# Attention Pyramid Module for Scene Recognition

Zhinan Qiao, Xiaohui Yuan, Chengyuan Zhuang, Abolfazl Meyarian
Department of Computer Science and Engineering, University of North Texas

## Abstract

The unrestricted open vocabulary and diverse substances of scenery images bring significant challenges to scene recognition. However, most deep learning architectures and attention methods are developed on general-purpose datasets and omit the characteristics of scene data. In this paper, we exploit the Attention Pyramid Module (APM) to tackle the predicament of scene recognition. Our method streamlines the multi-scale scene recognition pipeline, learns comprehensive scene features at various scales and locations, addresses the interdependency among scales, and further assists feature re-calibration as well as the aggregation process. APM is extremely light-weighted and can be plugged into existing network architectures in a parameter-efficient manner. By integrating APM into ResNet-50, we obtain a boost of top-1 accuracy by 3.54\% on the benchmark dataset. Our comprehensive experiments demonstrate that APM achieves much improved performance comparing with the state-of-the-art attention methods using significantly less computation budget.

## Background

### Scene Recognition:

Scenery images often represent a complex view that includes multiple objects at different scales and a complicated background.

(a) ImageNet (b) Place365

## Related work & Limitations:

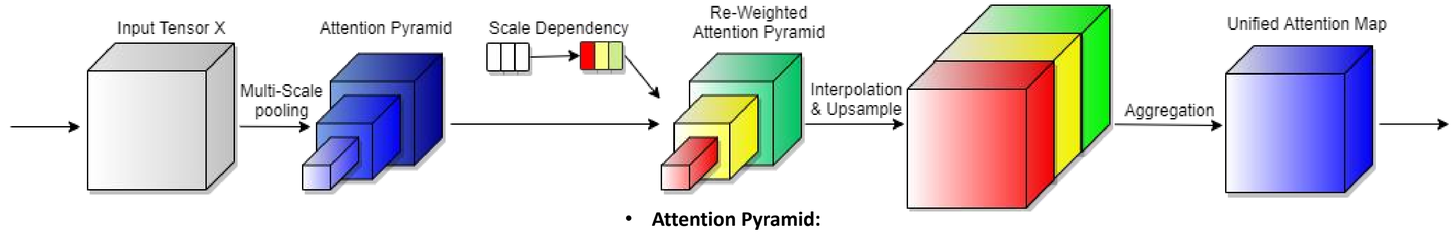
### Conventional multi-scale scene classification methods commonly follow a four-step pipeline:

Training multi-scale networks.
Extracting features .
Concatenating or summing the features.
Making the final prediction.

### Limitation:

Each level of the pyramid requires to train a separate network, these methods often face expensive computation cost, especially when the number of levels of the pyramid increases.

## Attention Pyramid Module (APM)




- **Attention Pyramid:**
  Scale-aware Attention Map Extraction.
- **Scale Dependency:**
  Learning the Weight of Scales.
- **Scale Aggregation:**
  Aggregating Re-weighted Scale-aware Attention Maps.

## Classification Results on Places365

| Network | ResNet-18 | | | ResNet-50 | | | ResNet-101 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Top-1 | Top-5 | Params | Top-1 | Top-5 | Params | Top-1 | Top-5 | Params |
| Vanilla | 53.693 | 83.778 | 11.36 M | 54.767 | 84.932 | 24.26 M | 56.471 | 86.249 | 43.25 M |
| APM | **54.978** | **84.786** | 11.37 M | **56.707** | **86.597** | 24.29 M | **56.740** | **86.770** | 43.31 M |

| | Top-1 | Top-5 | GFLOPs | Params |
|---|---|---|---|---|
| ResNet-50 | 54.767 | 84.932 | 4.12 | 24.26 M |
| GC-ResNet-50 [17] | 55.614 | 85.718 | **4.13** | 26.80 M |
| SK-ResNet-50 [14] | 56.142 | 86.274 | 4.18 | 24.85 M |
| GE-ResNet-50 [16] | 56.148 | 86.340 | 4.14 | 24.75 M |
| SE-ResNet-50 [12] | 56.162 | 86.258 | **4.13** | 26.79 M |
| CBAM-ResNet-50 [13] | 56.652 | 86.534 | 4.14 | 26.79 M |
| APM-ResNet-50 | **56.707** | **86.597** | **4.13** | **24.29** M |

**Highlight:** Our proposed module with the vanilla ResNet50 improves the performance by 3.54% top-1 classification accuracy, whereas almost no additional computations are introduced.

## Ablation Study

| | | | | | | |
|---|---|---|---|---|---|---|
| Pyramid Attention (Non-learnable) | | ✓ | ✓ | ✓ | ✓ | |
| Batch Normalization | | | ✓ | | ✓ | ✓ |
| Sigmoid | | | | ✓ | ✓ | ✓ |
| Pyramid Attention (Learnable) | | | | | | ✓ |
| Top-1 | 54.767 | 54.896 | 55.318 | 55.477 | 56.162 | **56.707** |
| Top-5 | 84.932 | 84.942 | 85.499 | 85.592 | 86.129 | **86.597** |

**Highlight:** 1) We can observe that using APM alongside all the components except the non-learnable module leads to the best result.
2) Switching a non-learnable multiscale module to APM achieved a 0.97% top-1 accuracy improvement.

## The influence of Scales of APM

| Model | ResNet-18 | | ResNet-50 | |
|---|---|---|---|---|
| | Top-1 | Top-5 | Top-1 | Top-5 |
| Vanilla | 53.693 | 83.778 | 54.767 | 84.932 |
| - $L_1$ | 54.523 | 84.838 | 56.019 | 86.011 |
| - $L_2$ | 54.636 | **84.978** | 56.482 | 86.518 |
| - $L_3$ | 54.540 | 84.767 | 56.099 | 86.403 |
| $L_1 + L_2 + L_3$ | **54.978** | 84.786 | **56.707** | **86.597** |

**Highlight:** The best result, in terms of top-1 accuracy, for ResNet18 and ResNet-50 is achieved when all three scales, L1,
L2, and L3 are involved.

## Conclusions

- We have presented a simple yet effective module, called APM for building attention pyramids inside benchmark networks and further assisting the task of scene recognition.
- The APM can be combined with any existing backbone architectures in a plug-and-play manner with marginal computation increase.
- We also experimentally demonstrated that our APM is more parameter efficient while achieving better performance against state-of-the-art attention modules.

## Contact

Zhinan Qiao
University of North Texas

Email: zhinanqiao@my.unt.edu
Website: https://github.com/ZN-Qiao/AP
Phone: (+1) 4696825915