Understanding Google's Transformer

Every month, Computer Vision News reviews a research from our field. This month we have chosen to review **Understanding Google's Transformer - the first sequence transduction model based entirely on attention**. In fact, a number of new articles from Google Research came out last week: "Attention is All You Need", "Separable Convolutions for Neural Machine Translation" and "One Model to Learn Them All", all based on an interesting new attention unit called Transformer.

"It took only a few days to add the parsing data-set generator to Transformer and achieve very good results after a week's training"

Transformer (T2T) streamlines the creation of open source models for a wide variety of machine learning tasks, such as natural language parsing, translation, image captioning, etc., making possible much more rapid exploration of a variety of innovative ideas.

In this issue, we shall review the tools and research developments that made the Transformer possible, so even readers with little or no background can get an overview of the field. We'll explore applications and code snippets where appropriate.

The review will be divided into 6 parts: (A) First, we'll go over standard RNNs. (B) Then, we'll turn to LSTM (Long Short-Term Memory) networks. (C) We'll explain the use of RNNs for sequence transduction, that is, encoder-decoder networks. (D) We'll discuss the addition of the attention unit to encoder-decoder networks. (E) We'll discuss attempted improvements of attention units, focusing on the Neural Turing Machine. And we'll conclude with (F) the latest innovation proposed in this area -- the Transformer.

(A) RNN:

Recurrent Neural Networks (RNNs) are a class of deep learning networks created to enable a network to more easily and naturally deal with continuous data, such as text, audio and video, that is, to use the information it has gleaned about a previous word in a sentence or frame of a video to understand the next word or frame. RNN methods have produced impressive initial results in handling a variety of tasks, achieved the ability to understand free text, and even created new original sequences from scratch. Nevertheless, there is still room for improvement.

To deal with the challenge, RNN processing (like human thinking, and unlike regular neural networks) is given persistence -- by building the networks with loops in them. In the diagram on the next page, the loop stretched out. The input x0 and outputs y0 also outputting z0 (the information deemed relevant to handling the next input) in a loop to itself,