Arcface Summary

Contents

List of Figures

Acronyms

Chapter 1

Introduction to Vision Transformers

Conventionally Convolutional Neural Networks are used when dealing images and spatial data for a while. But with the rise of Transformers they have started invading the field of Computer Vision as well and they have already conquered the field of Natural Language Processing especially the sector of machine translation.

Before discussing about Vision Transformers It's good to have an overview of the general architecture of Transformers.

1.1 A brief overview of an operation of a Transformer.

Transformer Neural Network is a novel deep learning architecture which uses the concept of attention for boost the speed and accuracy. Transformers were initially introduced for Natural Language Processing and it was able for outperform the deep learning model used by Google for Neural Machine Translation. Apart from the higher accuracy transformers have shown an incredible compatibility for parallelization.

1.1.1 Introducing the concept of Attention

Sequence to sequence Deep Learning models have shown a remarkable success in tasks like machine translation, text summarization and image captioning. In 2016, Google translate started using a Sequence to Sequence Learning model.

A typical sequence to sequence model consist of two parts called Encoder and Decoder and each of them consisted of Recurrent Neural Networks or LSTMs. In RNNs and LSTMs the memory is passed through the network as hidden states from time step to time step as shown in h1,h2 in figure1.1. And at the place where context of the sequence is passed to the Decoder from the Encoder, it is passed as the last hidden state of the final time step of the encoder portion.

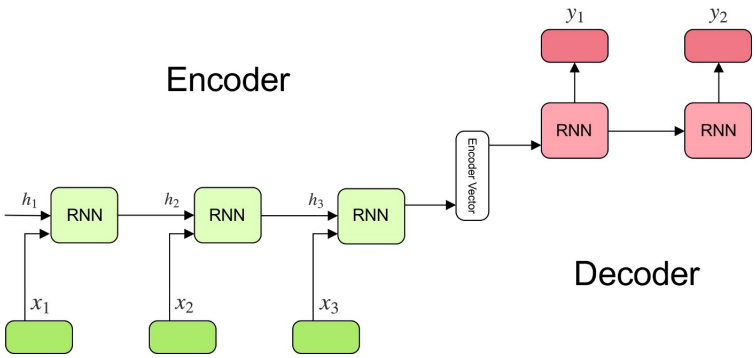

Figure 1.1: Sequence to Sequence model architecture

But, when the sequence is quite long some important information could loose while passing information via these hidden states and by using the concept of attention, Here instead of passing only the last hidden state from the Encoder, we feed all the hidden states from the start of the sequence into the Decoder. By adding this novel feature into sequence to sequence models, their performance for in working with longer sequence has improved enormously.

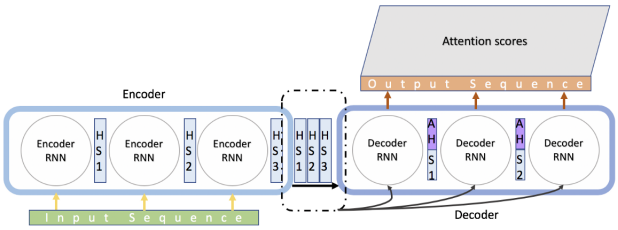

Figure 1.2: Sequence to Sequence model architecture

1.2 Ordinary Vision Transformers

1.3 Pyramid Vision Transformers

1.4 Compact Convolutional ViT

Bibliography