Graph Representation Learning Project

When are GATs attentive and how much?

In Section 3.4 of *Graph Attention Networks(Veličković et al.)*, authors report that GATs achieve state-of-the-art performance across *Cora* and *PPI*. Since GATs were able to improve upon GCNs by a margin of 1.5% on *Cora*, the authors suggest that "assigning different weights to nodes of a same neighbourhood may be beneficial". Furthermore, the authors report a significant improvement on the inductive *PPI* benchmark and "highlight the potential of attention-based models when dealing with arbitrarily structured graphs". The authors also state that GATs "are able to assign different importances to different nodes within a neighbourhood while dealing with different sized neighbourhoods", without concrete supporting evidence they actually do that on *PPI*. Since *PPI* is a protein interaction network and since those networks are usually heterophilic, I hypothesise that in such a setting, the attention is more beneficial and is unlikely to be uniform. This hypothesis could be verified by determining the probability distribution of scores learned by attention heads. The purpose of this study is to verify the hypothesis and provide evidence for the quoted claims.

To test the hypothesis, the attention score analysis is performed on the inductive PPI model proposed by Veličković et al. This model consists of three GAT layers. GAT layer is implemented from scratch, to exactly match the paper implementation. First two GAT layers use *ELU(Clevert et al.)* activation and have 4 attention heads, each generating 256 dimensional embeddings which are concatenated. The last layer has 6 attention heads, each generating 121 dimensional class logits which are averaged. The residual connection from the paper is used in last two layers. The experimental setup is identical to the setup from the Veličković et al. The model is initialized using *Glorot(Glorot and Bengio)*, seed 42, and trained for 400 epochs by minimising cross-entropy using *Adam(Kingma and Lei Ba)*, with a learning rate of 0.005. Early stopping had patience of 100 epochs. No other regularization is applied. The **attention distribution analysis** and **neighbourhood attention analysis** is performed on the best model according to the early stopping criterion, which is based on validation loss and micro F1 score, as in the paper. Here is a link to the code repository.

Methodology

Attention scores can be treated as a discrete probability distribution over each neighbourhood. The key idea of attention distribution analysis, based on one in DGL *docs*, is to compare the neighbourhood attention distribution to the corresponding uniform distribution. The corresponding uniform distribution is a discrete uniform distribution parametrized by $\frac{1}{|N(u)|}$, for each neighbourhood $N(u)$. We are interested in characterizing the attention distribution learned by each head, for each neighbourhood. Recall that the entropy of a distribution can be interpreted as a measure of its concentration. Thus, a very small neighbourhood attention entropy indicates that the model puts the most attention on a few nodes in the neighbourhood, while the high entropy attention implies uniform attention. To compute those entropies, the model is evaluated on the test set and attention scores assigned by each head, in each layer are captured. For each neighbourhood, entropy of captured attention scores and entropy of the corresponding uniform distribution are computed. The aggregated histogram of paired entropies computed on entire test set graph is constructed and plotted. To understand the type of attention distribution and verify the hypothesis that GATs assign different importances within a neighbourhood, regardless of its size, selected neighbourhoods are visualised with corresponding attention scores, across nodes of various degrees. Since large neighbourhoods are too difficult to visualise, the focus is on neighbourhoods of nodes with degrees in range $5 - 25$ which were manually selected. Node colors in visualisations are arbitrary, but the edge thickness indicates the attention score magnitude. Thicker the edge, larger its attention is.

Results

The inductive `GAT` model attains test set F1 score of 0.9753 ± 0.0022[1], while the almost identical model in which `GAT` layers were replaced with `GCN` layers attains test set F1 score of 0.9606 ± 0.0042[1]. These results are consistent with the results reported in the paper. This suggests benefits of the attention mechanism and motivates the following further analysis.

Attention distribution analysis

The key result of this analysis is the fact that entropy of attention scores is consistently and significantly different from entropy of the corresponding uniform distribution. Such a difference implies that the attention distribution is non-uniform, even in the first layer. As the layer increases, the attention entropy becomes much sharper. This observation is depicted in Figure 1, for `head = 0`. The result for other attention heads is almost identical. This goes in favour of the hypothesis that in this setting, `GAT` learns non-uniform attention in every layer. Interestingly, it seems that all heads, in every layer, learn the attention distribution with almost identical entropy. Moreover, the difference between attention entropies of initial and subsequent layers is very large. Since the final layer attends on latent representations from previous layers, those representations could contain information useful to assign the attention scores more accurately and more sharply, which may explain this gap. The classification layer has extremely low-entropy attention, indicating that the model pays attention only to one or to a few of neighbours. In neighbourhood visualisation, we often determine that model assigns the full attention to a single node, sometimes the node itself. This is consistent with the fact that protein networks are heterophilic, so the attention head seems to put the most emphasis on node itself and a (possibly) small number of relevant neighbours.

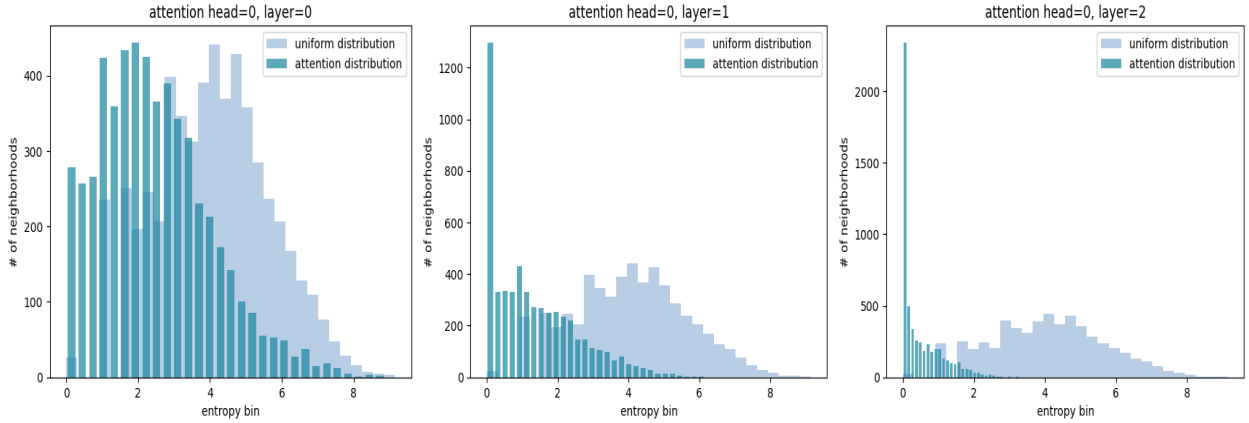


Figure 1: Neighbourhood attention entropy histogram for `head=0`, `layer=0`

Neighbourhood attention analysis

We noted that attention scores are non-uniform, even in the first layer. Figure 2 displays neighbourhoods of various sizes and the associated attention scores, assigned in the first layer. By looking at edge thickness, we can easily confirm the non-uniform attention. We know that features given to the first layer consist of positional and motif gene sets along with various immunological signatures. Since the proteins interacting with each other may differ in those structural properties, this attention pattern suggests the model is focusing on a small number of relevant neighbours, very early. This is expected for a heterophilic graph and could indeed be beneficial for learning. Moreover, the non-uniform neighbourhood attention is indeed independent of neighbourhood size.

[1] As in the paper, results were averaged over 10 runs, initialized with seed of value 42. Models on which the score was measured were selected by the already discussed early stopping strategy.

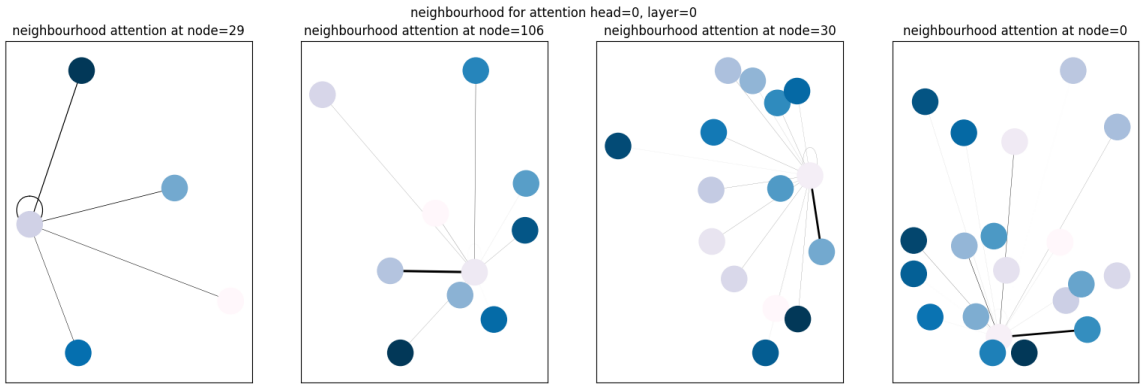
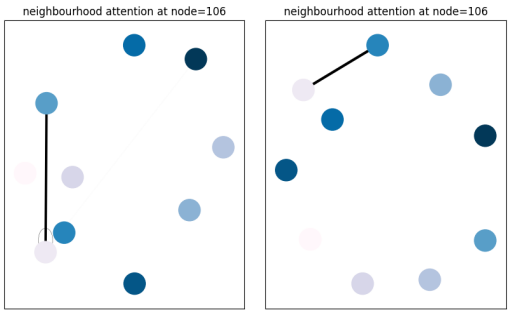


Figure 2: Neighbourhood visualisation for `head=0`, `layer=0`

In Figure 3, we look at the neighbourhood of a single node and attention scores assigned by later layers. The intermediate attention layer seems to have sharper attention than the initial layer. Note that the final layer attention is concentrated on a single node, which is consistent with the low-entropy distribution of this layer observed before. Interestingly, all selected nodes follow such an attention development pattern.



(a) Plot for `layer=1` (b) Plot for `layer=2`

node	max attention	min attention	layer
106	0.7445	0.0015	0
106	0.9114	0.0000	1
106	1.0000	0.0000	2

(c) Attention score statistics for `node=106`

Figure 3: Attention scores in neighbourhood of `node=106` across layers 1, 2 learned by `head=0`

The analysis above supports the hypothesis that `GAT`s assign different importances to different nodes within a neighbourhood, while demonstrating the ability to deal with neighbourhoods of different sizes. The analysis of attention distribution entropy provided strong evidence that `GAT`s learn non-uniform attention across various neighbourhoods, in the inductive setting on a heterophilic graph. However, the study was conducted on only one model and one dataset. To reach more conclusive statements, a similar study over a wide range of `GAT` models and datasets is necessary. Although convincing, the entropy analysis is not rigorous. For instance, by performing nonparametric statistical tests, it is possible to rigorously test uniformity of attention score distribution. However, the entropy analysis (adapted from DGL *docs*) could be useful in pruning and debugging attention models. It seems difficult to scale neighbourhood visualisation to larger graphs and more systematic and scalable method is needed. In every experiment run, attention heads seemed to learn an attention distribution with the almost identical entropy. It would be interesting to know why is this the case and are some heads redundant. According to *Brody et al.*, "increasing the number of heads strictly increases training accuracy, and thus, the expressivity". It would be interesting to know which attention distribution does their "dynamic attention" model, known as `GATv2`, learn.

Works Cited

Brody, Shaked, et al. How attentive are graph attention networks? Jan. 2022. arxiv.org/pdf/2105.14491.pdf.

Clevert, Djork-Arné, et al. Fast and Accurate Deep Network Learning By Exponential Linear Units (ELUS). Feb. 2016. arxiv.org/pdf/1511.07289v5.pdf. Accessed 14 Dec. 2022.

docs, DGL. Understand Graph Attention Network — DGL documentation. docs.dgl.ai/en/0.8.x/tutorials/models/1_gnn/9_gat.html. Accessed 14 Dec. 2022.

Glorot, Xavier, and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. 2010. proceedings.mlr.press/v9/glorot10a/glorot10a.pdf.

Kingma, Diederik, and Jimmy Lei Ba. ADAM: A Method For Stochastic Optimization. Jan. 2017. arxiv.org/pdf/1412.6980.pdf.

Veličković, Petar, et al. Graph Attention Networks. 2018. arxiv.org/pdf/1710.10903.pdf.