Attention Mechanisms and HAN architecture

Mohammed Reda Belfaida

October 15, 2025

Abstract

This document provides a structured exploration of Attention Mechanisms and Hierarchical Attention Network.

Table des matières

1 Attention Mechanisms (Discussions)

1.1 How can the basic self-attention mechanism be improved ?

Main Reference : Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. **A structured self-attentive sentence embedding**

The basic self-attention mechanism, which computes a single context vector by weighting encoder hidden states based on their self-referential relationships, faces significant limitations, particularly its inability to capture the full semantic range of long inputs or diverse sentence aspects, making it less effective for classification tasks like sentiment analysis. The paper "A Structured Self-Attentive Sentence Embedding" proposes an improvement by replacing this single vector with a structured self-attention approach, utilizing a bidirectional LSTM to generate hidden states H, followed by a 2D embedding matrix $M = A \times H$ with r attention heads to capture both local word interactions and broader semantic groups (e.g., coordinating conjunctions like "and"). This is enhanced by a penalization term that ensures diverse, non-overlapping attention weights, avoiding redundancy and improving interpretability, supported by two estimated matrices (W_s, W_a) and tunable hyperparameters (r, penalization strength). This method proves robust, language-agnostic, and computationally efficient compared to online-updating attention, with strong performance in tasks like textual entailment and sentiment classification, though it faces limits such as potential overfitting with high r, reliance on BiLSTM quality, and occasional over-focus on salient words (e.g., positive feedback in bad reviews), necessitating task-specific adjustments.

1.2 What are the main motivations for replacing recurrent operations with self-attention ?

Main Reference : Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need.

The paper "Attention Is All You Need" motivates the replacement of recurrent operations with self-attention by highlighting three key advantages : reduced computational complexity, enhanced parallelization, and shorter dependency path lengths. While RNNs exhibit a per-layer complexity of $O(nd)$ due to sequential processing, self-attention's $O(nd)$ complexity allows all word interactions to be computed simultaneously, leveraging GPU parallelism for faster training, despite higher memory demands for long sequences. Additionally, self-attention reduces the path length for capturing dependencies between distant words to a constant $O(1)$, compared to RNNs' $O(n)$, improving the modeling of both local and global relationships through multi-head attention. These benefits position self-attention as a robust alternative, driving the development of the Transformer architecture and enhancing performance across diverse applications.

2 Hirechical Attention Newtork (HAN)

2.1 Overview of the Architecture

Main Reference : Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics : human language technologies, pages 1480–1489, 2016.

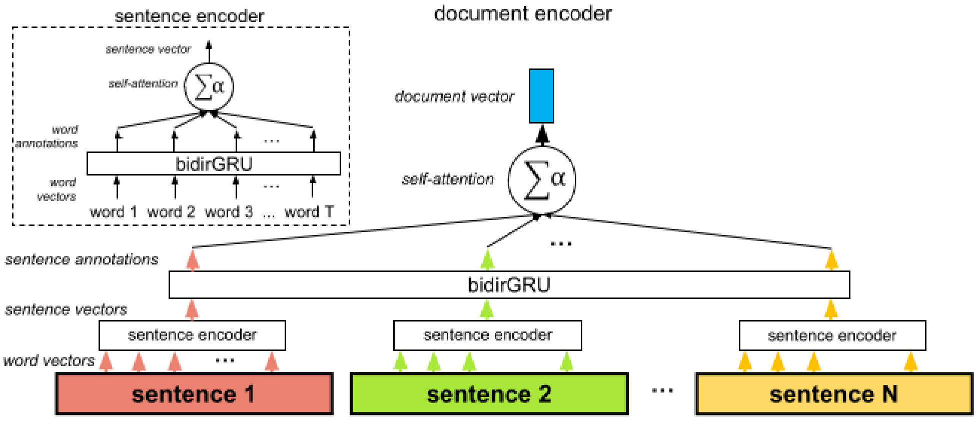

Figure 3: Hierarchical attention network.

FIGURE 1 – Overview of the architecture

An interesting application of self-attention was provided by and is illustrated in the figure above. This architecture is called HAN (for Hierarchical Attention Network). At each level, a RNN topped by a self-attention mechanism is used. First, the sentence embeddings are obtained. Then, a vector for the full document is computed from the sentence embeddings.

HAN makes sense for two reasons : first, it matches the natural hierarchical structure of documents (words → sentences → document). Second, in computing the encoding of the document, it allows the model to first determine which words are important in each sentence, and then, which sentences are important overall. By being able to re-weigh the word attentional coefficients by the sentence attentional coefficients, the model captures the fact that a given instance of a word may be very important.

2.2 Limitations

Main Reference : Bidirectional Context-Aware Hierarchical Attention Network for Document Understanding (Remy et al., 2019).

The main limitation of the original Hierarchical Attention Network lies in its context-independent sentence encoding. At the first level of the hierarchy, each sentence is encoded separately through a word-level attention mechanism that only considers the internal structure of that sentence. As a result, the attention decisions at this level are blind to inter-sentence interactions and the global meaning of the document.

This means that the same sentence will always produce the same representation, regardless of its position or role in different documents. Consequently, HAN struggles to capture long-range dependencies, discourse relations, or contextual nuances that emerge only when sentences are interpreted together.

In addition, because the document-level encoder cannot modify the already-computed sentence vectors, the model cannot adjust or refine local representations based on global context. This leads to redundant or incomplete document representations, especially when similar sentences appear multiple times.

Finally, the hierarchical structure of HAN, while linguistically motivated, introduces computational inefficiency and assumes a rigid text organization (words → sentences → document) that may not generalize well to all domains or languages.