Attention methods in Multimodal Image Fusion: From Human Perception to Sensing Applications

Dissertation

submitted in partial fulfillment of the requirements
for the degree of

Bachelor and Master of Technology

by

Satyam Mohla
Roll No: 150010028

under the guidance of

Prof. Subhasis Chaudhari

and

Prof. Biplab Banerjee

Department of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, India

June 2020

Abstract

Computer vision has attracted a lot of attention from multiple domains and researchers in the past decade due to its ability to perform better than humans in various tasks like recognition, segmentation, identification etc. A common strategy to develop these models involves a lot of training data, on which novel algorithms are trained. Researchers today have achieved superior performance on various tasks involving unimodal images, although this is still an active research area.

However, lately, with the advent of better imaging sensors, data from many different modalities is available in various domains like Remote Sensing (HSI-LiDAR DSM/HSI-MSI), Biomedical Imaging & Diagnostics (CT/MRI, CT/PET etc.), Autonomous Vehicles (Visible-D/LiDAR), Surveillance (Visible/IR) etc. This is very beneficial and has led to the beginning of multimodal computer vision algorithms which is a new research area.

These large multimodal image datasets contain a wealth of information which can assist researchers and experts in better classification and detection over unimodal methods. This is because multiple modalities provide richer representation in terms of information and generalizability due to the independent nature of modalities describing the same object/task.

However, this brings with itself two unanswered questions. Firstly, the most basic question would be whether the network is able to understand the different sources of information in a way that is expected. Secondly, how can they make maximal use of these different sources of information? The answer to this would be highly dependent on the underlying cost function and architecture of the algorithm. Our objective is thus to find a relevant architecture which would generate a rich, fused representation.

Attention has been a recent addition to the deep learning researchers toolbox, which came around 2015 and was first used in NLP (Attention is all you need), which was basically a gating mechanism which helped the network focus on task-relevant information. This gave the network a certain degree of control and selectability about what information to consider and what to discard. Similar ideas made their way to computer vision in 2018 as saliency maps, to perform neural feature selection, in which the generated attention maps highlighted essential task-relevant features.

Attention, thus a relatively novel concept in Computer Vision, has not been used much in multimodal image fusion, especially in the context of applications like Biomedical Imaging, Remote Sensing etc. The apparent reason is the difficulty in understanding the underlying representation of information and devising a meaningful fusion mechanism between different modalities. Another reason is the presence of modalities which contain multiple channels like HSI in Remote sensing and MRI in Biomedical Imaging. These multiple channels are highly correlated and huge in number.

Since each modality provides complementary information, a fused representation reduces the data. It provides a combined representation of different images, synergistically contributing to a more comprehensive understanding for human and machine perception. Thus, developing novel algorithms to obtain multimodal image fusion is of paramount importance for these applications.

In this work, we utilize attention to tackle both the questions in a from a new perspective, focusing firstly on multimodal image fusion by using attention to generate a fused rich representation to maximize accuracy. Secondly, we use attention as a tool to

quantify the importance of various streams and understand the recognition capabilities of the network bringing in aspects like explainability, robustness besides just accuracy. A lot of new ideas and methods are developed which are expanded ahead, leading to many contributions to academia.