# Visual Attention in Multi-Label Image Classification

Yan Luo
University of Minnesota
luoxx648@umn.edu

Ming Jiang
University of Minnesota
mjiang@umn.edu

Qi Zhao
University of Minnesota
qzhao@cs.umn.edu

## Abstract

*One of the most significant challenges in multi-label image classification is the learning of representative features that capture the rich semantic information in a cluttered scene. As an information bottleneck, the visual attention mechanism allows humans to selectively process the most important visual input, enabling rapid and accurate scene understanding. In this work, we study the correlation between visual attention and multi-label image classification, and exploit an extra attention pathway for improving multi-label image classification performance. Specifically, we propose a dual-stream neural network that consists of two sub-networks: one is a conventional classification model, and the other is a saliency prediction model trained with human fixations. Features computed with the two sub-networks are trained separately and then fine-tuned jointly using a multiple cross entropy loss. Experimental results show that the new saliency sub-network improves multi-label image classification performance on the MS COCO dataset. The improvement is consistent across various levels of scene clutteredness.*

trained with human fixations 是什么概念呢?

## 1. Introduction

Multi-label image classification is an essential computer vision task, aiming to recognize scene-level properties of an image from different aspects. Different from the extensively studied single-label image classification problem, multi-label image classification is more common and practical in real-world applications. An arbitrary image is likely to contain multiple objects and diverse information related to different visual and cognitive properties, such as appearance, emotions of human and animal, scene, interaction, viewpoint, scale, occlusion, and illumination. Therefore, one of the key problems in multi-label image classification is to capture the rich semantic information in complex and cluttered scenes [14].

To approach this problem, human visual system has developed a selective attention mechanism that allows us to effectively attend to interesting or important regions in a visually cluttered world [7]. Computational models of attention predict saliency (*i.e.*, features of importance in a scene) by mimicking such a selective attention mechanism [7]. Visual attention models have been empirically proved to be useful for various computer vision tasks, such as image retargeting [21], object recognition [25], video compression [3], tracking [16], image captioning [23], and so on. Although there are attempts to incorporate machine attention for multi-label image classification, *e.g.* [27], it is unknown to how human-like visual attention works in the context of multi-label image classification.

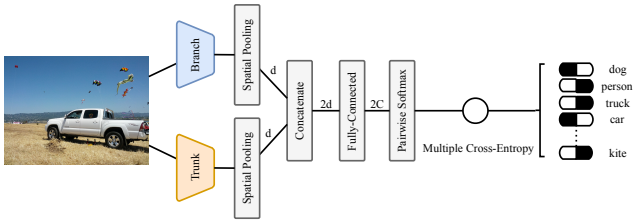



Figure 1: A dual-stream model is proposed to study the effect of visual attention on multi-label image classification. It consists of a sub-network (*i.e.*, trunk) to learn the features for classification while another sub-network (*i.e.*, branch) to be trained to predict image saliency. This model allows to quantify how much attention-related information contribute to multi-label image classification.

The objective of this work is to investigate the use of human-like visual attention in multi-label image classification. We first study the correlation between visual attention (*i.e.*, visual saliency predicting human gaze) and multi-label image classification through statistical analyses. Based on the analyses, we propose a dual-stream model to utilize human visual attention in the task of multi-label image classification. It consists of a sub-network that learns discriminative features for classification and another sub-network that learns saliency features for predicting human gaze. The proposed dual-stream model would yield its prediction based on the two types of features.

The contributions of this work are summarized as follows:

- We perform an extensive analysis to study the correlation between visual attention and multi-label image classification. The features characterizing visual attention are extracted and analyzed in the context of multi-label image classification, showing the usefulness of attention in multi-label image classification.
- We demonstrate that incorporating visual saliency model into the proposed dual-stream model can boost multi-label image classification performance.
- We provide quantitative and qualitative analyses to understand the influence of the attention-related features on multi-label image classification.

## 2. Related Works

In this section, we briefly review the related literature on multi-label classification and visual attention prediction.

**Multi-label classification**. Recently, Convolutional Neural Networks (CNNs) have made remarkable progress on single-label classification [9]. Due to its superior performance, CNN has been extensively applied to the problems of multi-label classification with localization techniques, such as region proposal [20] and localization [30]. Furthermore, recurrent neural networks (RNNs) are also widely used with CNNs to jointly characterize the semantic label dependency and relevance, such as [26, 2]. As discussed in [27], [26] may ignore the specific associations between semantic labels and the image content, and [27] introduces a framework unifying CNN and long short-term memory (LSTM), a special case of RNN, to fully exploit the spatial context in the images and associate the contents with semantic labels. A recurrent memorized-attentional module searches the attentional regions containing potential foreground objects. Interestingly, the attention module generalizes to various vision tasks, such as image captioning [29] and visual question answering [28]. Therefore, attention mechanisms have the potential to boost model performances in various vision tasks. Particularly, multi-label classification has a direct link to cognitive recognition and faces the challenge that is caused by diverse and rich context, where attention is needed. Several recent works have investigated the characteristics of the loss functions in multi-label classification task [12, 13]. As the cross entropy loss is simple, effective, and widely-used in CNNs[9, 4, 5], in this work, we adopt the multiple cross entropy (MCE) loss proposed by [13] in this work.

Similarly, numerous deep neural network models have been proposed to localize objects in images without additional human supervision. These models are learned end-to-end in a similar way as single-label image classification, while emphasizing the localization accuracy as well as the classification accuracy. Oquab et al. [18] apply a global max pooling to localize a point on objects, while Zhou et al. [34] argue that a global average pooling leads to better classification and localization performances with class activation maps (CAMs). However, neither of them targeted complex and cluttered scenes. Oquab et al. [18] scan the scene at multiple scales to find small objects, while Zhou et al. [34] only demonstrated single-label classification performance. Minh *et al*. present a recurrent network model that can sequentially attend to different locations within an image for image classification [17]. Xu *et al*. introduce hard and soft attention mechanism to generate words for salient objects in an image [29]. However, these are model-based attention mechanisms and it is still unknown that how a visual saliency learned from human fixations works in the multi-label image classification task.

**Deep learning based saliency prediction**. In the past years, we have witnessed the remarkable success of saliency modeling, especially using deep learning techniques. Kummerer et al. propose two deep saliency prediction networks: DeepGaze I [11] built upon the AlexNet [9] and DeepGaze II [10] built upon the VGG [22]. Liu et al. [15] introduce a multi-resolution CNN, which is fine-tuned over image patches centered on both attended and unattended locations. An model consisting of a deep neural network applied at two different scales are presented by Huang et al. [6]. Pan et al. [19] introduce SalGAN, a generative adversarial network for saliency prediction. In [1], an LSTM-based deep network is proposed to refine the predicted saliency map iteratively. The most architectures of these works are complicated. For simplicity and generalizability, we follow [6] to use a similar network, which is based on ResNet-50 and takes single-scale images as inputs, for saliency modeling in this work.

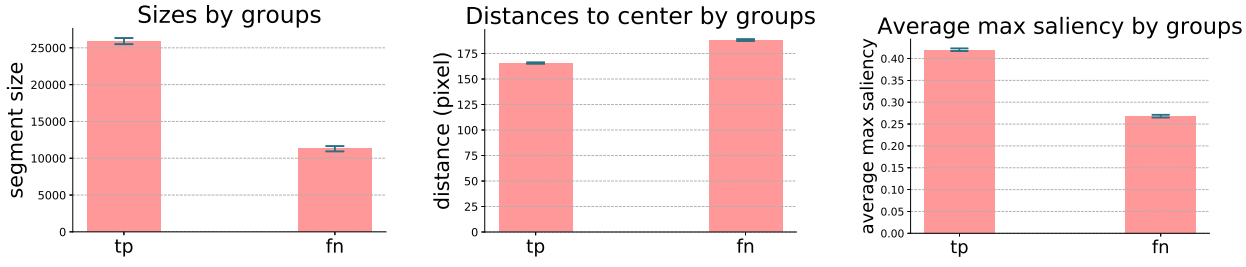
## 3. Analysis

In this section, we analyze the correlation between visual saliency and multi-label classification. To this end, we will first build a classification model, based on the ResNet-50 architecture. We present the comparison of various object-level, image-level, and class-level statistics of the model predictions with the corresponding visual saliency ground truths.

### 3.1. Baseline and Performance Metrics

In this work, for the multi-label classification task, we use a baseline ResNet-50 network with a multiple cross entropy loss [13]. This baseline model is trained on the MS COCO training set and evaluated on its validation set. We use the same metrics as the related works [27, 35], *i.e*. classwise/overall precision (C-P/O-P), classwise/overall recall (C-R/O-R), classwise/overall F1 score (C-F1/O-F1), and mean average precision (mAP).

For the saliency prediction task, we use a variant of the SALICON saliency model [6] which is also based on a




(a) (b) (c)

Figure 2: Analysis of two groups, *i.e.*, true positives (tp) and false negatives (fn) of classification, in terms of object size (a), object distance to the image center (b), and saliency value on the objects (c). Error bars indicate standard error of mean.

ResNet-50 backbone. It is trained on the SALICON [8] training set and predict saliency maps of MS COCO validation images. Normalized between 0 and 1, the values of the saliency maps indicate the likelihood that humans look at the corresponding pixels, which will be used in the following analysis.

### 3.2. Correlation between Visual Saliency and Multi-label Classification

To study the effect of various factors on a classifier's performance, we separate objects into two groups according to classification results: objects correctly recognized (*i.e.* true positives) and objects incorrectly recognized (*i.e.* false negatives). Between the two groups, we compare the objects' size, distance to image center, and the *max object saliency* (*i.e.* the maximum saliency value in an object's mask). As shown in Figure 2a, true positives (tp) are significantly larger than false negatives (fn), suggesting the larger objects are easier to classify (unpaired t-test, t=25.2469, p<0.0001). Figure 2b shows that objects closer to the image center are classified more correctly (unpaired t-test, t=-21.0477, p<0.0001). Further, correctly classified objects are also more salient (see Figure 2c, unpaired t-test, t=34.8849, p<0.0001). Another observation on Figure 2c is that visual saliency is prone to attend the objects whose labels are correctly classified. The objects whose labels are misclassified get less saliencies.

Moreover, we analyze the above statistics for each class, and compute their correlations with the class-wise mAP scores. The classification performance (mAP) is found positively correlated with the object size (Pearson's $\rho$=0.2784, p=0.0124) and negatively with the distance to image center (Pearson's $\rho$=-0.5010, <0.0001). It is also positively correlated with both the ground-truth "max object saliency" (Pearson's $\rho$=0.3651, p<0.0001) and model predictions (Pearson's $\rho$=0.3276, p=0.0030). These findings suggest strong connections between the classification performance and an object's size, location and saliency.

## 4. Methodology

According to the analysis in Section 3, we know that the correctly classified objects are more salient than the incorrectly classified objects. This implies that the visual saliency is prone to attend the objects which should be classified as a certain class. To verify this point, we propose a dual-stream model for multi-label classification, which uses saliency information as a complementary modality to complement conventional multi-label classification models. The overall architecture of the proposed dual-stream model is shown in Figure 1.

### 4.1. Network Architecture

Similar to [35], we adopt the ResNet-50 [4] as the trunk and the branch with the fully-connected layers removed. The trunk is a stack of convolutional layers to generate the feature maps. The features generated by the trunk are concatenated with the output of the branch that has the same architecture as the trunk. The resulting features are used for the inference of multiple labels. The proposed dual-stream model is a unified framework and is trained in an end-to-end manner.

The $7 \times 7 \times 2048$ feature maps (for $224 \times 224$ input images) generated by the last block in ResNet-50 is used as input for the spatial pooling layer. We denote $\boldsymbol{I}$ as an input image at size $224 \times 224$ with ground-truth labels $\boldsymbol{y} = [y^1, y^2, \ldots, y^C]$, where $y^l \in \{0, 1\}$ and $C$ is the number of categories in the dataset. Assuming that the feature maps $\boldsymbol{X}_t$ and $\boldsymbol{X}_b$ are generated by the last block of the trunk and the branch, respectively, they would be passed to two spatial pooling layers, respectively, leading to an output vector $\boldsymbol{x}_t$ and $\boldsymbol{x}_b$. This procedure can be written as follows,

$$\boldsymbol{x}_{source} = f_{source}(\boldsymbol{I}; \theta), \quad \boldsymbol{x} \in \mathbb{R}^{2048}, \quad source = \{t, b\} \tag{1}$$

where $f_t$ (or $f_b$) is the mapping of the trunk network (or the branch network) and $\theta$ is the weights of this network. Next, the trunk resulting vector $\boldsymbol{x_t}$ would be concatenated

with the branch resulting vector $\boldsymbol{x_b}$ to yield $\boldsymbol{x}_{cat} \in \mathbb{R}^{4096}$. Then, a linear transformation would be done before passing its results to the loss criterion that evaluates the discrepancy between the predictions and the ground truths. This procedure can be formulated as follows,

$$\boldsymbol{x}_{cat} = \text{concatenate}(x_t, x_b) \tag{2}$$

$$\boldsymbol{x} = \boldsymbol{x}_{cat} \times \boldsymbol{W}, \quad \boldsymbol{W} \in \mathbb{R}^{4096 \times 2C} \tag{3}$$

where $c$ is the number of the feature channels.

The advantages of our architectural design are threefold: 1) The incorporation of two networks is consistent with the Feature Integration Theory [24] of human visual attention, as the top-down attention (learned from the classification task) and the bottom-up attention (learned from the saliency prediction task) are integrated in parallel. 2) Our design is relatively simple compared with [35, 27], without complex layers that could reduce the speed of model training and inference. 3) The main components of the proposed architecture can be extended or replaced with more advanced designs.

### 4.2. Multiple Cross Entropy

Different from single-label loss as in ImageNet dataset, multinomial logistic loss (softmax loss) cannot be used directly in the multi-label classification task. This is because exponential normalization in softmax function will increase the distance among all candidates, *i.e*. the confidences w.r.t. each category, to highlight the largest candidate. This mechanism is suitable in single label classification which expects to have only one predicted label, but it is hard to align to the nature of multi-label classification task.

In this work, we adopt Multiple Cross Entropy (MCE) [13] as the loss function. First, we will compute the confidence

$$\hat{y}_j^i = \frac{\exp(x_j^i)}{\sum_{j=0}^{1} \exp(x_j^i)} \tag{4}$$

where $x_j^i$ is the feature w.r.t. $i$-th class from the last layer, $j \in \{0, 1\}$ indicates the index of positive confidence and negative confidence w.r.t. a class, and $\hat{y}$ is the confidence. After the confidences are computed, multi-class cross entropy $\ell$ would be computed as follows

$$\ell = -\sum_i \left( y^i \log \hat{y}_1^i + (1 - y^i) \log \hat{y}_0^i \right) \tag{5}$$

where $y^i$ is the $i$-th ground-truth label, it can either be 1 or 0.

## 5. Experiments

In this section, we introduce the experimental setup and present the results of our proposed model. Qualitative examples will be presented to help understand the characteristics of the proposed model in practice.

### 5.1. Experimental Setup

**Dataset**. MS COCO [14] is well-known for its rich contextual information and widely-used for multi-label classification. On the other hand, SALICON [8] is a visual saliency dataset which is built on a subset of MS COCO images to enable joint studies of image saliency and semantics. In this work, we use the MS COCO dataset for the multi-label classification task and the SALICON dataset to pre-train the sub-network for visual saliency.

**Training details**. Training of the proposed model consists of three phases: 1). the baseline model pre-trained on ImageNet is fine-tuned on MS COCO for the multi-label classification task. As a result, the resulting model would be used as classification trunk in Figure 1. Similarly, the baseline model also is fine-tuned on the SALICON dataset for saliency prediction task as the SALICON saliency [6] model did. By removing the last convolutional layer and adding a spatial pooling layer and a fully-connected layer, the resulting saliency model can be fine-tuned on MS COCO training set for the multi-label classification task. Instead of using two-scale images as the input in SALICON saliency model, we use a single-scale image as the input for simplicity. The resulting classification model would be used as the classification branch in Figure 1. 2). The features of the trunk and branch will be concatenated together and followed by a spatial pooling layer and a fully-connected layer to fulfill multi-label classification task. In this phase, the weights of the trunk and the branch are frozen to merely fine-tune the fully-connected layer. 3). The resulting model of Phase 2 is used to fine-tune on MS COCO training set again, but without freezing weights and with a smaller learning rate. The momentum and weight decay in this work are the same as the ones in [4], *i.e*. 0.9 and 0.0001. Due to the different data nature in the MS COCO and ImageNet datasets, we use a small learning rate ($1e-05$ in Phase 2 and $1e-06$ in Phase 3), instead of 0.1 in [4], to prevent training from skyrocketing caused by gradient explosion.

**Baseline model**. As we use the ResNet-50 [4] as the backbone architecture of the proposed model, we consider it as a baseline for a fair comparison. Following [18], we use global max pooling (GMP) in this work. To comprehensively evaluate the baseline model, we apply a GMP on the feature maps generated by the last building block before proceeding to the fully-connected operation.

**Multi-label classification metrics**. We use the same evaluation metrics as [35], i.e., mean average precision (mAP), classwise precision, recall, F1 (C-P, C-R, C-F1), and overall precision, recall, F1 (O-P, O-R, O-F1). C-P, C-R, C-F1,

Table 1: Performance on the validation set of MS COCO. C-P, C-R, and C-F1 stand for per-class precision, recall, and F-1 measure, respectively. O-P, O-R, and O-F1 stand for overall precision, recall, and F-1 measure, respectively. All the numbers are presented in percentage (%). RSN-50 Dual-stream Baseline is the model that uses the same architecture as the proposed Dual-stream model but is not finetuned on SALICON. In this way, it can show the contribution of saliency information to multi-label classification performance.

| | C-P | C-R | C-F1 | O-P | O-R | O-F1 | mAP |
|---|---|---|---|---|---|---|---|
| VGG MCE [13] | - | - | - | - | - | - | 70.2 |
| Weak sup[18] | - | - | - | - | - | - | 62.8 |
| CNN-RNN[26] | 66.0 | 55.6 | 60.4 | 69.2 | 66.4 | 67.8 | - |
| RGNN [33] | - | - | - | - | - | - | 73.0 |
| WELDON [2] | - | - | - | - | - | - | 68.8 |
| Multi-CNN [32] | 54.8 | 51.4 | 53.1 | 56.7 | 58.6 | 57.6 | 60.4 |
| CNN+LSTM [32] | 62.1 | 51.2 | 56.1 | 68.1 | 56.6 | 61.8 | 61.8 |
| MCG-CNN+LSTM [32] | 64.2 | 53.1 | 58.1 | 61.3 | 59.3 | 61.3 | 64.4 |
| RLSD [32] | 67.6 | 57.2 | 62.0 | 70.1 | 63.4 | 66.5 | 68.2 |
| Pairwise ranking [12] | 73.5 | 56.4 | - | 76.3 | 61.8 | - | - |
| MIML-FCN [31] | - | - | - | - | - | - | 66.2 |
| RDAR [27] | 79.1 | 58.7 | 67.4 | 84.0 | 63.0 | 72.0 | 72.2 |
| RSN-50 Baseline | 64.6 | 77.9 | 57.9 | 70.8 | 80.5 | 63.2 | 71.5 |
| RSN-50 Dual-stream Baseline | 66.7 | 75.4 | 61.3 | 71.9 | 79.4 | 65.8 | 72.1 |
| RSN-50 Dual-stream | 66.7 | 78.4 | 60.3 | 72.1 | 80.6 | 65.2 | 72.5 |

O-P, O-R, and O-F1 are defined as follows

$$\text{C-P} = \frac{1}{C}\sum_i \frac{N_i^c}{N_i^p} \quad \text{C-R} = \frac{1}{C}\sum_i \frac{N_i^c}{N_i^m} \quad \text{C-F} = \frac{2\text{C-P} \times \text{C-R}}{\text{C-P} + \text{C-R}}$$
$$\text{O-P} = \frac{\sum_i N_i^c}{\sum_i N_i^p} \quad \text{O-R} = \frac{\sum_i N_i^c}{\sum_i N_i^m} \quad \text{O-F} = \frac{2\text{O-P} \times \text{O-R}}{\text{O-P} + \text{O-R}} \tag{6}$$

where $C$ is the number of labels, $N_i^c$ is the number of images that correctly predicted for the $i$-th class, $N_i^p$ is the number of predicted images for the $i$-th label, $N_i^g$ is the number of ground truth images for the $i$-th label. More concretely, average precision is defined as follows

$$AP_i = \frac{\sum_{k=1}^{R} \hat{P}_i(k) r_i(k)}{\sum_{k=1}^{R} r_i(k)} \tag{7}$$

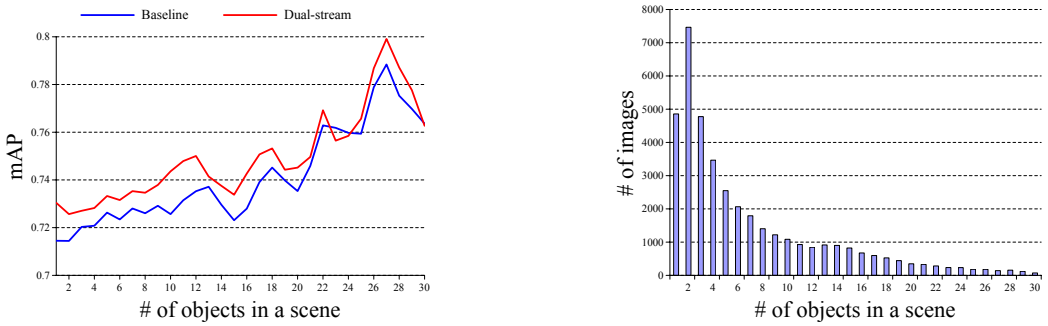
To compute the mAP, we collect all the predicted probabilities for each class of all the images. The corresponding predicted $i$-th labels over all images are sorted in descending order. The average precision of the $i$-th class is the mean of the average of precision of correctly predicted $i$-th labels. $\hat{P}_i(k)$ is the precision ranked at $k$ over all predicted $i$-th labels. $R$ denotes the number of predicted $i$-th labels. Finally, the mAP is obtained by averaging AP over all classes.

## 5.2. Performance

We compare the proposed with several state-of-the-art multi-label classification models. Experimental results are shown in Table 1. For a fair comparison, we also report a variant of the proposed model, *i.e*. the dual-stream model, which follows the same training procedure as the proposed model, but without fine-tuning on the SALICON dataset in Phase 2. By comparing the proposed model to the dual-stream model, we can quantify how much saliency information from saliency dataset contributes to multi-label classification. We can see that concatenating two RSN-50 baselines (*i.e*. RSN-50 Dual-stream Baseline) can improve the mAP to $72.1\%$ from $71.5\%$. The additional saliency information learned by the proposed RSN-50 dual-stream model would further improve the mAP to $72.5\%$. Similarly, the proposed RSN-50 dual-stream model overall achieves better performances than the RSN-50 baseline in C-P, C-R, C-F1, O-P, O-R, and O-F1. C-F1 and O-F1 of the proposed RSN-50 dual-stream model are $60.3\%$ and $65.2\%$, whereas the ones of the RSN-50 baseline are $57.9\%$ and $63.2\%$, respectively. The improvement implies that visual saliency information is helpful for multi-label classification. Then, it is interesting to know the object size in favor of the proposed dual-stream model.

## 5.3. Effects on Various Cluttered Scenes

To understand how saliency information working in different levels of cluttered scenes, we experiment the proposed dual-stream model with the images in different levels in terms of clutteredness and see how the performance varies in response to this factor. To quantify the level of clutteredness of an image, the number of objects/instances in the scene is one of the feasible indicators. Therefore,




(a) (b)

Figure 3: (a) The increase of mAPs with numbers of objects in a scene. The mAPs are smoothed by convoluting with a moving average $[0.3333, 0.3333, 0.3333]$. The start and the end of the mAPs vector are padded with the boundary value. This reveals how saliency information works with the numbers of objects. (b) The histogram of number of objects in a scene. It shows that the most images on MS COCO has less than 20 objects.

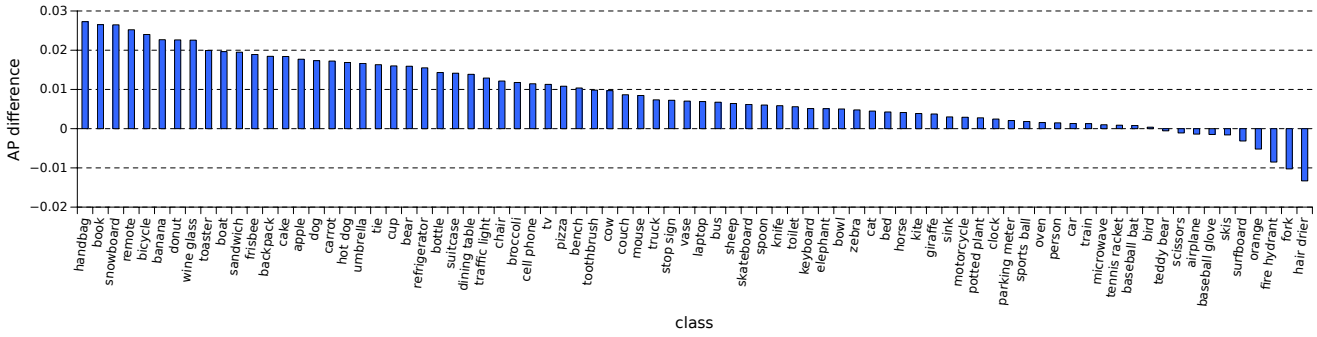


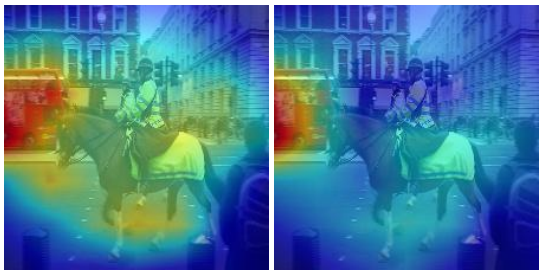

Figure 4: Improvement in AP w.r.t. each category with the proposed dual-stream model. It is sorted by the improvement and can be observed that APs of 70 categories out of 80 are improved by the proposed dual-stream model. This implies that the feature learned for visual saliency prediction is useful for multi-label image classification.

we first group the images in the validation set according to its number of objects. Then, we compute mAP w.r.t. each group of images.

Figure 3a shows the increase of the classification performance (mAP) with the number of objects in the image. The mAPs are smoothed by convoluting with a moving average $[0.3333, 0.3333, 0.3333]$. The start and the end of the mAPs vector are padded with the boundary value. It can be seen that the proposed dual-stream model consistently outperforms the baseline in images with fewer objects, while for more cluttered scenes, the performance improvement is less significant. Particularly, the additional saliency information improves the mAP from 0.7145 to 0.7304 when there is only one object in an image. This is because that saliency information captures the important regions of the object for better classification. Since the mAP is calculated upon all samples instead of an individual sample, we report the histogram of the number of objects in Figure 3b to show the distributions of the groups of images. Most images contain less than 20 objects, and their mAPs are relatively unbiased thanks to the scale of samples.

### 5.4. Categorical Performance

To compare the roles of saliency information across different object categories, we plot the improvements in AP (AP of the proposed dual-stream model − AP of the baseline) for each category. As shown in Figure 4, the additional saliency information provides more performance gain on the categories 'handbag', 'snowboard', 'wine glass', 'banana', 'donut', 'remote', and 'book', which on average occupy relatively small areas. On the other hand, the AP improvement of the category 'hair drier' is decreased, and its average size is relatively large and ranked at the 34th largest category. Therefore, we believe visual saliency improved the classification performance by localizing small but salient objects, which tend to be overlooked by conventional classification models.



(a) Bus: 0.2042 v.s. 0.5920

(b) Tie: 0.1866 v.s. 0.6666


(c) Handbag: 0.2393 v.s. 0.8325

(d) Clock: 0.3488 v.s. 0.6884

Figure 5: Class activation maps and classification confidences of the baseline (left) and the proposed dual-stream model (right). Specifically, the images are misclassified by the baseline, but correctly classified by the proposed model. This implies that saliency information helps the model locate the regions of interest w.r.t. a certain label.

### 5.5. Visualization

In this section, we visualize class activation maps (CAMs) to illustrate how neural network features are changed by the proposed dual-stream model. As introduced by Zhou [34], by combining all features maps with class-specific weights, the CAMs localize the discriminative regions for each object category, bridging the gap between image regions and semantic labels. Figure 5, shows typical examples that are misclassified by the baseline model but correctly classified by the proposed dual-stream model. As can be seen, in the dual-stream model, CAMs focus on the relevant regions of interest, to correctly classify the corresponding objects, while the baseline localizes and classify the objects incorrectly. For instance, the dual-stream model does not attend to the man and the horse in Figure 5a when classifying the bus. Similar observations can be obtained in Figures 5b-5d.

## 6. Conclusion

In this work, we analyze the correlation between visual attention and multi-label image classification. We observe that visual saliency is correlated to the results of multi-label classification, due to better localization of semantically-related regions. Inspired by the observation, we propose a dual-stream model to integrate visual attention models into multi-label image classification network. The proposed dual-stream model exploits the advantages of human visual attention data. Experimental results on the MS COCO dataset shows that the proposed dual-stream model achieves better performance to the baseline. Moreover, our analysis shows that visual saliency feature is helpful for various levels of scene clutteredness.

## Acknowledgment

This research was funded by the NSF under Grants 1849107 and 1763761, and the University of Minnesota Department of Computer Science and Engineering Start-up Fund (QZ).

## References

[1] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *arXiv preprint arXiv:1611.09571*, 2016.

[2] T. Durand, N. Thome, and M. Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4743–4752, 2016.

[3] H. Hadizadeh and I. V. Bajic. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings*

*of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015.

[7] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10):1489–1506, 2000.

[8] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] M. Kümmerer, T. S. Wallis, and M. Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

[11] M. Kmmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In *ICLR Workshop*, May 2015.

[12] Y. Li, Y. Song, and J. Luo. Improving pairwise ranking for multi-label image classification. *CoRR*, abs/1704.03135, 2017.

[13] Z. Li, W. Lu, Z. Sun, and W. Xing. Improving multi-label classification using scene cues. *Multimedia Tools and Applications*, pages 1–16, 2017.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 362–370, 2015.

[16] V. Mahadevan and N. Vasconcelos. Biologically inspired object tracking using center-surround saliency mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):541–554, 2013.

[17] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–694, 2015.

[19] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[21] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch. Automatic image retargeting. In *Proceedings of the 4th international conference on Mobile and ubiquitous multimedia*, pages 59–68. ACM, 2005.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Y. Sugano and A. Bulling. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*, 2016.

[24] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[25] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognitiona gentle way. In *Biologically motivated computer vision*, pages 251–267, 2002.

[26] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016.

[27] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin. Multi-label image recognition by recurrently discovering attentional regions. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[29] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[30] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai. Exploit bounding box annotations for multi-label object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–288, 2016.

[31] H. Yang, J. T. Zhou, J. Cai, and Y. Ong. MIML-FCN+: multi-instance multi-label learning via fully convolutional networks with privileged information. *CoRR*, abs/1702.08681, 2017.

[32] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu. Multi-label image classification with regional latent semantic dependencies. *arXiv preprint arXiv:1612.01082*, 2016.

[33] R.-W. Zhao, J. Li, Y. Chen, J.-M. Liu, Y.-G. Jiang, and X. Xue. Regional gating neural networks for multi-label image classification. In *BMVC*, 2016.

[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[35] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. *CoRR*, abs/1702.05891, 2017.