Conference paper at ICLR 2018

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

GRAPH ATTENTION NETWORKS

Petar Velickovic, Guillem Cucurull, Arantxa Casanova,
Adriana Romero, Pietro Lio y Yoshua Bengio

Presentador: Jorge Díaz Ramírez
ICT3115 - Sistemas Urbanos Inteligentes
Fecha: 19/05/2022

Contenidos

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

1. INTRODUCCIÓN
2. ARQUITECTURA GAT
3. CONFIGURACIÓN DE LOS EXPERIMENTOS
4. EVALUACIÓN
5. CONCLUSIONES

Introducción

Graph attention networks [PDF] arxiv.org

P Veličković, G Cucurull, A Casanova… - arXiv preprint arXiv …, 2017 - arxiv.org

… 2.1 **GRAPH ATTENTIONAL** LAYER We will start by describing a single **graph attentional** layer,

… The particular **attentional** setup utilized by us closely follows the work of Bahdanau et al. (…

☆ Guardar 99 Citar Citado por 3363 Artículos relacionados Las 8 versiones »

[PDF] **Graph attention networks** [PDF] utdallas.edu

P Velickovic, G Cucurull, A Casanova, A Romero, P Lio… - stat, 2017 - utdallas.edu

… 2.1 **GRAPH ATTENTIONAL** LAYER We will start by describing a single **graph attentional** layer,

… The particular **attentional** setup utilized by us closely follows the work of Bahdanau et al. (…

☆ Guardar 99 Citar Citado por 5935 Artículos relacionados »

https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&authuser=2&q=GRAPH+ATTENTION+NETWORKS&btnG=

Introducción

Graph Attention Networks

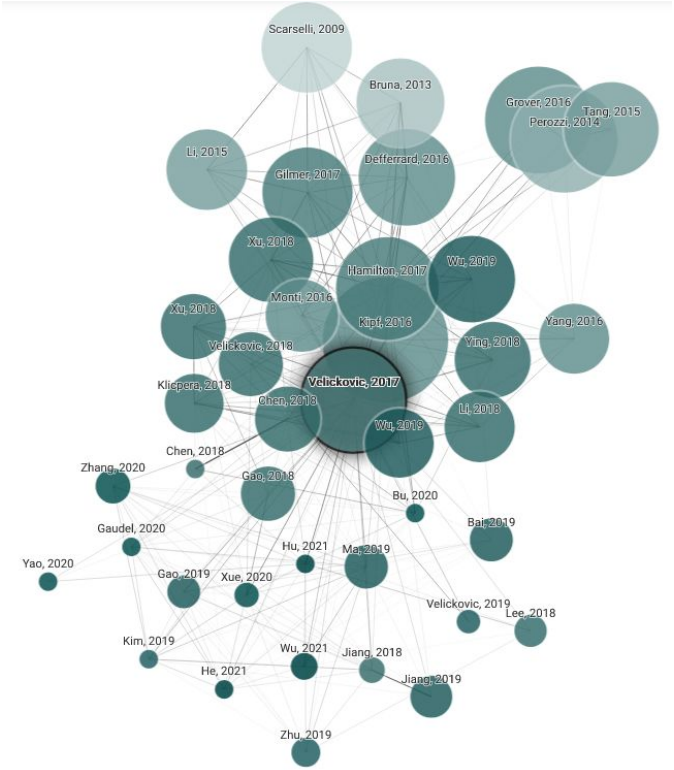
Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, P. Lio', Yoshua Bengio

ICLR 2018.

4991 Citations, 47 References

Save

We present graph attention networks (GATs), novel neural network architectures that operate on graph-structured data, leveraging masked self-attentional layers to address the shortcomings of prior methods based on graph convolutions or their approximations. By stacking layers in whi... Show more



https://www.connectedpapers.com/main/33998aff64ce51df8dee45989cdca4b6b1329ec4/Graph-Attention-Networks/graph

INTRODUCCIÓN

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

- Las redes neuronales convolucionales se han aplicado con éxito para abordar problemas como la clasificación de imágenes, la segmentación semántica o traducción automática,
  - **PERO** los datos muy estructurados (grilla)
  - Por ejemplo Mallas 3D, redes sociales, redes de telecomunicación, entre otras -> grafos

INTRODUCCIÓN

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

- Las redes neuronales de grafos (GNN) se introdujeron en Gori et al. (2005) y Scarselli et al. (2009) como una generalización de las redes neuronales recursivas.
- Los mecanismos de atención se han convertido casi en un estándar de facto en muchas tareas basadas en secuencias (Bahdanau et al., 2015; Gehring et al., 2016).
    - Cuando se utiliza para calcular una representación de una única secuencia, se suele denominar autoatención o intra-atención.

INTRODUCCIÓN

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

- Basado en lo anterior:
  - La idea es calcular las representaciones ocultas de cada nodo del grafo, atendiendo a sus vecinos, siguiendo una estrategia de autoatención.

ARQUITECTURA GAT

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Capa Atencional de grafo

- Entrada: Conjunto de características de nodos $\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\}$
  - Donde:
    - $\vec{h}_i \in \mathbb{R}^F$
    - *N* es el número de nodos y *F* es el número de características de cada nodo

- Salida: Nuevo conjunto de características de nodos $\mathbf{h}' = \{\vec{h}'_1, \vec{h}'_2, \ldots, \vec{h}'_N\}$
  - Donde:
    - $\vec{h}'_i \in \mathbb{R}^{F'}$
    - Cardinalidad *F'* potencialmente distinta a *F*

ARQUITECTURA GAT

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Capa Atencional de grafo

- Transformación Lineal: Parametrizada por la matriz de pesos $\mathbf{W} \in \mathbb{R}^{F' \times F}$
  - Aplicada a cada nodo

- Auto-atención en los nodos:
  - Mecanismo atencional compartido: $a : \mathbb{R}^{F'} \times \mathbb{R}^{F'} \rightarrow \mathbb{R}$
  - calcula los coeficientes de atención: $e_{ij} = a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$
    - Indicando la importancia de las características del nodo j para el nodo i
    - Atendiendo a cada otro nodo
  - Atención enmascarada:
    - Solo calcular e_{ij} para los nodos $j \in \mathcal{N}_i$, donde $\mathcal{N}_i$ es alguna vecindad del nodo *i* en el gráfico

ARQUITECTURA GAT

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Capa Atencional de grafo

- Auto-atención en los nodos:
  - Se normalizan todas las opciones de j usando Softmax, para hacer comparables todos los coeficientes:

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

En la práctica

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T [\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T [\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k]\right)\right)}$$

Donde, .T trasposición, || Concatenación

ARQUITECTURA GAT

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Capa Atencional de grafo

- Auto-atención en los nodos:
  - Los coeficientes de atención normalizados se utilizan para calcular una combinación lineal de las características que les corresponden, para que sirvan como características finales de salida para cada nodo:

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

ARQUITECTURA GAT

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Capa Atencional de grafo

- Auto-atención en los nodos:
  - Para estabilizar el proceso de aprendizaje del mecanismo de atención utilizan Atención multi-cabezas.
  - K mecanismos de atención independientes ejecutan la transformación de la ecuación anterior

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

Donde, || Concatenación

α_{ij}^k Son los coeficientes de atención normalizados calculados por el k-ésimo mecanismo de atención (a^k)

W^k es la matriz de pesos de la transformación lineal de entrada correspondiente

ARQUITECTURA GAT

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Capa Atencional de grafo

- Auto-atención en los nodos:
  - En especial, si se realiza atención con multicabeza en la capa final (predicción) de la red, la concatenación ya no es sensible. En su lugar, se emplea el promedio y se retrasa la aplicación de la no linealidad final (softmax o un sigmoide logístico para problemas de clasificación):

$$\vec{h}'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{k}\mathbf{W}^{k}\vec{h}_j\right)$$

ARQUITECTURA GAT

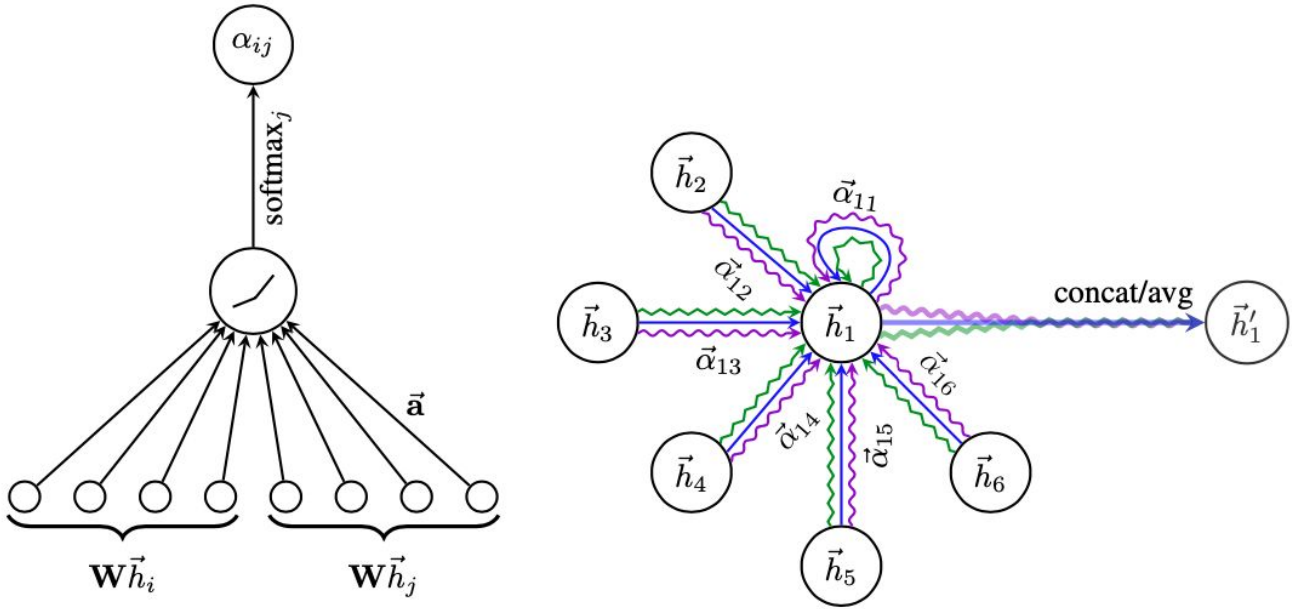


Figure 1: **Left:** The attention mechanism $a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$ employed by our model, parametrized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F'}$, applying a LeakyReLU activation. **Right:** An illustration of multi-head attention (with $K = 3$ heads) by node 1 on its neighborhood. Different arrow styles and colors denote independent attention computations. The aggregated features from each head are concatenated or averaged to obtain $\vec{h}'_1$.

CONFIGURACIÓN DE LOS EXPERIMENTOS

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Aprendizaje Transductivo
    a. Dataset Cora y Citeseer
        i. Modelo con 2 capas GAT
        ii. 1era capa con cabezas de atención K=8
        iii. Calcular F’=8 características, seguida de función de no lineal ELU (Exponential Linear Unit)
        iv. 2da capa con solo 1 cabeza de atención para calcular las clases, seguido por la activación SoftMax
        v. $\lambda = 0.0005$
        vi. Dropout=0.6 para las dos capas

CONFIGURACIÓN DE LOS EXPERIMENTOS

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Aprendizaje Transductivo
    a. Dataset Pubmed
        i. Modelo con 2 capas GAT
        ii. 1era capa con cabezas de atención K=8
        iii. Calcular F'=8 características, seguida de función de no lineal ELU (Exponential Linear Unit)
        iv. 2da capa con solo 8 cabeza de atención para calcular las clases, seguido por la activación SoftMax
        v. λ = 0.001
        vi. Dropout=0.6 para las 2 capas

CONFIGURACIÓN DE LOS EXPERIMENTOS

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

1. Aprendizaje Inductivo
    a. Datasets
        i. Modelo con 3 capas GAT
        ii. Las 1ras dos capas con cabezas de atención K=4
        iii. Calcular F'=256 características, seguida de función de no lineal ELU (Exponential Linear Unit)
        iv. 3ra capa con solo 6 cabeza de atención para calcular las clases, seguido por la activación Sigmoide Logística
        v. Sin Regularización
        vi. Sin Dropout
        vii. Skip connections en las capas atencionales intermedias
        viii. $\lambda = 0.01$ (Pubmed) y $\lambda = 0.005$ (Cora y Citeseer)
        ix. Early stopping strategy
        x. 100 épocas

Evaluación

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)

Evaluación

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Table 2: Summary of results in terms of classification accuracies, for Cora, Citeseer and Pubmed. GCN-64* corresponds to the best GCN result computing 64 hidden features (using ReLU or ELU).

	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)	$81.7 \pm 0.5\%$	—	$78.8 \pm 0.3\%$
GCN-64*	$81.4 \pm 0.5\%$	$70.9 \pm 0.5\%$	**79.0** $\pm 0.3\%$
GAT (ours)	**83.0** $\pm 0.7\%$	**72.5** $\pm 0.7\%$	**79.0** $\pm 0.3\%$

Evaluación

Table 3: Summary of results in terms of micro-averaged F_1 scores, for the PPI dataset. GraphSAGE* corresponds to the best GraphSAGE result we were able to obtain by just modifying its architecture. Const-GAT corresponds to a model with the same architecture as GAT, but with a constant attention mechanism (assigning same importance to each neighbor; GCN-like inductive operator).

	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)	$\mathbf{0.973} \pm 0.002$

Conclusiones

- (GATs), novedosa red neuronal de tipo convolucional que operan con datos estructurados en forma de grafos, aprovechando las capas de autoatención enmascaradas.

- Unión de varias técnicas (convolución y atención) en grafos

Conference paper at ICLR 2018

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

GRAPH ATTENTION NETWORKS

Petar Velickovic, Guillem Cucurull, Arantxa Casanova,
Adriana Romero, Pietro Lio y Yoshua Bengio

Presentador: Jorge Díaz Ramírez
ICT3115 - Sistemas Urbanos Inteligentes
Fecha: 19/05/2022