Graph Attention Network

ICLR 2018

Petar Veli˘ ckovic∗
Department of Computer Science and Technology
University of Cambridge

Guillem Cucurull∗
Centre de Visi´ o per Computador, UAB

Arantxa Casanova∗
Centre de Visi´ o per Computador, UAB

Adriana Romero
Montr´ eal Institute for Learning Algorithms

Pietro Lio
Department of Computer Science and Technology
University of Cambridge

Yoshua Bengio
Montr´ eal Institute for Learning Algorithms

2025.07.29
DMAIS Master Course
박세준

INDEX

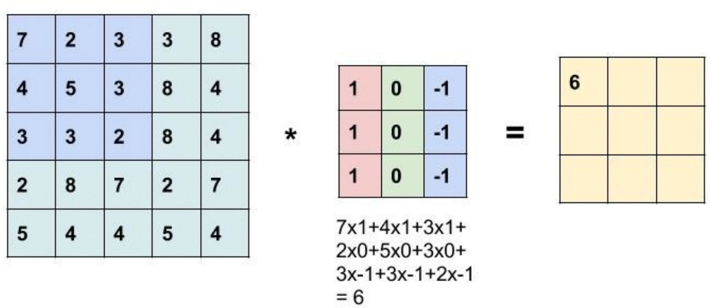
- Background
- Motivation
- GAT
- Comparisons
- Experiments
- Conclusions

TERM

- **Transductive learning**
  - Observed specific training cases to specific test cases
- **Inductive learning**
  - Observed training cases to general rules, which are then applied to the test cases.

BACKGROUND

- **Convolutional Neural Network**
  - Successfully applied to data, which representation has a grid structure
    - Image, Sequence data…
  - Not appropriate to data which can usually be represented in the form of graphs
    - 3D meshes, social networks, telecommunication networks, biological networks, brain connectomes



BACKGROUND

- **Graph Neural Network-Spectral Approaches**
  - With sprectral representation of the graphs
  - Successfully applied in the context of node classification
  - Learned filter is depend on Laplacia eigenbias, which depend on graph structure
  - Inductive learning is impossible

- **Limitation of GCN**
  - Uniform weighting to neighbor nodes
  - Same weight matrix W to every nodes
  - Strict neighbor structure

BACKGROUND

- **Graph Neural Network-Non Spectral Approaches**
    - Define convolutions directly on the graph
    - Operating groups of spatially close neighbors
    - Independent of graph structure → inductive learning possible
    - Define operator which works with different sized neighborhoods
    - Maintain the weight sharing property of CNNs

- **Limitations of GraphSAGE**
    - Data sampling
    - LSTM aggregator can’t guarantee permutation invariant

MOTIVATION

- **Attention Mechanism**
  - Calculate the attention weight and assign different weights to each
  - No sampling required
  - Weights can be calculated regardless of order, permutation invariance is guaranteed

- **Graph Attention Network**
  - Apply attention to node classification of graph structured data
  - Attention can solve both approaches problems

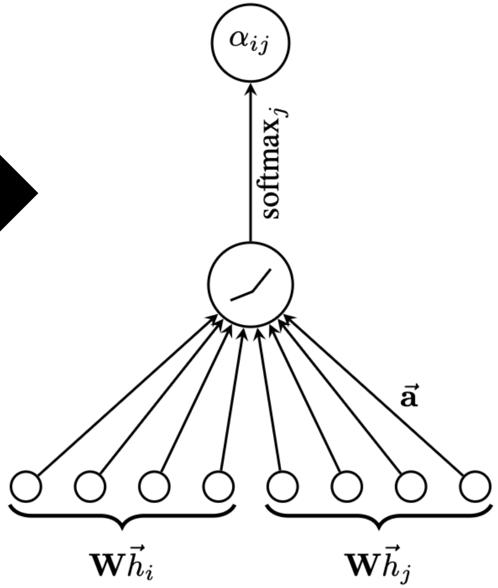
GAT

- **Graph Attention Layer**
  - Input: $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\}, \vec{h}_i \in \mathbb{R}^F$
  - Output: $\mathbf{h}' = \{\vec{h}'_1, \vec{h}'_2, \ldots, \vec{h}'_N\}, \vec{h}'_i \in \mathbb{R}^{F'}$

- **Calculate Attention Score**
  - Attention score
    - $e_{ij} = a(W\vec{h}_i, W\vec{h}_j)$
    - Indicate the importance of node $j's$ features to node i
    - Inefficient
    - Solve non spectral degree problem
    - Masked attention to calculate just neighborhood of node i
  - Softmax nomalize
    - $\alpha_{ij} = \mathrm{softmax_j}(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_{ik})}$

GAT

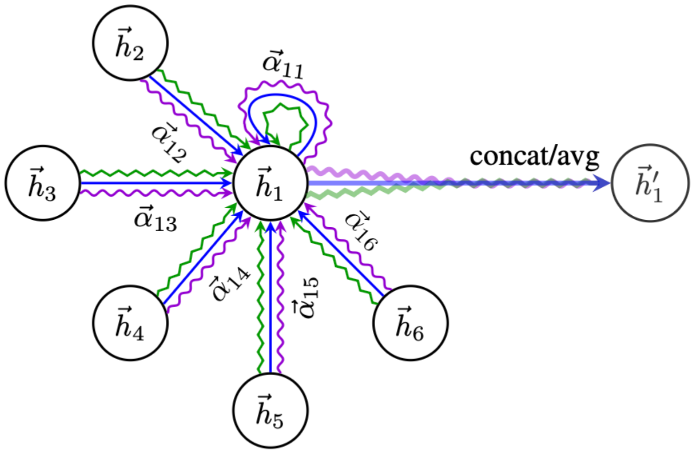
- **Calculate Attention Score**
  - Attention mechanism
    - $\alpha_{ij} = \frac{\exp\left(\text{LeackyReLU}\left(\vec{a}^T[W\vec{h}_i || W\vec{h}_j]\right)\right)}{\sum_{k\in N_i}\exp\left(\text{LeackyReLU}\left(\vec{a}^T[W\vec{h}_i || W\vec{h}_k]\right)\right)}$



- **Attain Final Feature**
  - Basic
    - $\vec{h}'_i = \sigma\left(\sum_{j\in N_i}\alpha_{ij}W\vec{h}_j\right)$
    - To serve as the final output features for every node

GAT

- **Attain Final Feature**
  - Multi head attention with concat
    - $\vec{h}_i' = \|^k{}_{k=1}\sigma\left(\sum_{j\in N_i}\alpha_{ij}^k W^k \vec{h}_j\right)$
    - Single attention head may be limited, multiple heads are used to capture a wider variety of patterns.
  - Multi head attention with average (for prediction layer)
    - $\vec{h}_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{k}\sum_{j\in N_i}\alpha_{ij}^k W^k \vec{h}_j\right)$
    - In prediction layer concatenation is no longer sensible
    - Use averaging to predict final output



COMPARISONS

- **Computationally Highly Efficient**
  - No eigendecompositions or similar costly matrix operations are required
  - GAT is good at parallelism
- **Comparison With GCN**
  - Time complexity is similar but more parallelism
  - Model allows for assigning different importances to nodes of a same neighborhood
  - Flexible structure
  - Good at inductive learning

COMPARISONS

- **Comparison With GraphSAGE**
  - No data sampling required
  - Permutation invariant is guaranted

EXPERIMENTS

- **Dataset**

Table 1: Summary of the datasets used in our experiments.

	Cora	Citeseer	Pubmed	PPI
Task	Transductive	Transductive	Transductive	Inductive
# Nodes	2708 (1 graph)	3327 (1 graph)	19717 (1 graph)	56944 (24 graphs)
# Edges	5429	4732	44338	818716
# Features/Node	1433	3703	500	50
# Classes	7	6	3	121 (multilabel)
# Training Nodes	140	120	60	44906 (20 graphs)
# Validation Nodes	500	500	500	6514 (2 graphs)
# Test Nodes	1000	1000	1000	5524 (2 graphs)

EXPERIMENTS

- **Transductive Learning**

Transductive

Method	Cora	Citeseer	Pubmed
MLP	55.1%	46.5%	71.4%
ManiReg (Belkin et al., 2006)	59.5%	60.1%	70.7%
SemiEmb (Weston et al., 2012)	59.0%	59.6%	71.7%
LP (Zhu et al., 2003)	68.0%	45.3%	63.0%
DeepWalk (Perozzi et al., 2014)	67.2%	43.2%	65.3%
ICA (Lu & Getoor, 2003)	75.1%	69.1%	73.9%
Planetoid (Yang et al., 2016)	75.7%	64.7%	77.2%
Chebyshev (Defferrard et al., 2016)	81.2%	69.8%	74.4%
GCN (Kipf & Welling, 2017)	81.5%	70.3%	**79.0%**
MoNet (Monti et al., 2016)	81.7 ± 0.5%	—	78.8 ± 0.3%
GCN-64*	81.4 ± 0.5%	70.9 ± 0.5%	**79.0** ± 0.3%
GAT (ours)	**83.0** ± 0.7%	**72.5** ± 0.7%	**79.0** ± 0.3%

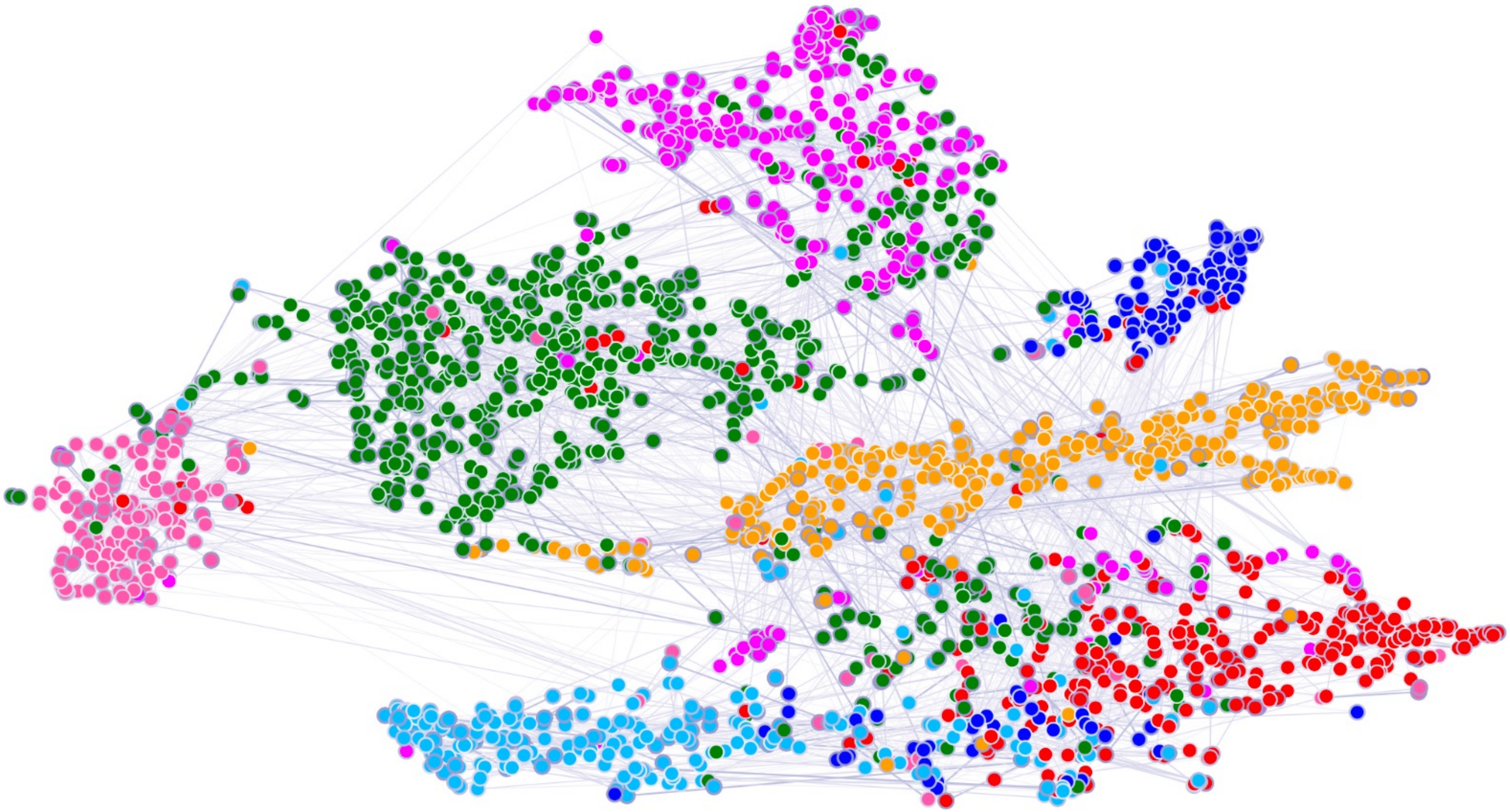
EXPERIMENTS

- **Inductive**

Inductive	
Method	**PPI**
Random	0.396
MLP	0.422
GraphSAGE-GCN (Hamilton et al., 2017)	0.500
GraphSAGE-mean (Hamilton et al., 2017)	0.598
GraphSAGE-LSTM (Hamilton et al., 2017)	0.612
GraphSAGE-pool (Hamilton et al., 2017)	0.600
GraphSAGE*	0.768
Const-GAT (ours)	0.934 ± 0.006
GAT (ours)	$\mathbf{0.973} \pm 0.002$

EXPERIMENTS

- **Visualize Cora Data**



CONCLUSION

- **Contribution**
  - Apply attention to graph
- **Performance**
  - Demonstrates better performance than all Sota models
- **Restirct**
  - Still have restrictions, so it seems that further research is needed

FINISH