Literature Review for the GAT Paper:  
(Word Count: 295 )

GAT architecture works very well on node classification and link predictions than preciously existing ones. It introduces a learnable weight matrix over the neighbour over the GCN architecture. Hence allowing it to capture more important information. Importantly it does not suffer from over-smoothing problem of the GCN. The key mechanism of the attention weight is derived from the Transformer model used for NLP tasks. It is an advancement from the GraphSAGE and the GCN model. The attention mechanism allows GAT to assign different importance to different neighbors based on their relevance to the target node, capturing more fine-grained and context-aware information from the graph structure. This helps us to train the model on a better scale. Unlike the GCN which takes all the information uniformly (by its normalization) , the GAT allows “relevance to particular features much more” which actually are important. GAT's ability to adaptively assign attention weights makes it more robust to variations in graph topology and node attributes, leading to better generalization and performance on real-world datasets. In experimental datasets such as Cora, Citesser, and PubMed, GAT has outperformed GCN and other state-of-the-art methods, achieving a much higher accuracy and efficiency. Traditional attention mechanisms, like those used in sequence-to-sequence models, compute a single attention distribution over the entire input sequence. However, in graph data, different parts of the graph might require different attention patterns. The multi-head attention mechanism addresses this by allowing the model to attend to different parts of the input graph simultaneously. A very important key which I find interesting is the reformulation to a particular instance of MoNet. It also Is cost efficient as it does not require any high cost tasks such as matrix multiplications and is parallelizable across all the nodes of the Graph.