

AttentionNet: Aggregating Weak Directions for Accurate Object Detection

Donggeun Yoo
dgyoo@rcv.kaist.ac.kr
KAIST

Sunggyun Park
sunggyun@kaist.ac.kr
KAIST

Joon-Young Lee*
jylee@rcv.kaist.ac.kr
KAIST

Anthony S. Paek
apaek@lunit.io
Lunit Inc.

In So Kweon
iskweon@kaist.ac.kr
KAIST

Abstract

We present a novel detection method using a deep convolutional neural network (CNN), named AttentionNet. We cast an object detection problem as an iterative classification problem, which is the most suitable form of a CNN. AttentionNet provides quantized weak directions pointing a target object and the ensemble of iterative predictions from AttentionNet converges to an accurate object boundary box. Since AttentionNet is a unified network for object detection, it detects objects without any separated models from the object proposal to the post bounding-box regression. We evaluate AttentionNet by a human detection task and achieve the state-of-the-art performance of 65% (AP) on PASCAL VOC 2007/2012 with an 8-layered architecture only.

1. Introduction

After the recent advance [16] of deep convolutional neural network (CNN) [7], CNN based object classification methods in computer vision has reached human-level performance on ILSVRC classification [18]: top-5 error of 4.94% [15], 6.67% [22], 6.8% [21], and 5.1% for human [18]. These successful methods, however, have limited range of application since most of the images are not object-centered. Thus, current research focus in visual recognition is moving towards richer image understanding such as object detection or pixel-level object segmentation. Our focus lies in the object detection problem.

Although many researchers had proposed various CNN-based techniques for object detection [24, 19, 12, 9], it is still challenge to estimate object bounding boxes beyond the object existence. Even the most successful framework so far, Region-CNN [12] reported top scores [17, 21, 22] in ILSVRC'14, but it is relatively far from human-level accuracy compared to the classification. One major limitation of this framework is that the proposal quality highly affects the detection performance. Another side of detection with CNN, regression models are also applied to detection

*This work was done when he was in KAIST. He is currently working in Adobe Research.



Figure 1. Real detection examples of our detection framework. Starting from an image boundary (dark blue bounding box), our detection system iteratively narrows the bounding box down to a final human location (red bounding box).

[24, 19, 9], but direct mapping from an image to an exact bounding box is relatively difficult for a CNN. We thought that there must be a room for modification and improvement for the use of CNN as a regressor.

In this paper, we introduce a novel and straightforward detection method by integrating object existence estimation and bounding box optimization into a single convolutional network. Our model is on a similar line of a detection-

by-regression approach, but we adopt a successful CNN-classification model rather than a CNN-regression model. We estimate an exact bounding box by aggregating many weak predictions from the classification model, such as an ensemble method combines many weak learners to produce a strong learner. We modify a traditional classification model to a suitable form, named *AttentionNet*, for estimating an exact bounding box. This network provides quantized directions pointing a target object from the top-left (TL) and bottom-right (BR) corner of an image.

Fig. 1 shows real detection examples of our method. Starting from an entire image, the image is recursively cropped according to the predicted directions at TL/BR and fed to *AttentionNet* again, until the network converges to a bounding box fitting a target object. *Each direction is weak but the ensemble of directions is sufficient to estimate an exact bounding box.* The difficulty of estimating an exact bounding box at once with a regression model is solved by a classification model. For multiple instances, we also place our framework on the sliding window paradigm but introduce an efficient method to cope with those.

Compared with the previous state-of-the-art methods, a single *AttentionNet* does everything but yields state-of-the-art detection performance. Our framework does not involve any separated models for object proposals nor post bounding box regression. A single *AttentionNet* 1) detects regions where a single instance is included only, 2) provides directions to an instance at each region, 3) and also correct the miss-localizations such as bounding box regression [11].

AttentionNet we demonstrate in this paper is not scalable to multiple classes yet. Extension of this method to multiple classes with a single *AttentionNet* is ongoing. We therefore demonstrate single-class object detection on public datasets to verify the strength of our model. We primarily evaluate our detection framework by the human detection task and extend to a non-human class.

Contributions Our contributions are three folds:

1. We suggest a novel detection method, which estimates an exact bounding box by aggregating weak predictions from *AttentionNet*.
2. Our method does not include any separated models such as the object proposal, object classifiers and post bounding box regressor. *AttentionNet* does all these.
3. We achieve the state-of-the-art performance on single-class object detection tasks.

2. Related Works

Object detection has been actively studied for the last few decades. One of the most popular approaches is part-based models due to their strength in handling pose varia-

tions and occlusions. Deformable Part Model (DPM), proposed by Felzenszwalb *et al.* [11], is a flexible model composed of object parts combined with deformation cost to manage severe deformations. Poselets, proposed by Bourdev *et al.* [2], is another part-based model demonstrating competitive performance. Poselets has numerous part-wise HOG detectors covering wide pose variations. Activated parts vote the location of objects.

In recent years, CNN-based approach leads to the successful development of object detection with drastic advances of deep CNN [7]. Large scale ImageNet [8] database and the raise of parallel computing contribute to a breakthrough in detection [24, 19, 12, 17, 21, 22] as well as classification [16].

The state-of-the-art method in object detection is the Region-CNN (R-CNN), which represents a local region by CNN activations [12]. Specifically, R-CNN framework proceeds as follows. First it extracts local regions which probably contain an object by using an object proposal methods [26]. The local regions, called object proposals, are warped into a fixed size and fed to a pre-trained CNN. Then each proposal is represented by mid-level CNN activations (e.g. 7th convolutional layer) and evaluated by separated classifiers (e.g. SVMs). The object proposals are then merged and fed to a bounding box regressor [11] to correct miss-localizations. Despite its efficiency and successful performance, it has a limitation that proposal quality highly affects detection performance. If the proposal model fails to propose a suitable candidate, the rest procedures will not have the opportunity to detect it. For this reason, [9, 23] proposed a new class-agnostic proposal method with a CNN regression model to improve the proposal quality while reducing the number of proposals. Also, R-CNN is a cascaded model composed of individual components as object proposal, feature extraction, classification, and bounding box regression, therefore these should be individually engineered for the best performance.

Apart from R-CNN framework, there is another CNN-based approach, which considers object detection as a regression problem. Szegedy *et al.* [24] trains a CNN which maps an image to a rectangular mask of an object. Sermanet *et al.* [19] also employ a similar approach but their CNN directly estimates bounding box coordinates. These methods are free from object proposals, but it is still debatable to leave all to a CNN trained with a mean-square cost to produce an exact bounding box.

Compared to the previous CNN methods, our method does not rely on object proposals since we actively explore objects by iterative classifications. Also, the proposed network has a unified classification architecture, which is verified from many CNN applications and also does not need to tune up individual components.

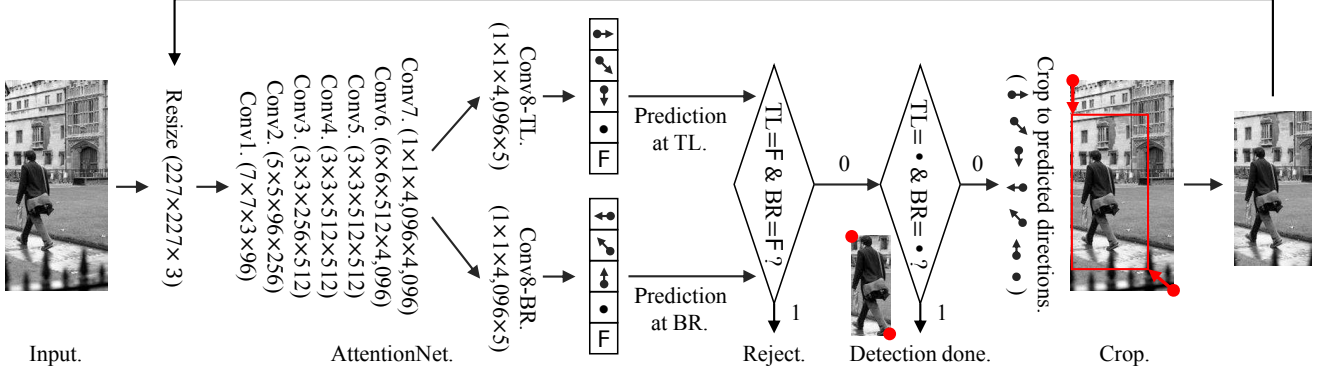


Figure 2. A pipeline of our detection framework. AttentionNet is composed of two final layers for top-left (TL) and bottom-right (BR) of the input image domain. Each of them outputs a direction ($\rightarrow \searrow \downarrow$ for TL, $\leftarrow \nwarrow \uparrow$ for BR) where each corner of the image should go to for the next step, or a “stop” sign (\bullet), or “non-human” sign (F). When AttentionNet outputs “non-human” in both layers, the image is rejected. The image is cropped according to the weak directions and fed to AttentionNet again, until it meets “stop” in both layers.

3. Detection with AttentionNet

We introduce how our detection framework operates under a constrained condition where an image includes a single target instance only. Extension of this detection system to multiple instances will be described in Sec. 4.

3.1. Overview

We summarize the core algorithm of our detection framework in Fig. 2. We warp an input image into a fixed CNN input size and feed it to AttentionNet. AttentionNet outputs two directional predictions corresponding the top-left (TL) corner and the bottom-right (BR) corner of the input image. For example, possible predictions for TL are the following: “go to right (\rightarrow)”, “go to right-down (\searrow)”, “go to down (\downarrow)”, “stop here (\bullet)” and “no instance in this image (F)”. Let us assume the prediction of AttentionNet indicates \downarrow_{TL} and \nwarrow_{BR} as shown in Fig. 2. We then crop the input image to the corresponding directional vectors of a fixed length l , and feed the cropped image to AttentionNet again to get next directions. This procedure is repeated until the image meets the following conditions: F in both corners, or \bullet in both corners. If AttentionNet returns F at both corners, the test ends with a final decision of “no target instance in this image”. Also, if AttentionNet returns \bullet at both corners, the test ends with the current image as a detection result. This detected image can be back-projected to a corresponding bounding box in the original input image domain. Given a stopped (detected) bounding box b and corresponding output activations $y_{TL}, y_{BR} \in \mathbb{R}^5$, the detection score s^b is discriminatively defined as,

$$\begin{aligned}
 s^b &= s_{TL}^b + s_{BR}^b, \quad \text{s.t.} \\
 s_{TL}^b &= y_{TL}^\bullet - (y_{TL}^{\rightarrow} + y_{TL}^{\searrow} + y_{TL}^{\downarrow} + y_{TL}^F), \\
 s_{BR}^b &= y_{BR}^\bullet - (y_{BR}^{\leftarrow} + y_{BR}^{\nwarrow} + y_{BR}^{\uparrow} + y_{BR}^F).
 \end{aligned} \tag{1}$$

This detection framework has several benefits. Most portion of object proposal windows generated by traditional proposal methods truncate target objects. However, previous methods depends on these object proposals [6, 5, 12, 17, 21] grade these windows by SVM scores only, without an intrinsic model to carefully handle the problem. A maximally scored object proposal through SVM does not guarantee that it is fitting an entire target object. This issue will be discussed in Sec. 3.4 again. In contrast, starting from a boundary out of target object, our method reaches at a terminal point with obvious stop signals in both corners like the examples in Fig. 1. Compared with previous methods solving the detection problem by a CNN-based regression [19, 24], our method adopts a more prosperous classification model with a soft-max cost rather than a regression model with a mean-square cost. Even if the model provides weak direction vectors which are quantized and have a fixed short length, our prediction for object detection becomes stronger as the bounding box is iteratively narrowed down to a target object.

3.2. Network architecture

The architecture of AttentionNet is illustrated in Fig. 2 but we drop pooling and normalization layers from the figure for simplicity. Layers from the first to the seventh convolution (also called seventh fully-connected layer) are the same layers in Chatfield *et al.*’s VGG-M network [4]. In this network, the stride and filter size are smaller at the first layer but the stride at the second convolution layer is larger, compared with Krizhevsky *et al.*’s network [16]. This is also similar to the CNN suggested by Zeiler and Fergus [28]. We adopt this model due to its superior performance on the ILSVRC classification (Top-5 error of 16.1 on a single model) without significant increase in training time. Please refer to [4] for more details of this architecture.



Figure 3. Real examples of crop-augmentation for training AttentionNet. The target instance is the right man. Dashed cyan bounding boxes are ground-truths, and the blue bounding boxes are the augmented regions. Red arrows/dots denote their ground truths.

Our detection framework requires two predictions for TL and BR, but we prefer a CNN for classification to be trained with soft-max loss. While the regression with a mean-square loss forces the model to a difficult bounding box, classification of quantized directions with soft-max loss forces the model to be maximally activated at a true direction rather than exact values of a bounding box coordinates. We therefore separate the output layers of TL and BR, because this setting enables us to use the soft-max loss with which each branch returns each prediction. These final layers are denoted by Conv8-TL and Conv8-BR in Fig. 2. Confidences of the 5 decisions at TL (BR) are computed with corresponding 5 filters of $1 \times 1 \times 4,096$ size at Conv8-TL (Conv8-BR). These two layers are followed by the final ReLU layer.

3.3. Training

To make AttentionNet operate in the scenario we devise, it is quite important to process original training images to a suitable form. During the test stage, starting from an initial test over the entire image boundary to a final decision of “stop” or “no instance”, the number of possible decision pairs is 17 ($=4 \times 4 + 1$) such as $\{\rightarrow, \searrow, \downarrow, \bullet\}_{TL} \times \{\leftarrow, \swarrow, \uparrow, \bullet\}_{BR}$ for positive regions and $\{F_{TL}, F_{BR}\}$ for negative regions. We therefore must augment the original training images into a reformed training set evenly covering these 17 cases. Fig. 3 shows real examples how we process an original training image to multiple augmented regions. We randomly generate positive regions which satisfy the following three rules.

1. A positive region must include at least 50% of the area of a target instance.

2. A positive region can include multiple instances (as the top-left example in Fig. 3), but the target instance must occupy the biggest area. Within a cropped region, the area of the target instance must be at least 1.5-times larger than that of the other instances.
3. Regions are cropped in varying aspect ratios as well as varying scales.

The second rule is important for complex instance layouts in the multiple instance scenario (to be introduced in Sec. 4). Without this rule in the scenario, a final bounding box is prone to fit multiple instances at once. In order to make AttentionNet always narrow the bounding box down to the largest instances among multiple instances, we must follow the second rule in generating positive regions. The third rule is also necessary because aspect ratio and scale change during the iterative crops in a test stage. We extract negative regions which are not overlapped with bounding boxes of a target class, or overlapped with bounding boxes of non-target classes. The negative regions also have varying aspect ratios and scales.

When we compose a batch to train the CNN, we select positive and negative regions in an equal portion. In a batch, each of the 16 ($=4 \times 4$) cases for positive regions occupies a portion of $1/(2 \times 16)$, and the negative regions occupy the remaining portion of $1/2$. The loss for training AttentionNet is an average of the two soft-max losses computed independently in TL and BR.

3.4. Verification of AttentionNet

Before we extend AttentionNet to multiple-instances, we verify the effectiveness of our top-down approach, against to the object proposal [26] based framework [12]. As studied by Agrawal *et al.* [1], strong mid-level activations in a CNN come from object “parts” that is distinctive to other object classes. Because Region-CNN based detection relies on each region score coming from the CNN activations, it is prone to focus on discriminative object “part” (e.g. face) rather than “entire object” (e.g. entire body).

To analyze this issue, we demonstrate a toy experiment of human detection. As an object proposal based method, we represent object proposals by activations from a fine-tuned CNN and score them by a SVM. Then, we choose the bounding box which has the maximum SVM score as a detection result. We compare this setting with our AttentionNet framework in Fig. 2. The Region-CNN, SVM, and AttentionNet are trained with the same data including ILSVRC’12 classification and PASCAL VOC 2007. The test images which contain a single human with a reasonable scale are selected from PASCAL VOC 2007 testset. Then, we compute the average precision.

In this experiment, the object proposal based setting shows 79.4% while AttentionNet shows 89.5%. The object

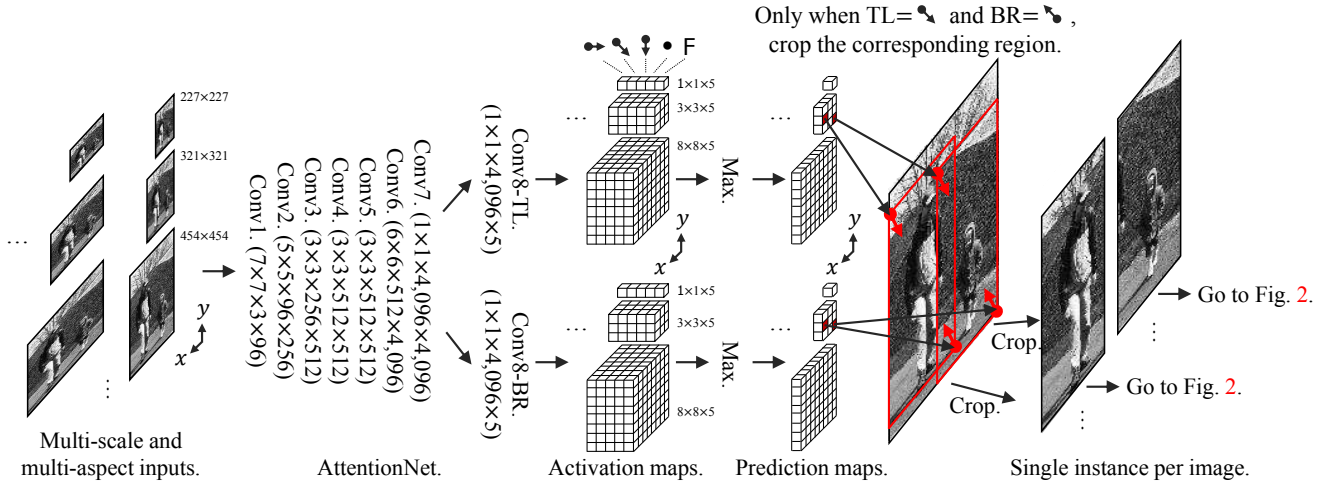


Figure 4. Extracting single-instance regions, where a single instance is included only. Multiple inputs with multiple scales/aspects are fed to AttentionNet, and prediction maps are produced. Only image regions satisfying $\{\searrow_{TL}, \swarrow_{BR}\}$ are regarded as the single-instance regions. These regions are fed to AttentionNet again for final detection. Note, the CNN here and that in Fig. 2 are the same one, not separated.

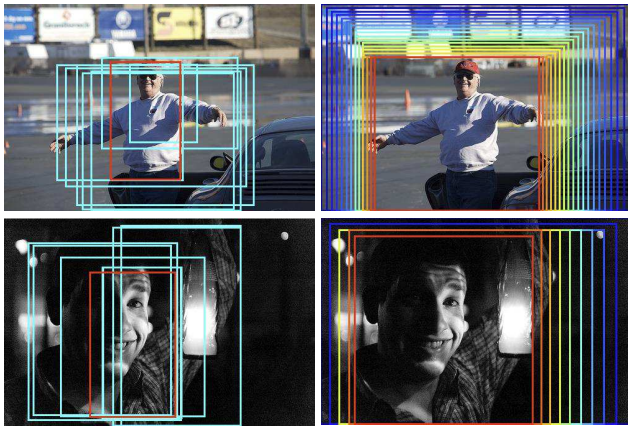


Figure 5. Real detection examples of the object proposal based method (left) and AttentionNet (right). In the left column, a red bounding box is the top-1 detected region among top-10 object proposals (cyan) with the maximum SVM score.

proposal based method shows much lower detection performance because of the relatively weak correlation between strong activation and “entire” human body. In contrast, AttentionNet reaches a terminal point well starting from a boundary out of a target object. As shown in Fig. 5, the maximally scored object proposal is prone to focus on discriminative faces rather than “entire” human body.

4. Extension to Multiple Instances

AttentionNet is trained to detect a single instance in an image and operates in this way. In this section, we introduce an efficient method to extend our detection framework to a practical situation where an image involves multiple

instances. Our straightforward solution is to propose candidate regions where only a single instance is included. We reuse AttentionNet for the single instance region proposal, therefore no separated model is necessary. We then detect each instance from each region proposal, and merge the results into a reduced number of bounding boxes followed by a final refinement procedure where AttentionNet is also reused again.

4.1. Efficient single instance region proposal

Let us assume we have an arbitrary region in an image. If we feed this region to AttentionNet, the 17 decision combinations in TL/BR are possible. Among them, only the output of $\{\searrow_{TL}, \swarrow_{BR}\}$ guarantees that the region includes entire body of a target instance with proper margins. In the other decision combinations, the region is probably truncating a target instance or does not contain an instance. This is the logic to make sure whether an entire single instance is included or not.

To boost the recall of the single instance region proposals, we place our framework in the sliding window paradigm, but we do not naively crop and feed each of sliding windows to AttentionNet. Following a technique successfully used in [25, 27], we also utilize the property that a CNN does not require fixed size input only because a fully connected layer is a convolution layer composed of filters of 1×1 size. For example, if an image 2-times larger (321×321) than a regular CNN input (227×227) is fed to AttentionNet, a spatial activation map of $3 \times 3 \times 5$ size is produced from an output layer as shown in Fig. 4. Here, the 9 activation vectors ($\in \mathbb{R}^5$) correspond to 9 local patches of 227×227 size with a stride of 32. In this way, we can speed up the sliding window method with AttentionNet.

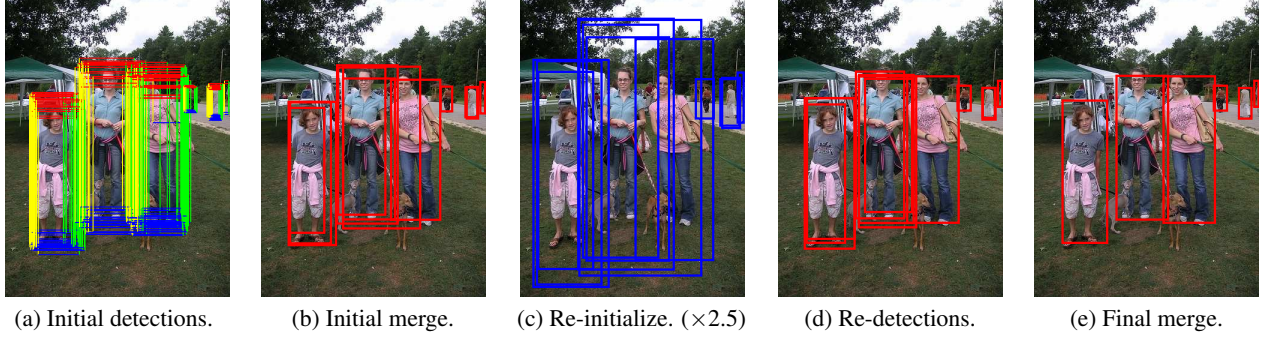


Figure 6. Real examples of our detection procedure, including initial results (a~b) and refinement (c~e). Initially detected candidates come from Fig. 4 followed by Fig. 2 are merged by an intersection over union (IoU) of 0.8. We extend each merged box to 2.5-times larger size, and feed them to Fig. 2 again. Finally we merge the second results by an IoU of 0.5.

Since instances in images are possibly diverse in aspect ratio and scale, multiple scales/aspects are also required for the sliding windows. We therefore feed multi-scale/aspect images to AttentionNet, and obtain predictions of thousands of sliding windows, as depicted in Fig. 4. Given the prediction maps, only corresponding regions of the prediction of $\{\neg_{TL}, \neg_{BR}\}$ are cropped and fed to AttentionNet again for the final detection. The number of scales/aspects can be deterministically set according to the size/aspect statistics of ground-truth bounding boxes in a training set.

4.2. Initial detection and final refinement

Each single instance region proposal produced in Sec. 4.1 is fed to AttentionNet iteratively until it meets $\{\bullet_{TL}, \bullet_{BR}\}$ or $\{F_{TL}, F_{BR}\}$. The first image in Fig. 6 shows a real example of the initial detections. These bounding boxes are merged to a decreased number by single-linkage clustering: a group of bounding boxes satisfying a minimum intersection over union (IoU) of α_0 are averaged into one with their scores of Eq. (1).

To refine the result, [11, 12] employ a bounding box regression, which finally re-localizes the bounding boxes. This is a linear regression model which maps a given feature of a bound box to a new detection window.

In our case, we can employ AttentionNet again as the role for bounding box regression for further refinement. We re-scale each bounding box in Fig. 6-(b) to a new initial window in Fig. 6-(c) by a re-scaling factor of β . These re-initialized windows are fed to AttentionNet again and result in new bounding boxes as shown Fig. 6-(d). This re-detection procedure gives us one more chance to reject false positives as well as fine localization. These bounding boxes are finally merged to final results by an IoU of α_1 .

5. Evaluation

We perform the object detection task on public datasets to verify the strength of AttentionNet. We primarily apply

our detection framework to a human detection problem and extend it to a non-human class.

Among a wide range of object classes, it is beyond question that the class “human” has taken the center stage in object detection for decades because of its wide and attractive applications. Nonetheless, human detection on uncontrolled natural images is still challenging due to severe deformations caused by pose variation, occlusion, and overlapped layout. For rigorous verification in such challenging settings, we choose human as our primary target class.

Datasets We select PASCAL VOC 2007 and 2012 [10] where images are completely uncontrolled because they are composed of arbitrary web images from Flickr. Humans in these sets are severely occluded, truncated and overlapped with diverse pose variations as well as scales. Following the standard setting in previous human detection over these sets, we use all the “trainval” images for training, and report an average precision (AP). For PASCAL VOC 2007, we rigorously use the evaluation function provided by the development toolkit. For PASCAL VOC 2012, we submit our result to the evaluation server and obtain the AP value.

Training Our framework requires only one training stage for AttentionNet. The selected datasets provide relatively small number of training images, while AttentionNet is composed of many weights to be optimized. Thus, we pre-train the model with ILSVRC 2012 classification dataset [18] and transfer the model to our target datasets. Following [1], the learning rate of the pre-trained layers (e.g. 0.001 for conv1 to conv7) is 10-times smaller than that of the re-initialized layer (e.g. 0.01 for conv8-TL and conv8-BR).

Parameters in test stage Our method requires several parameters to be determined before the test stage. We tuned all these parameters over the validation set of PASCAL VOC 2007. After we determine the proper parameter values, we

Method	Extra data	VOC'07	VOC'12
AttentionNet	ImNet	61.7	62.8
AttentionNet + Refine	ImNet	65.0	65.6
AttentionNet + R-CNN	ImNet	66.4	69.0
AttentionNet + Refine + R-CNN	ImNet	69.8	72.0
Person R-CNN + BBReg	ImNet	59.7	N/A
Person R-CNN + BBReg $\times 2$	ImNet	59.8	N/A
Person R-CNN + BBReg $\times 3$	ImNet	59.7	N/A
Felzenszwalb <i>et al.</i> '10 [11]	None.	41.9	N/A
Bourdev <i>et al.</i> '10 [2]	H3D	46.9	N/A
Szegedy <i>et al.</i> '13 [24]	VOC'12	26.2	N/A
Erhan <i>et al.</i> '14 [9]	None.	37.5	N/A
Gkioxari <i>et al.</i> '14 [13]	VOC'12	45.6	N/A
Bourdev <i>et al.</i> '14 [3]	ImNet + H3D	59.3	58.7
He <i>et al.</i> '14 [14]	ImNet	57.6	N/A
Girshick <i>et al.</i> '14 [12]	ImNet	58.7	57.8
Girshick <i>et al.</i> '14 [12]	ImNet	64.2*	N/A
Shen and Xue '14 [20]	ImNet	59.1	60.2

*Very deep model of 16 convolution layers [21] is used.

Table 1. Human detection performance on PASCAL VOC 2007/2012. ImNet denotes ILSVRC 2012 classification set. AP(%) is reported.

apply the same parameters to the whole experiments in this paper regardless of datasets. We use the parameters as follows. We set the length l of each direction vector to 30 pixels, and we limit the maximum number of iterative feed-forwards to 50 to prevent the possibility of divergence. To increase the chance for detecting large instances (e.g. an image fully filled with a face), we first re-size the average image to $(2h \times 2w)$ -size and then place an input image of $(h \times w)$ -size at the center of the magnified average image before feed-forwarding. We use 7 scales with a scale step of 2, and 3 aspect ratios of $\{\frac{h}{w} \mid 1.0, 1.5, 2.0\}$, according to the statistics of ground-truth bounding boxes in the training set. We set the merging parameters of α_0 and α_1 to 0.8 and 0.5, respectively. We set the re-scaling factor β in Fig. 6-(c) to 2.5. When we do not perform the refinement step of Fig. 6-(c~e), we set the initial merging parameter α_0 to 0.6.

We first evaluate the performance of human detection on PASCAL VOC 2007/2012. We compare our method against the recent state-of-the-art methods, and Table 1 shows the result. Without the refinement step of Fig. 6-(c~e), our method achieves 61.7% and 62.8% for each dataset. When we equip the refinement step, we achieve a new state-of-the-art score of 65.0% and 65.6% with an additional improvement of +3.3% and +2.8%. For the refinement step, we simply re-use AttentionNet for re-localization, therefore we do not need an extra model.

Similar to ours, Poselets-based methods [2, 13, 3] are also limited to a single-class (e.g. human) object detection. Our method outperforms these methods with a large margin. Though we does not include an intrinsic human

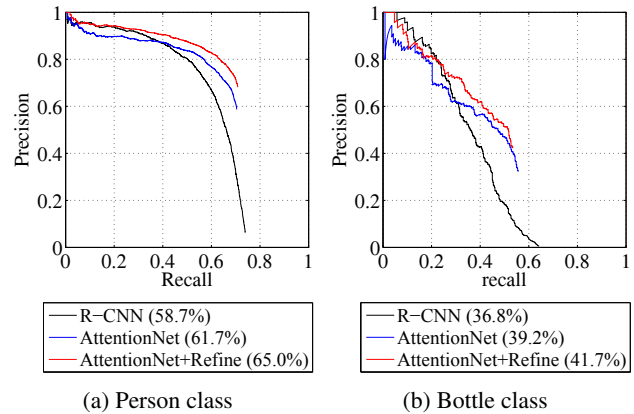


Figure 7. Precision-recall curves on PASCAL VOC 2007.

model to handle diverse poses and severe occlusions, our framework successfully converged to the window fitting the human from a window coarsely covering an entire human body. Our top-down approach is robust to the diverse human poses (Fig. 1) because our model is trained to operate in that way with carefully augmented training samples.

In the detection-by-regression manner, most similar works to ours are [24, 9]. CNN-regression models combined with a mean square error are trained to produce a target object mask [24], or bounding box coordinates [9] for the purpose of class-agnostic object proposals. Our AttentionNet clearly outperforms these methods, and it verifies the strength of the ensemble of weak directions from a promising CNN-classification model.

R-CNN [12] has been the best performing detection framework which is used for [14, 20] showing state-of-the-art performances, but our method also beats these methods which use the 8-layered architecture. Our result is even slightly better (+0.8%) than that of R-CNN [12] equipped with a 16-layered very deep network [4], which yields the top-5 error of 7.0% in ILSVRC'12 classification. However, one obvious strength of R-CNN is class-scalability while yielding high performance. Extension to multiple classes with a single AttentionNet is our primary direction.

To make the comparison completely fair, we trained R-CNN on the setting of “person”-versus-all by using the R-CNN code¹ provided by the authors. It is noted by “Person R-CNN + BBReg”. Our method still shows better results, demonstrating the significant margin of +5.3% on PASCAL VOC 2007. We also tried to iterate the bounding-box regression in R-CNN. It is noted by “BBReg $\times N$ ”. The improvement is negligible: +0.1% and +0.0% for second and third iterations. These imply the benefit of our stacked classification strategy against to the stacked regression.

Fig. 7-(a) shows the precision-recall curves of human de-

¹<https://github.com/rbgirshick/rcnn>

Method	Extra data	VOC'07	VOC'12
AttentionNet	ImNet	39.2	41.2
AttentionNet + Refine	ImNet	41.7	42.5
AttentionNet + R-CNN	ImNet	42.9	43.2
AttentionNet + Refine + R-CNN	ImNet	45.5	45.0
Bottle R-CNN + BBReg	ImNet	32.4	N/A
Bottle R-CNN + BBReg $\times 2$	ImNet	32.4	N/A
Bottle R-CNN + BBReg $\times 3$	ImNet	32.5	N/A
He <i>et al.</i> '14 [14]	ImNet	40.5	N/A
Girshick <i>et al.</i> '14 [12]	ImNet	36.8	32.6
Girshick <i>et al.</i> '14 [12]	ImNet	44.6*	N/A
Shen and Xue '14 [20]	ImNet	36.3	33.1

*Very deep model of 16 convolution layers [21] is used.

Table 2. Bottle detection performance on PASCAL VOC 2007/2012. AP(%) is reported.

tection on PASCAL VOC 2007. We plot the curve of R-CNN by using the source code from the authors. The curve clearly shows both of strength and weakness of AttentionNet. Compared with R-CNN, our method yields much high precision, but the tail is truncated early. It implies lower recall than R-CNN. This tendency comes from our hard decision strategy: AttentionNet 1) accepts a region as an initial candidate only if the both corners satisfy $\{\searrow_{TL}, \swarrow_{BR}\}$ at the same time, and 2) takes a final positive decision only if the both corners indicate “stop”. This hard criteria gives us significantly reduced number of bounding boxes with a quite strong confidence, but results in weak recall. *Our AP of 65.0% is a result only with 4,863 boxes, while R-CNN yields 58.7% with 53,624 boxes.*

We can derive a positive meaning from this observation: detection results from the two different approaches are complementary. We therefore combine the two results as in the following. Because our bounding boxes are more confident, we reject R-CNN bounding boxes which are overlapped to ours more than an IoU of 0.7, and add a bias to our bounding box scores to be followed by the remaining R-CNN bounding boxes. As reported in Table 1, we achieve the significantly boosted performance of 69.8% from the combination of AttentionNet and R-CNN.

One AttentionNet is currently handle only one object class yet, however its application does not limited to specific classes since it does not include any crafted class-specific model. We therefore demonstrate performance on another object class. Among the 20 classes in PASCAL VOC series, “bottle” is one of the most challenging object class. Bottle is the smallest class, occupying only 8k pixels in average, while the other classes occupy 38k pixels in average. Its appearance is also indistinct due to its transparent material. Thus, we select “bottle” and verify our method for this challenging case. We use the exactly same parameters used in human detection without any tuning.

Table 2 shows the result of bottle detection on PASCAL VOC 2007/2012. We observe a similar tendency to human detection in overall. Except R-CNN [12] equipped with very large CNN [21], our method yields the best scores in both datasets. Specifically, AttentionNet with refinement beats the previous methods by the gaps more than +9.4% in VOC 2012. Our method combined with the complementary R-CNN still gives us performance improvements.

Fig. 7-(b) shows the precision-recall curves of bottle detection on PASCAL VOC 2007. Although R-CNN shows slightly better precision in low recall, their precision steeply decreases while our precision decreases with a less slope. Our curve is truncated early with a lower recall than R-CNN. Further positive mining and bootstrapping could be possible solutions to boost recall.

6. Conclusions

In this paper, we have proposed a novel method for object detection. We adopted a well-studied classification technique into object detection and presented AttentionNet to predict weak directions towards a target object. Since we actively explorer an exact bounding box of a target object in a top-down approach, we does not suffer from the quality of initial object proposals and provide accurate object localization.

Through this work, we have two important observations. Firstly, we achieve the new state-of-the-art performance by changing the way of object detection. Second, our top-down approach is complementary to the previous state-of-the-art method using a bottom-up approach, therefore combining two approaches boosts the performance of object detection.

Limitations and future work We have two limitations in this study. Our AttentionNet is not scalable to multiple classes. However, it is a fact that AttentionNet has a potential for extension to generic object classes, because this model does not include any crafted class-specific models. The other thing is low recall. It is caused by our hard decision strategy as we stated. We believe there is a room for boosting recall. For example, we can loosen the hard decision criteria by employing a thresholding strategy to the direction scores from AttentionNet. Positive mining and bootstrapping can also be a promising candidate solutions. We also leave these issues as a future work.

7. Acknowledgement

This work was supported by the Technology Innovation Program (No. 10048320), funded by Korea government (MOTIE). This work was also supported by the National Research Foundation of Korea (No. 2010-0028680), funded by Korea government (MSIP).

References

- [1] P. Agrawal, R. B. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.
- [2] L. D. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2010.
- [3] L. D. Bourdev, F. Yang, and R. Fergus. Deep poselets for human detection. *Computing Research Repository (CoRR)*, abs/1407.0717, 2014.
- [4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of British Machine Vision Conference (BMVC)*, 2014.
- [5] Q. Chen, Z. Song, R. Feris, A. Datta, L. Cao, Z. Huang, and S. Yan. Efficient maximum appearance search for large-scale object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [6] R. G. Cinbis, J. J. Verbeek, and C. Schmid. Segmentation driven object detection with fisher vectors. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2013.
- [7] Y. L. Cun, B. Boser, J. S. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.
- [8] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. F. Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [9] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal on Computer Vision (IJCV)*, 88(2):303–338, June 2010.
- [11] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1627–1645, 2010.
- [12] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [13] G. Gkioxari, B. Hariharan, R. B. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.
- [15] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [17] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, and X. Tang. DeepID-net: Deformable deep convolutional neural networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal on Computer Vision (IJCV)*, pages 1–42, April 2015.
- [19] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2014.
- [20] Z. Shen and X. Xue. Do more dropouts in pool5 feature maps for better object detection. *Computing Research Repository (CoRR)*, abs/1409.6911, 2014.
- [21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository (CoRR)*, abs/1409.1556, 2014.
- [22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [23] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. *Computing Research Repository (CoRR)*, abs/1412.1441, 2014.
- [24] C. Szegedy, A. Toshev, D. Erhan, and G. Inc. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [25] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Computing Research Repository (CoRR)*, abs/1406.2984, 2014.
- [26] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal on Computer Vision (IJCV)*, 104(2):154–171, 2013.
- [27] D. Yoo, S. Park, J.-Y. Lee, and I. S. Kweon. Fisher kernel for deep neural activations. *Computing Research Repository (CoRR)*, 2014.
- [28] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.