Data Challenge ALTeGraD
Molecule Retrieval with Natural Language Queries
Team Numpy

Adib Habbou   Steven Zheng   Teddy Alexandre

MVA - ENS Paris-Saclay

February 14, 2024

Overview

Challenge Objective

- Developing a model for retrieving molecules using natural language.
- Address the challenge of integrating text and graph information.

$\Rightarrow$ Dual Encoder : Graph + Text

Graph Encoder

- Our main idea was to use attention mechanism for graphs.
- To do that, we used 3 different types of Graph Attention Layers.

$\Rightarrow$ GAT, GATv2 and SuperGAT

GAT

By allowing nodes to attend over their neighborhood's features, the model can specify different weights to different nodes without costly operations.

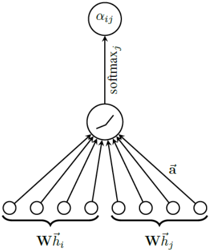

Figure: Attention mechanism utilized by GAT

$$e(h_i, h_j) = \text{LeakyReLU}\left(a^T \cdot [\mathbf{W}h_i || \mathbf{W}h_j]\right) \quad (1)$$

Multi-Head GAT

With multi-head, the features are aggregated or averaged from each head.

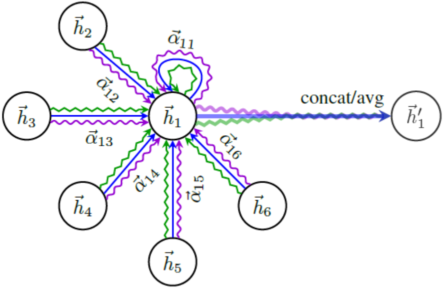

Figure: Multi-head Attention mechanism utilized by GAT

GATv2

- GAT compute a restricted form of attention called static attention.
- GATv2 is a dynamic graph attention variant, that just modifies the order of operations, is demonstrated to be strictly more expressive.

$$e(h_i, h_j) = a^T \mathrm{LeakyReLU}\left(\mathbf{W} \cdot [h_i || h_j]\right) \quad (2)$$

SuperGAT

SuperGAT leverages self-supervised tasks to predict edges, using the presence/absence of edges to inform the relationships between nodes.

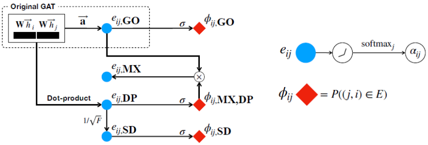

Figure: Attention mechanism utilized by SuperGAT

$$
\begin{aligned}
e_{GO}(h_i, h_j) &= \mathrm{LeakyReLU}(a^T \cdot [\mathbf{W}h_i || \mathbf{W}h_j]) \\
e_{DP}(h_i, h_j) &= \mathrm{LeakyReLU}((\mathbf{W}h_i)^T \cdot \mathbf{W}h_j) \\
e_{SD}(h_i, h_j) &= e_{DP}(h_i, h_j)/\sqrt{F} \quad \text{(F: number of features)} \\
e_{MX}(h_i, h_j) &= e_{GO}(h_i, h_j) \times \sigma(e_{DP}(h_i, h_j)) \quad (\sigma\text{: sigmoid})
\end{aligned} \tag{3}
$$

Text Encoder

- Comparison of various BERT-based models (BERT, DistilBERT, SciBERT, BioBERT, BioMegatron...)
- DistilBERT (66M): distilled version of BERT, size reduced by 40%
- SciBERT (110M): pretrained on scientific text corpus
- BioBERT (110M): pretrained on biomedical text corpus
- BioMegatron (345M): pretrained on biomedical and clinical NLP data

Ensemble Method

- Combining strengths of different models for improved accuracy.
- Averaging technique for output aggregation.

$\Rightarrow$ Alone you go faster, together you go further!

Validation Loss

Surprisingly, GAT exhibits a rapid and consistent reduction in validation loss after epoch 20, outperforming GATv2 and SuperGAT.

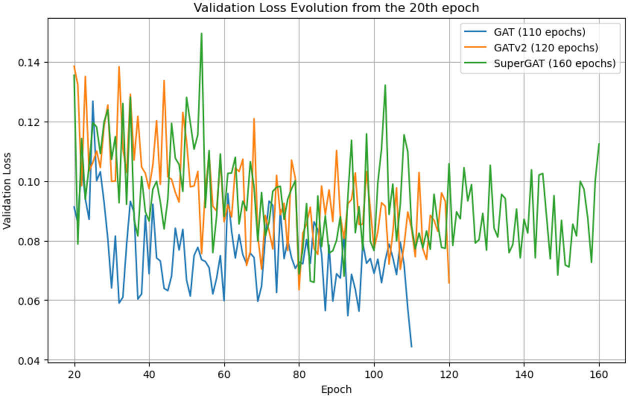

Figure: Evolution of the validation loss starting from 20th

Improving Model Efficiency with Accelerate

- Utilized the Accelerate library from Hugging Face and Automatic Mixed Precision (AMP) for optimized training.
- AMP enables dynamic switching between FP32 and FP16, enhancing computational speed (about x2.25) while maintaining precision.
- This combination significantly boosted training throughput and enabled larger batch sizes, optimizing GPU utilization.

Table: Batch Size Comparison

Graph Encoder	**Without Accelerate**	**With Accelerate**
GAT	16	96
GATv2	16	96
SuperGAT	16	80

Results

We trained our models on both the train and validation dataset.

Model	Score
GAT	0.8538
GATv2	0.8356
SuperGAT	0.8371
GAT + GATv2	0.9009
GATv2 + SuperGAT	0.8988
GAT + SuperGAT	0.8977
GAT + GATv2 + SuperGAT	0.9387

Table: Predictions scores for different models

Challenges

- GPU availability and memory constraints
- Difficulty of adapting the masking strategy
- A lot of unsuccessful experiments (Layers, Activation Function...)

Figure: Rank and score on the public Leaderboard

Key Takeaways

- Demonstrated the effectiveness of Graph Attention Networks (GAT, GATv2, SuperGAT) in capturing complex molecular structures.
- Employed DistilBERT for efficient processing of natural language descriptions, balancing performance and computational efficiency.
- Utilized the Accelerate library from Hugging Face and Automatic Mixed Precision (AMP) to optimize training pipeline.
- The ensemble method combining GAT, GATv2, and SuperGAT outperformed individual models, showcasing the power of aggregation.
- Highlighted the importance of contrastive learning in aligning textual and graphical representations for accurate retrieval tasks.

References I

- Carl Edwards ChengXiang Zhai, Heng Ji (2021). “”'Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries”” . In: *"2021 Conference on Empirical Methods in Natural Language Processing, pages 595–607"*.
- Dongkwan Kim, Alice Oh (2021). “”'How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision”” . In: *arXiv:2204.04879, Published as a conference paper at ICLR*.
- Hoo-Chang Shin Yang Zhang, Evelina Bakhturina Raul Puri Mostofa Patwary Mohammad Shoeybi Raghav Mani (2020). “”'BioMegatron: Larger Biomedical Domain Language Model”” . In: *"arXiv:2010.06060"*.
- Iz Beltagy Kyle Lo, Arman Cohan (2019). “”'SciBERT: A Pretrained Language Model for Scientific Text”” . In: *"arXiv:1903.10676"*.
- Jacob Devlin Ming-Wei Chang, Kenton Lee Kristina Toutanova (2019). “”'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”” . In: *"arXiv:1810.04805"*.

References II

- Jinhyuk Lee Wonjin Yoon, Sungdong Kim Donghyeon Kim Sunkyu Kim Chan Ho So and Jaewoo Kang (2019). “”BioBERT: a pre-trained biomedical language representation model for biomedical text mining””. In: *”Bioinformatics, Volume 36, Issue 4, Pages 1234–1240”*.
- Petar Veličković Guillem Cucurull, Arantxa Casanova Adriana Romero Pietro Liò Yoshua Bengio (2018). “”Graph Attention Networks””. In: *arXiv:1710.10903, Published as a conference paper at ICLR*.
- Shaked Brody Uri Alon, Eran Yahav (2022). “”How Attentive are Graph Attention Networks?””. In: *arXiv:2105.14491, Published as a conference paper at ICLR*.
- Victor SANH Lysandre DEBUT, Julien CHAUMOND Thomas WOLF (2020). “”DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter””. In: *”arXiv:1910.01108”*.

Q&A

Thank you for your attention!

Time for questions...