Notes on Graph Attention Networks

Rom Parnichkun

August 18, 2022

1 Introduction

- This work introduces an attention-based architecture to perform node classification of graph-structured data
- The idea is to compute the hidden representation of each node in the graph, by attending over its neighbors, following a *self-attention* strategy.

2 Graph Attention Layer

- Input to a layer is $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, ..., \vec{h}_N\}$, $\vec{h}_i \in \mathbb{R}^F$.
- Output from a layer is $\mathbf{h}' = \{\vec{h}'_1, \vec{h}'_2, ..., \vec{h}'_N\}$, $\vec{h}'_i \in \mathbb{R}^{F'}$.
- *Weight matrix* $\mathbf{W} \in \mathbb{R}^{F' \times F}$, is initially applied to every node $\mathbf{W}\vec{h}$.
- *Self-attention* $a : \mathbb{R}^{F'} \times \mathbb{R}^{F} \to \mathbb{R}$ is then applied to pairs of neighboring nodes to compute the *attention coefficients*:

  $$e_{ij} = a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j), \tag{1}$$

  where

  $$a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j) = \text{LeakyReLU}(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_j)]), \tag{2}$$

  where $\parallel$ represents the concatenation operation and $\vec{\mathbf{a}} \in \mathbb{R}^{2F'}$ is a weight vector shared across the graph.
- Softmax is then applied to the attention coefficients for normalization

  $$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ij})}, \tag{3}$$

  where $\mathcal{N}_i$ is the set of node i's neighboring nodes.
- Layer output is computed as follows:[1]

  $$\vec{h}'_i = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\vec{h}_j\Big) \tag{4}$$

- For increased stability, they employ *multi-head attention*

  $$\vec{h}'_i = \parallel_{k=1}^{K} \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\vec{h}_j\Big) \tag{5}$$

[1]Note that they do not use different weight matrices for the key, query, and value; which is interesting.

3 Comparisons to related work

- As opposed to GCN's, GAT's allow for (implictly) assigning *difference importances* to nodes of the same neighborhood.
- Attention mechanism is applied in a shared manner to all edges in the graph. Therefore, it doesn't require the global graph structure like the adjacency matrix.
- Graph may be undirected or directed