Molecule Retrieval with Natural Language Queries

Sofiane Ezzehi[1] **and Bastien Le Chenadec**[1]

[1]École des Ponts ParisTech

Contribution Statement

1 Introduction

The goal of this challenge is to retrieve molecules from a database using natural language queries. Each sample in the dataset is constituted of a ChEBI description of a molecule, which is a text describing its structure and properties, and an undirected graph representing the molecule with embeddings for each node. The embeddings are pre-computed using the Mol2Vec algorithm [1]. Given a textual query, the goal is to retrieve the molecule that best matches the query. The evaluation metric is the label ranking average precision score (LRAP) which is equivalent to the mean reciprocal rank (MRR) in our case.

The challenging part of this task is to find a way to combine two very different modalities : texts and graphs. One way to achieve this is to use contrastive learning : one model encodes the text and the other encodes the graph. The two encoders are then trained to project similar samples close to each other in the embedding space. This approach has been shown to be effective in many tasks [2, 3].

2 Data

3 Method

3.1 Graph Attention Networks

[4]

3.2 DiffPool

[5]

3.3 Language modelling

4 Training

5 Results

REFERENCES

[1] Sabrina Jaeger, Simone Fulle, and Samo Turk. Mol2vec: Unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, 2018. URL https://pubs.acs.org/doi/10.1021/acs.jcim.7b00616.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *arXiv (Cornell University)*, 2 2020. URL http://export.arxiv.org/pdf/2002.05709.

[3] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SIMCSE: Simple Contrastive Learning of Sentence Embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1 2021. doi: 10.18653/v1/2021.emnlp-main.552. URL https://doi.org/10.18653/v1/2021.emnlp-main.552.

[4] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Píetro Lió, and Yoshua Bengio. Graph attention networks. *arXiv (Cornell University)*, 2 2018. doi: 10.17863/cam.48429. URL https://arxiv.org/pdf/1710.10903.pdf.

[5] Zhitao Ying, Jiaxuan You, Christopher J. Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *neural information processing systems*, 31: 4805–4815, 12 2018. URL https://arxiv.org/pdf/1710.10903.pdf.