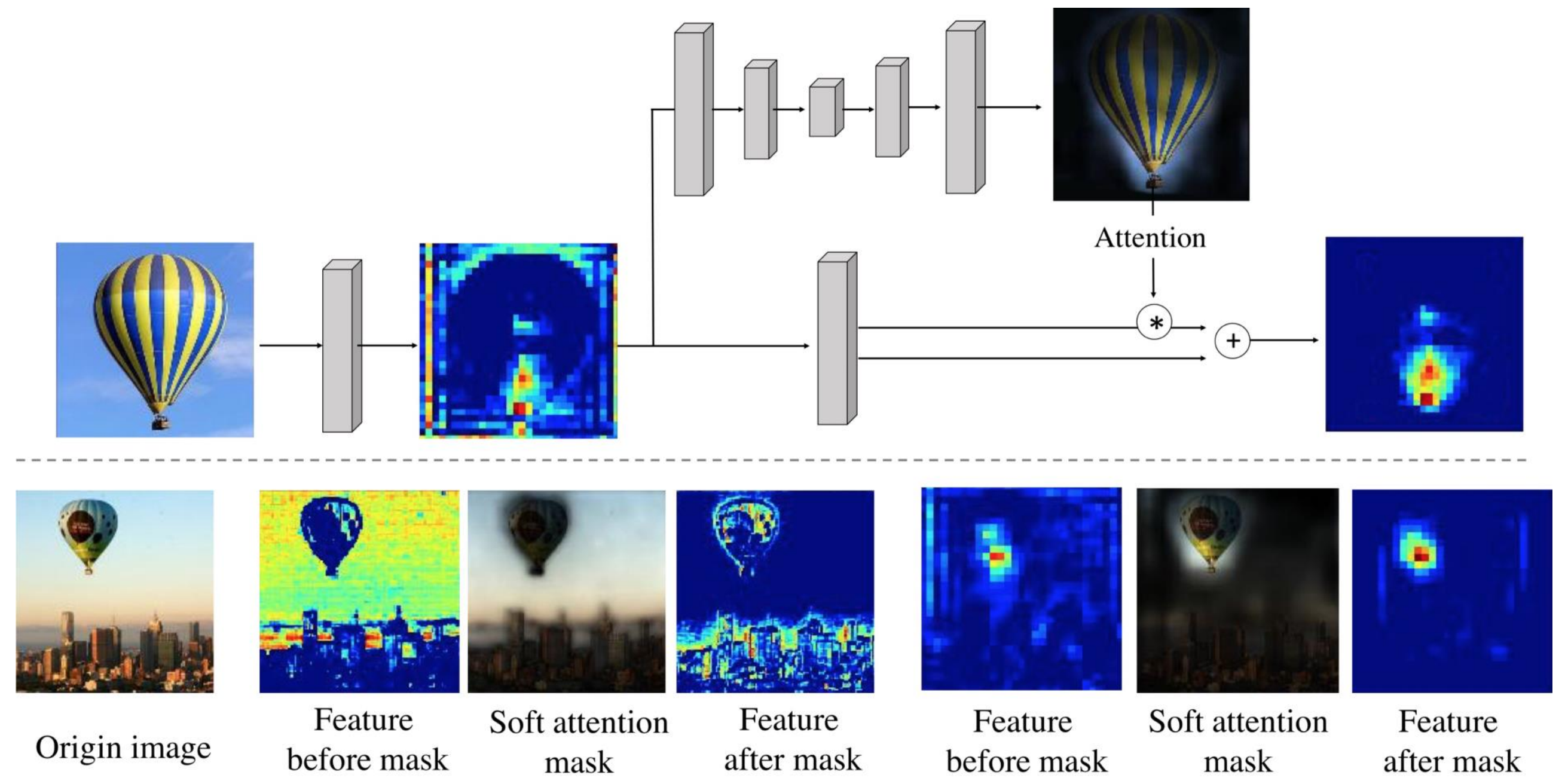
# Residual Attention Network for Image Classification

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Chen Li
Honggang Zhang, Xiaogang Wang, Xiaoou Tang
SenseTime Group Limited, Tsinghua University,
The Chinese University of Hong Kong, Beijing University of Posts and Telecommunications

AttentionNet

IEEE 2017 Conference on Computer Vision and Pattern Recognition

## 1. Introduction

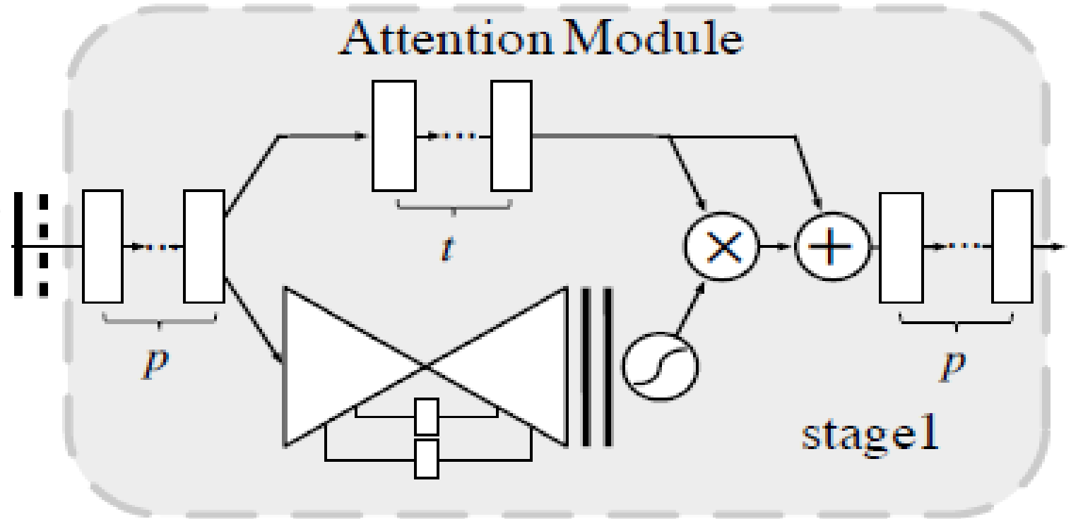


- Top: an example shows the interaction between features and attention masks.
- Bottom: example images illustrate that different features have different corresponding attention masks in our network.

## 2. Contributions

- We propose attention module which can mimic bottom-up fast feedforward process and top-down attention feedback in a single feedforward process.
- We propose attention residual learning mechanism to optimize very deep Residual Attention Network with hundreds of layers.
- Residual Attention Network can be viewed as attaching multiple mask branches (from different attention module) to trunk branches which can be arbitrary structures (resnet, resnext, inception and so on).

## 3.1 Residual attention module



- Trunk branch learns the discriminative features.
- Attention branch not only serves as a feature selector during forward inference, but also as a gradient update filter during back propagation.
- Basic unit can be any state-of-the-art convolution modules such as residual unit, resnext and inception.

## 3.2 Attention residual learning

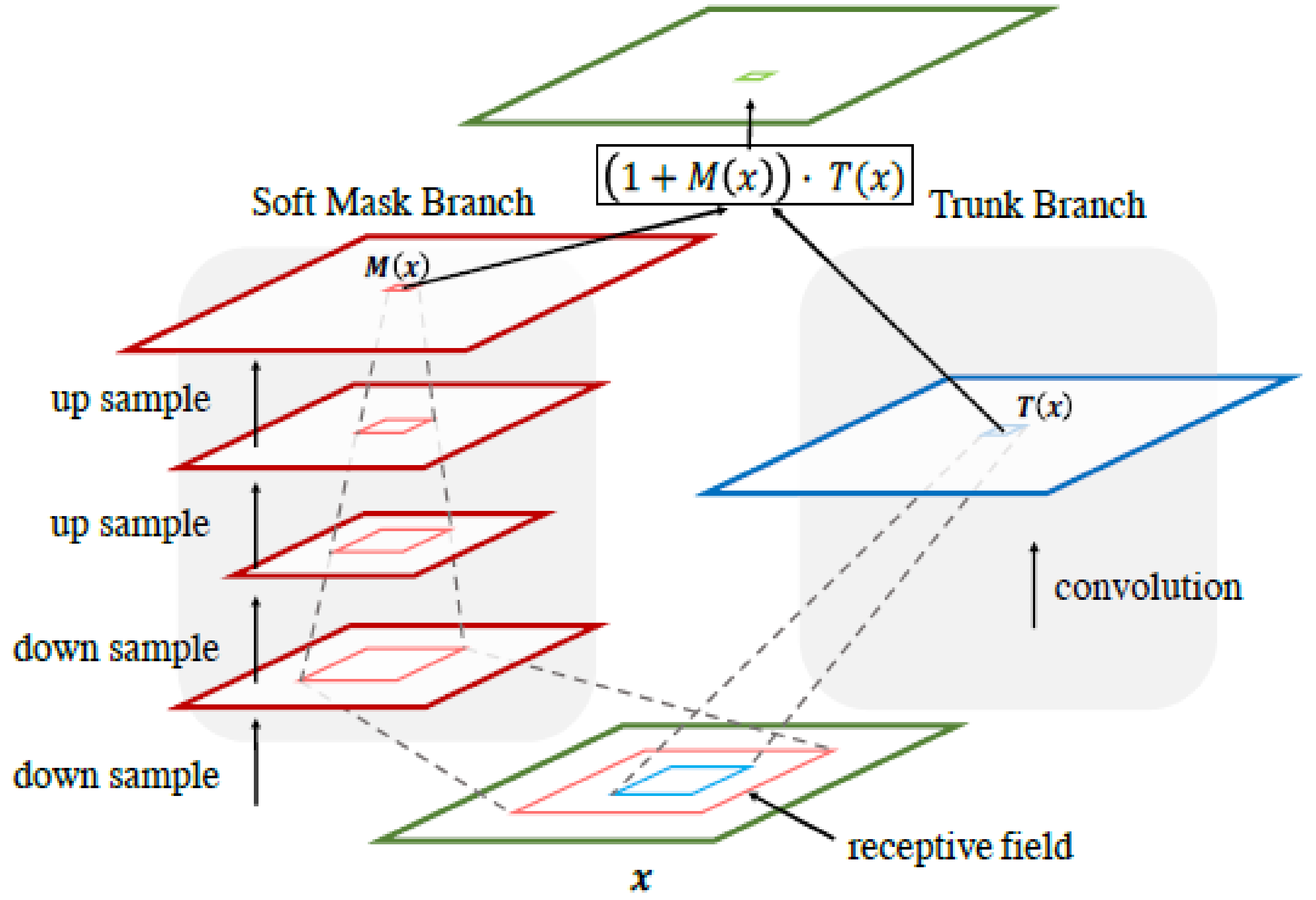
$$H_{i,c}(x) = \left(1 + M_{i,c}(x)\right) * F_{i,c}(x)$$

- M(x) ranges from [0, 1], with M(x) approximating 0, H(x) will approximate original features F(x).

| Network | ARL Top-1 err | NAL Top-1 err |
|---|---|---|
| Attention-56 | **5.52%** | 5.89% |
| Attention-92 | **4.99%** | 5.35% |
| Attention-128 | **4.44%** | 5.57% |
| Attention-164 | **4.31%** | 7.18% |

- Naive attention learning will lead to severe degradation of both useful and useless information.
- Attention residual learning can keep good properties of original features, but also gives them the ability to bypass soft mask branch and forward to top layers to weaken mask branch's feature selection ability.
- Stacked Attention Modules can gradually refine the feature maps.

## 3.3 Soft mask branch



- The Encoder-Decoder structure can effectively capture global information discussed extensively in the previous works of attention mechanism.

## 4.1 Mask influence evaluation

- We compare Attention-56 with ResNet-152 to evaluate the mask influence to the single crop performance on ImageNet.
- Attention-56 network achieves better performance with only 52% parameters and 56% FLOPs compared with ResNet-152.

| Network | Params | FLOPs | Top-1 err | Top-5 err |
|---|---|---|---|---|
| ResNet-152 | $60.2\times 10^{6}$ | $11.3\times 10^{9}$ | 22.16% | 6.16% |
| **Attention-56** | $\mathbf{31.9\times 10^{6}}$ | $\mathbf{6.3\times 10^{9}}$ | **21.76%** | **5.9%** |

## 4.2 Basic unit evaluation

- We use different basic unit to construct our Residual Attention Network to verify that Attention Network can be applied to any advanced basic unit.
- The AttentionNeXt-56 network performance is the same as ResNeXt-101 while the parameters and FLOPs are significantly fewer than ResNeXt-101.

| Network | Params | FLOPs | Top-1 err | Top-5 err |
|---|---|---|---|---|
| ResNeXt-101 | $44.5\times 10^{6}$ | $7.8\times 10^{9}$ | 21.2% | 5.6% |
| **AttentionNeXt-56** | $\mathbf{31.9\times 10^{6}}$ | $\mathbf{6.3\times 10^{9}}$ | **21.2%** | **5.6%** |
| Inception-ResNet-v1 | - | - | 21.3% | 5.5% |
| **AttentionInception-56** | $\mathbf{31.9\times 10^{6}}$ | $\mathbf{6.3\times 10^{9}}$ | **20.36%** | **5.29%** |

## 4.3 Comparisons with advanced network

- We compare our Attention-92 evaluated using single crop on the ImageNet with state-of-the-art algorithms.

| Network | Params | FLOPs | Top-1 err | Top-5 err |
|---|---|---|---|---|
| ResNet-200 | $64.7\times 10^{6}$ | $15.0\times 10^{9}$ | 20.1% | 4.8% |
| Inception-ResNet-v2 | $31.9\times 10^{6}$ | $6.3\times 10^{9}$ | 19.9% | 4.9% |
| **Attention-92** | $\mathbf{51.3\times 10^{6}}$ | $\mathbf{10.4\times 10^{9}}$ | **19.5%** | **4.8%** |