Geometriai és topologikus mélytanulás

Házi Feladat Dokumentáció

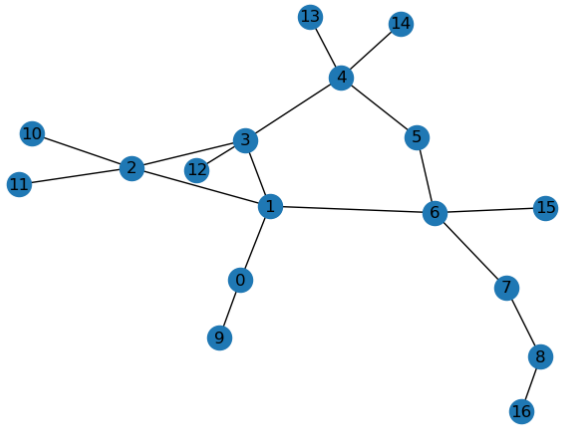
Wittmajer Dávid

Bevezetés

A házi feladat célja molekuláris tulajdonságok regressziós előrejelzése gráf-alapú gépi tanulási módszerek alkalmazásával. A kísérletek során a QM7b és QM9 kvantumkémiai benchmark adatsorokat használtam fel, melyek közül végül a QM9-re helyeztem a hangsúlyt. A kezdeti vizsgálatok a HOMO és LUMO energiák – vagyis a legmagasabb betöltött és a legalacsonyabb betöltetlen molekulapályák energiái – predikciójára irányultak, a későbbiekben pedig a vizsgálatot több célváltozóra is kiterjesztettem.

Adathalmazok

Mindkét adathalmaz molekulák gráfreprezentációit és az ezekhez tartozó jellemzőket tartalmazza, és kifejezetten molekuláris gépi tanulási modellek összehasonlító tesztelésére, azaz benchmarkolására készült. A QM7b esetében azonban hiányoznak a csúcsokra (atomokra) és élekre (kötésekre) vonatkozó jellemzők, ami jelentősen megnehezítette volna a gráfneurális hálókon alapuló módszereim alkalmazását. Emiatt végül a QM9 adathalmazra koncentráltam, amely részletesebb szerkezeti információkat biztosít.



1. Figure: Egy molekula gráfja a QM9 adathalmazból

A QM9 amellett, hogy részletesebb szerkezeti információkat tartalmaz, jóval nagyobb is: mintegy 130 ezer molekulagráfot foglal magában. Ezek a gráfok átlagosan 18 csúcsot (atomot) és 37 élt (kötést) tartalmaznak. Minden példához 19 különböző molekuláris tulajdonság (célváltozó) tartozik. A teljes adathalmazt 80–10–10%-os arányban osztottam fel tanító, validációs és teszt halmazokra.

Alkalmazott módszerek

Embedding

Először a Node2Vec algoritmust szerettem volna kipróbálni és összehasonlítani gráf-neurális hálókkal, azonban ez a megközelítés elsősorban egyetlen nagy gráf csúcspontjainak beágyazására alkalmas. Sok kisméretű, egymástól független gráf esetén a használata nehézkes, mivel előbb az összes molekulagráfot egy nagy, nem összefüggő gráffá kell összeilleszteni, majd ezen kell betanítani a modellt. Ezután a kis gráfokhoz tartozó csúcspontok embeddingjeit kell valamilyen módon aggregálni, hogy ezekből regressziós predikciók készülhessenek a molekulák szintjén.

Helyette a Graph2Vec modellt választottam, amelyet kifejezetten teljes gráfok beágyazására terveztek, így jól illeszkedik a feladathoz. Bár találtam egy jól működő nyílt forráskódú implementációt a KarateClub könyvtárban, sajnos ezt nem tudtam használni, mivel rendkívül összetett és problémás függőségei voltak, amelyek teljesen felborították a conda környezetemet. Ezért végül a meglévő koncepció alapján saját Graph2Vec implementációt készítettem.

Miután elkészültek a gráfok beágyazott (embedding) reprezentációi, ezekből egy három rétegű, teljesen összekötött neurális hálóval (MLP – Multi-Layer Perceptron) becsültem meg a célváltozókat.

Gráf neurális hálók (GNNs)

Az embedding-alapú megközelítést különböző gráf-neurális hálózatokkal szerettem volna összehasonlítani, hogy jobban látható legyen a különbség a klasszikus és a strukturális tanulást végző modellek között. Az összehasonlítás részeként kipróbáltam több, kifejezetten regressziós feladatokra tervezett gráfos modellt is, többek között a GCN, GAT, GIN, MPNN, SchNet és DimeNet (illetve DimeNet++) architektúrákat. A SchNet implementáció sajnos nem működött megfelelően, és nem volt lehetőségem kijavítani az időkereten belül. A DimeNet++ modell pedig rendkívül lassúnak bizonyult, és a validációs teljesítménye sem indokolta, hogy jelentős erőforrást fordítsak a további kísérletezésre. Ennek ellenére a többi

modell – különösen a GCN, GIN és MPNN – elegendő összehasonlítási alapot nyújtott az embedding-alapú módszer értékeléséhez.

A **GCN** (Graph Convolutional Network) a gráfkonvolúciós műveletek klasszikus változata, amely aggregálja a szomszédok információit a csúcsok frissítéséhez. A **GAT** (Graph Attention Network) ezt továbbfejleszti figyelemmechanizmus alkalmazásával, ami súlyozottabbá teszi a különböző szomszédok hozzájárulását.

A **GIN** (Graph Isomorphism Network) egy olyan gráf-neurális hálózat, amelyet kifejezetten a gráfok izomorfizmusának felismerésére fejlesztettek ki. Ez a modell a szomszédos csúcsok információit egy összeadáson alapuló aggregációs művelettel gyűjti össze, majd nemlineáris transzformációkon keresztül frissíti a csúcsok reprezentációját. A GIN erőssége, hogy képes megkülönböztetni olyan gráfokat is, amelyeket sok más GNN típus nem tud, ezáltal közelebb áll a gráfizomorfizmus probléma megoldásához. Emiatt hatékonyan alkalmazható komplex szerkezetek felismerésére, ami molekuláris adatok elemzésénél kiemelten fontos.

Az **MPNN** (Message Passing Neural Network) egy általános és rugalmas gráf-neurális hálózati architektúra, amely iteratív üzenetküldési (message passing) folyamatokra épül. Az MPNN működésének alapja, hogy minden csúcs a szomszédaitól kap üzeneteket, amelyeket összegyűjt, majd saját állapotát frissíti ezek alapján. Ez a folyamat többször ismétlődik, így a csúcsok reprezentációja fokozatosan egyre gazdagabbá válik. Az MPNN jól alkalmazkodik a molekuláris gráfok komplex kölcsönhatásaihoz, képes megragadni a kémiai kötések és atomok finom összefüggéseit, ezért gyakran használják kvantumkémiai regressziós feladatokban.

Megvalósítás és eredmények

Kicsit optimalizáltam néhány modellt, köztük az embedding-alapú Graph2Vec megközelítést is. A tapasztalatok szerint minél nagyobb dimenzióba ágyaztam be a gráfokat, annál jobb teljesítményt ért el a rendszer. Emellett növeltem a tanulási rátát és a downsampling arányt, valamint a Weisfeiler-Lehman (WL) iterációk számát 2-ről 3-ra emeltem. Természetesen a beágyazások dimenziójának növelésével párhuzamosan a regressziót végző MLP rejtett rétegeinek méretét is megnöveltem, hogy a hálózat képes legyen kihasználni a megnövekedett információs kapacitást.

Ezzel a módszerrel így sikerült elérni körülbelül 0,08-as MSE értéket, ami igen kedvező eredménynek számít, különösen annak fényében, hogy a vizsgált regressziós modellek közül csak egy-kettő tudott ennél jobb teljesítményt felmutatni.

Mivel a HOMO-LUMO predikciónál ilyen jó eredményeket sikerült elérni, ezért kiterjesztettem a vizsgálatot az első öt molekuláris tulajdonság becslésére is. Ebben az esetben a rendszer 4,44-es MSE értéket produkált, ami figyelembe véve, hogy egyszerre öt különböző értéket becsültünk, talán még elfogadhatónak tekinthető, bár természetesen közel sem olyan pontos, mint a korábbi, kétváltozós predikció.

Ezután a GCN, GAT, GIN és MPNN modelleket vizsgáltam meg részletesebben. Meglepetésemre ezek közül a GAT teljesített a leggyengébben, mindössze 0,4468-as MSE értéket ért el, még kis optimalizációval is. A GCN jobb eredményt mutatott, 0,2292-es MSE-t ért el optimalizáció nélkül, azonban én inkább a jobbnak ígérkező modellek – GIN és MPNN – finomhangolására koncentráltam.

A GIN modell kezdeti MSE értéke 0,16 volt, amelyet a különböző optimalizációs technikák alkalmazásával sikerült 0,07-re csökkenteni. Ugyanakkor a tanítás során a validációs hibák jelentős ingadozásokat mutattak, és időnként extrém magas, 2000 fölötti kiugrások is előfordultak, ami a modell megbízhatóságát megkérdőjelezte. Emiatt a továbbiakban nem fordítottam rá több figyelmet.

Az elején a legígéretesebb modellnek az MPNN bizonyult, amely már kezdetben is 0,07-es MSE értékkel rendelkezett, és megfelelő optimalizációval ezt 0,04-ig sikerült csökkenteni. Az optimalizált modell az első öt molekuláris tulajdonság becslésére 1,144-es MSE-t ért el, ami jelentősen jobb eredmény a korábbi embedding-alapú megközelítéshez képest. Ezt követően más célváltozókkal is teszteltem a modellt, és ez talán az egyik legizgalmasabb része a kutatásnak, hiszen még rengeteg kísérletet lehetne végezni, ha több idő állna rendelkezésre.

Modell	**Task**	**MSE (kezdeti)**	**MSE (optimalizált)**
Graph2Vec + MLP	HOMO-LUMO	0.85	0.079
	1-5 property	-	4.44
GCN	HOMO-LUMO	0.229	-
GAT	HOMO-LUMO	0.50	0.45
GIN	HOMO-LUMO	0.16	0.07 (instabil)
MPNN	HOMO-LUMO	0.085	0.04
	1-5 property	-	1.144
	12–15 property	-	20.878
	16–18 property	-	1.696

2. Figure: Az elért eredmények

Ha túl sok célváltozót próbálunk egyszerre megtanítani – például a 0–9 vagy 7–10 indexű property-ket próbáltam –, az MPNN modellnek, akkor egyáltalán nem képes hatékonyan megtanulni a becsülni, hiszen a végső teszt veszteség több milliós nagyságrendű lett. Ez arra utal, hogy bizonyos molekuláris tulajdonságok önmagukban is nehezebben becsülhetők meg.

A HOMO-LUMO jó becslési eredményei alapján azt gondoltam, hogy bár egyetlen modell nem feltétlenül képes megbízhatóan előre jelezni az összes 19 célváltozót, érdemes lehet kisebb, összetartozó tulajdonságcsoportokra külön modelleket tanítani. Így összességében megkaphatjuk a teljes predikciót. Ugyanakkor a kísérletek azt is jelzik, hogy vannak olyan paraméterek, amelyek különösen nehezen becsülhetők, ezért ezek egyenkénti vizsgálata is indokolt lenne, bár ez az idő hiányában elmaradt.

A kód elérhető itt: https://github.com/wittmajerd/Graph-DL-HW

Források:

- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). *Neural message passing for quantum chemistry*. In Proceedings of the 34th International Conference on Machine Learning (ICML). https://arxiv.org/abs/1704.01212
- Klicpera, J., Groß, J., & Günnemann, S. (2020). *Directional message passing for molecular graphs*. In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2003.03123
- Kipf, T. N., & Welling, M. (2017). *Semi-supervised classification with graph convolutional networks*. In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1609.02907
- Ramakrishnan, R., Dral, P. O., Rupp, M., & von Lilienfeld, O. A. (2014). *Quantum chemistry structures and properties of 134 kilo molecules*. Scientific Data, 1, 140022. https://doi.org/10.1038/sdata.2014.22
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph attention networks*. In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1710.10903
- Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). *MoleculeNet: A benchmark for molecular machine learning*. Chemical Science, 9(2), 513–530. https://doi.org/10.1039/C7SC02664A
- Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). *How powerful are graph neural networks?* In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1810.00826
- Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). *Graph neural networks: A review of methods and applications*. AI Open, 1, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001
- **PyTorch Geometric QM9 dataset documentation**. (n.d.). *torch_geometric.datasets.QM9*. Retrieved from https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.QM9.html