# Enhanced soft attention mechanism with an inception-like module for image captioning

# 图像标题生成任务中携带类Inception模块的增强软注意力机制

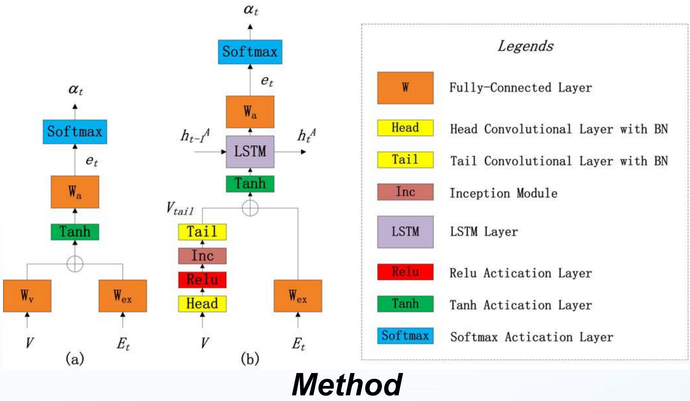
Zheng Lian, Haichang Li, Rui Wang, Xiaohui Hu

2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), 9-11 November 2020

{lianzheng2017, haichang, wangrui, hxh}@iscas.ac.cn

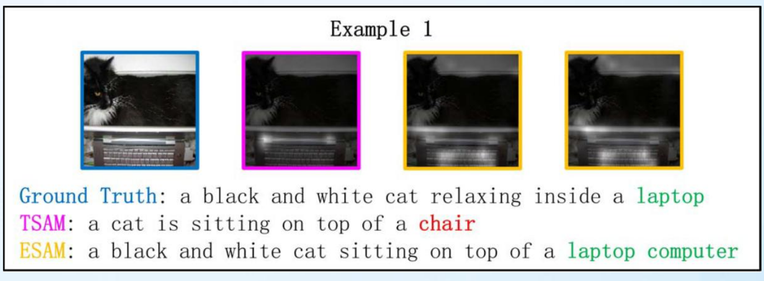
## *Introduction & Motivation*

- Image captioning is a multimodal task connecting computer vision and natural language processing.
- Traditional Soft Attention Mechanism (TSAM) assigns a weight to a certain region by only taking as input its own feature maps and some external characteristics, which will lead to unreasonable weight distribution due to the lack of surrounding information related to the region.




## *Method*

- In this paper, we propose an *Enhanced Soft Attention Mechanism* (ESAM), which improves the architecture of the TSAM. Instead of a single layer perceptron, we implement the transformation of the regional features through an inception-like module, which can capture additional information from surrounding regions.
- We further add an Attention LSTM to process the attended features, which can catch the previous attention distribution and provide better representations of attention weights to the followed fully connected layer.




## *An Example*

- The caption model with TSAM generates the wrong word “chair” rather than the ground truth “laptop”. This is due to the fact that the TSAM does not consider the information of adjacent regions, which leads to a relatively gathered attention distribution.
- With the help of the inception-like module, our ESAM balances the attention weights between adjacent regions. The attention map produced by our ESAM covers almost the whole “laptop” and then leads the language module to generate a more reasonable descriptive sentence.