abstract concepts, but are also very noisy. We picked ∼18 K most frequent tags for our dictionary. We empirically found that the first few tags of each image are usually more relevant, and consequently use only the first 5 tags of an image in the training.

Tag classifier training. We use the pre-trained `GoogleNet` model from Caffe Model Zoo, and fine-tune the model using the multi-label cross-entropy objective function for the 18 K tags. Images are padded to square shape by mirror padding and upsampled to 256×256. Random flipping and cropping are used for data augmentation. We use SGD with a batch size of 64 and a starting learning rate of 0.01. The learning rate is lowered by a factor of 0.1 when the validation error plateaus. The training process passes through the data for three epochs and takes ∼55 h on an NVIDIA K40c GPU.

Dataset and evaluation. To quantitatively evaluate our top-down attention method and the baselines in text-to-region association, we use the recently proposed Flickr30k Entities (Flickr30k) dataset [12]. Evaluation is performed on the test split of Flickr30k (1000 images), where every image has five sentential descriptions. Each Noun Phrase (NP) in a sentence is manually associated with the bounding box (bbox) regions it refers to in the image. NPs are grouped into

[4] https://stock.adobe.com.

Table 4. Performance comparison on the Flickr30k Entities dataset. We report performance for both the whole dataset and a subset of small instances. The R@N refers to the overall recall rate regardless of phrase types. mAP (Group) and mAP (Phrase) should be interpreted differently, because most phrases belong to the group `people`. CCA* refers to the precomputed results provided by [12], while CCA and SPE are the results reported in the original paper. MCG_base is the performance using MCG's original proposal scores. EB is EdgeBoxes [39].

	opt. γ	R@1	R@5	R@10	mAP (Group)	mAP (Phrase)
MCG_base	–	10.7/ 7.7	30.3/22.4	40.5/30.3	6.9/ 4.5	16.8/12.9
Grad (MCG)	0.50	24.3/ 7.6	49.6/32.9	59.7/45.8	10.2/ 3.8	28.8/15.6
Deconv (MCG)	0.50	21.5/11.3	48.4/34.5	58.5/46.0	10.0/ 4.0	26.5/16.7
LRP (MCG)	0.50	24.3/11.8	51.6/36.8	61.3/48.5	10.3/ 4.3	28.9/18.1
CAM (MCG)	0.75	21.7/ 6.5	47.1/27.9	56.1/39.1	7.5/ 2.0	26.0/11.9
MWP (MCG)	0.50	**28.5**/15.0	52.7/39.1	61.3/49.8	11.8/ 5.3	**31.1**/20.3
c-MWP (MCG)	0.50	26.2/**21.2**	**54.3**/**43.4**	**62.2**/**51.7**	**15.2**/**10.8**	30.8/**24.0**
CCA* [12] (EB)	–	25.2/**21.8**	**50.3**/**41.0**	58.1/**47.3**	12.8/**11.5**	28.8/**23.6**
CCA [12] (EB)	–	25.3/ –	–	**59.7**/ –	11.2/ –	–
c-MWP (EB)	0.25	**27.0**/18.4	49.9/35.2	57.7/43.9	**13.2**/ 8.1	**29.4**/20.0

Table 5. Per group recall@5 (%) on the Flickr30k Entities dataset. The mean scores are computed over different group types, which are different from the overall recall rates reported in Table 4.

	People	Clothing	Bodypart	Animal	Vehicle	Instrument	Scene	Other	Mean
MCG_base	36.1	30.1	9.9	50.8	37.8	26.5	31.5	19.1	30.3
Grad (MCG)	65.0	32.4	14.0	**70.1**	63.0	40.7	58.8	32.5	47.1
Deconv (MCG)	65.4	31.6	18.7	67.0	64.0	46.9	53.6	28.9	47.0
LRN (MCG)	64.6	37.7	16.4	62.9	63.5	45.7	59.4	37.9	48.5
CAM (MCG)	60.5	28.4	9.6	57.0	57.5	37.0	**64.4**	32.7	43.4
MWP (MCG)	**68.6**	37.7	16.1	68.7	66.3	53.7	54.5	36.8	50.3
c-MWP (MCG)	63.5	**47.6**	**24.5**	69.9	**72.0**	**54.3**	61.0	**40.2**	**54.1**
CCA* [12] (EB)	**63.6**	**43.7**	**22.9**	57.0	69.0	50.6	45.0	**36.2**	48.5
c-MWP (EB)	62.8	35.0	17.6	**65.1**	**73.5**	**58.6**	**53.2**	**36.2**	**50.3**

eight types (see [12]). Given an NP, the task is to provide a list of scored bboxes, which will be measured by the recall rate (similar to the object proposal metric) or per-group/per-phrase Average Precision (AP) (similar to the object detection metric). We use the evaluation code from [12].

To generate scored bboxes for an NP, we first compute the word attention map for each word in the NP using our tag classifier. Images are resized to 300 pixels in the shortest dimension to better localize small objects. Then we simply average the word attention maps to get an NP attention map. Advanced language models can be used for better fusing the word attention maps, but we adopt the simplest fusion scheme to demonstrate the effectiveness of our top-down

attention model. We skip a small proportion of words that are not covered by our 18 K dictionary. MCG [38] is used to generate 500 segment proposals, which are re-scored based on the phrase attention map. The re-scored segments are then converted to bboxes, and redundant bboxes are removed via Non-maximum Suppression using the IOU threshold of 0.7.

The segment scoring function is defined as $f(R) = S_R/A_R^\gamma$ where S_R is the sum of the values inside the segment proposal R on the given attention map and A_R is the segment's area. The parameter γ is to control the penalty of the segment's area, which is optimized for each method in the range $[0 : 0.25 : 1]$.

Results. The recall rates and mAP scores are reported in Table 4. For our method and the baselines, we additionally report the performance on a subset of small instances whose bbox area is below 0.25 of the image size, as we find small regions are much more difficult to localize. Our c-MWP consistently outperforms all the attention map baselines across different metrics. In particular, the group-level mAP of our method is better than the second best by a large margin.

We also compare with a recent fully supervised method [12], which is trained directly on the Flickr30k Entities dataset using CNN features. For fair comparison, we use the same bbox proposals used in [12], which are generated by EdgeBoxes (EB) [39]. These proposals are pre-computed and provided by [12]. Our performance using EB is lower than using MCG, mainly due to the lower accuracy of the EB's bbox proposals. Compared with the segmentation proposals, the bbox proposals can also affect our ranking function for small and thin objects. However, our method still attains competitive performance against [12]. Note that our method is weakly supervised and does not use any training data from the Flickr30k Entities dataset.

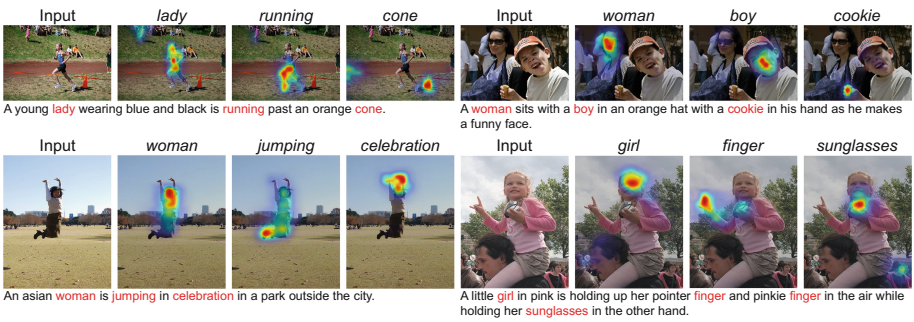

Fig. 7. Word attention maps obtained by c-MWP using our image tag classifier. For each test image, one of its caption annotations from Flickr30k Entities is displayed below. We show the attention maps for the words in red in each caption. By leveraging a large-scale weakly labeled dataset, our method can localize a large number of visual concepts, *e.g.* objects (cone, sunglasses and cookie), fine-grain categories of people (woman and boy), body parts (finger) and actions (jumping, running and celebration). More examples are included in our supplementary material.

We further report the per-group Recall@5 score in Table 5. Our method achieves promising results in many group types, *e.g.* `vehicle` and `instrument`. Note that the fully supervised CCA (EB) [12] gives significantly worse performance than c-MWP (EB) in `animal`, `vehicle` and `instrument`, which are the three rarest types in the Flickr30k Entities dataset. This again shows the limitation of fully-supervised approaches due to the lack of fully-annotated data.

Some example word attention maps are shown in Fig. 7 to demonstrate the localization ability of our method. As we can see, our method can localize not only noun phrases but also actions verbs in the text.

5 Conclusion

We propose a probabilistic Winner-Take-All formulation to model the top-down neural attention for CNN classifiers. Based on our formulation, a novel propagation method, Excitation Backprop, is presented to compute the Marginal Winning Probability of each neuron. Using Excitation Backprop, highly discriminative attention maps can be efficiently computed by propagating a pair of contrastive top-down signals via a single backward pass in the network. We demonstrate the accuracy and the generalizability of our method in a large-scale Pointing Game. We further show the usefulness of our method in localizing dominant objects. Moreover, without using any localization supervision or language model, our neural attention based method attains competitive localization performance *vs.* the state-of-the-art fully supervised methods on the challenging Flickr30k Entities dataset.

Acknowledgments. This research was supported in part by Adobe Research, US NSF grants 0910908 and 1029430, and gifts from NVIDIA.

References

1. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Vaina, L.M. (ed.) Matters of Intelligence. Conceptual Structures in Cognitive Neuroscience. Synthese Library, vol. 188, pp. 115–141. Springer, New York (1987)
2. Anderson, C.H., Van Essen, D.C.: Shifter circuits: a computational strategy for dynamic aspects of visual processing. Proc. Natl. Acad. Sci. **84**(17), 6297–6301 (1987)
3. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artif. Intell. **78**(1), 507–545 (1995)
4. Wolfe, J.M.: Guided search 2.0 a revised model of visual search. Psychon. Bull. Rev. **1**(2), 202–238 (1994)
5. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: ICLR Workshop (2014)
6. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_53

7. Cao, C., Liu, X., Yang, Y., Yu, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., Xu, W., et al.: Look and think twice: capturing top-down visual attention with feedback convolutional neural networks. In: ICCV (2015)
8. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization (2016)
9. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One **10**(7), e0130140 (2015)
10. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
11. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_48
12. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: CVPR (2015)
13. Baluch, F., Itti, L.: Mechanisms of top-down attention. Trends Neurosci. **34**(4), 210–224 (2011)
14. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**(1), 97–136 (1980)
15. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Ann. Rev. Neurosci. **18**(1), 193–222 (1995)
16. Reynolds, J.H., Heeger, D.J.: The normalization model of attention. Neuron **61**(2), 168–185 (2009)
17. Abrial, J.-R.: On B. In: Bert, D. (ed.) B 1998. LNCS, vol. 1393, pp. 1–8. Springer, Heidelberg (1998). doi:10.1007/BFb0053350
18. Beck, D.M., Kastner, S.: Top-down and bottom-up mechanisms in biasing competition in the human brain. Vis. Res. **49**(10), 1154–1165 (2009)
19. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. In: ICLR (2015)
20. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv preprint (2014). arXiv:1412.6806
21. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: CVPR (2015)
22. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: ICCV (2015)
23. Papandreou, G., Chen, L.C., Murphy, K., Yuille, A.L.: Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In: ICCV (2015)
24. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR (2015)
25. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: CVPR (2015)
26. Pinheiro, P.H., Collobert, R.: Recurrent convolutional neural networks for scene parsing. In: ICLR (2014)
27. Kemeny, J.G., Snell, J.L., et al.: Finite Markov Chains. Springer, New York, Berlin, Heidelberg, Tokyo (1960)
28. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)

29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
30. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
31. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS (2014)
32. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint (2015). arXiv:1506.06579
33. Caffe: convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia (2014)
34. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: BMVC (2014)
35. Caffe Model Zoo. https://github.com/BVLC/caffe/wiki/Model-Zoo
36. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
37. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. In: ICLR (2014)
38. Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: CVPR (2014)
39. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_26