Residual Attention Network for Image Classification

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, Xiaoou Tang

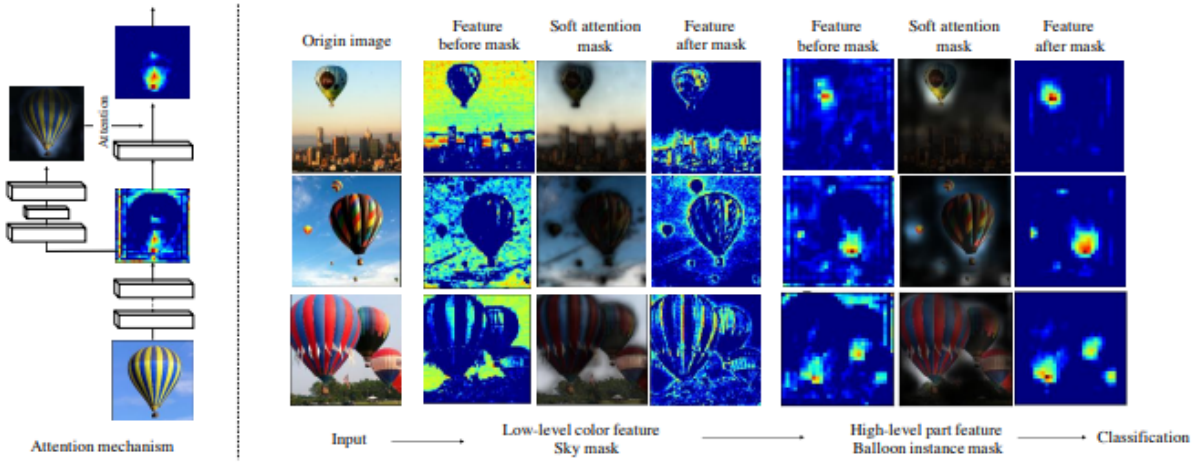
https://arxiv.org/abs/1704.06904
https://arxiv.org/pdf/1704.06904.pdf

TL;DR

"Residual Attention Network" or ResNet, is a convolutional neural network that incorporates the attention mechanism into its architecture. Attention modules are stacked to generate attention-aware features. ResNet achieves state-of-the-art object recognition performance on CIFAR-10, imageNet and CIFAR-100. 3.90%, 4.8% and 20.45% error respectively.

Introduction

Many different things draw our attention; it not only serves to select a focused location but also enhances different representations of objects at a location. ResNet incorporates the attention mechanism into a convolutional neural network to generate attention-aware features. These attention-aware features change adaptively for each layer. This mechanism brings more discriminative feature representation, brings more consistent performance, and easily incorporates deep network structures in an end-to-end training fashion. ResNet's depth can be easily stretched to hundreds of layers.



Results are made possible because of the stacked network structure, attention residual learning to optimize very deep Resnet's with hundreds of layers, and bottom-up top-down feedforward attention.

Residual Attention network

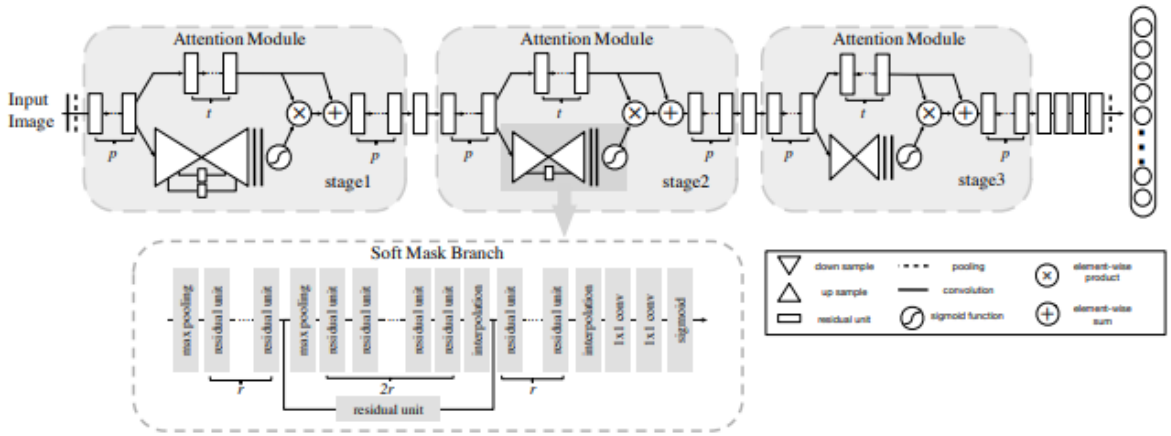


Figure 2: Example architecture of the proposed network for ImageNet. We use three hyper-parameters for the design of Attention Module: p, t and r. The hyper-parameter p denotes the number of pre-processing Residual Units before splitting into trunk branch and mask branch. t denotes the number of Residual Units in trunk branch. r denotes the number of Residual Units between adjacent pooling layer in the mask branch. In our experiments, we use the following hyper-parameters setting: $\{p = 1, t = 2, r = 1\}$. The number of channels in the soft mask Residual Unit and corresponding trunk branches is the same.

Each attention module is divided into a mask branch and a trunk branch. The trunk branch performs feature processing. ResNet uses pre-activation Residual Unit, ResNeXt and Inception to construct an Attention Module. The mask branch uses a bottom-up top-down structure to weight output features and control neurons of the trunk branch. The attention mask can serve as a feature selection. Lower layers may identify different colours, while top layers masks find instances of individual hot air balloons.

Discussion

ResNet can capture mixed attention and is an extensible CNN. Different attention modules capture different types of attention to guide feature learning. Attention Modules can also be combined to form larger network structures. ResNet also allows training very deep networks.