物質探索のためのグラフ深層学習の事前学習モデル

佐々木 勇和*

Pre-trained Model on Graph Neural Networks for Material Discovery

Yuya SASAKI*

物質探索における人工知能の活躍は近年目覚ましい．一方で，人工知能モデルを訓練するための実験データを大量に準備することは途方もなく時間がかかるため，自身のタスクに特化した高精度な人工知能モデルを構築することは困難である．本研究は，事前学習モデルを構築することで少量の実験データで高精度なモデル構築を目指す．ここでは，データの共有なしで共同的にモデルを構築する技術である連合学習を用いてその可能性を検証する．計算機実験により，データを共有せずに構築したモデルでも高精度な予測できることを示す．

1．背景

2024年，人工知能の研究者であるデミス・ハサビス氏とジョン・ジャンパー氏がノーベル化学賞を受賞した．両名の貢献はアルファフォールドと呼ばれる人工知能モデルの開発であり，このモデルによりタンパク質の立体構造を高精度に予測することで，新たなタンパク質設計の効率化を達成した．この例に限らず，新たな化学物質や材料の開発において人工知能の発展は目覚ましい．人工知能モデルが実験前に所望の物性の予測を行うことで対象物質を絞り込むことができ，実験の優先度が低い物質を特定することができる．新たな物質に探索において人工知能技術の活用は益々加速すると予想され，においてもより広く人工知能技術を活用することが期待されている．

様々な人工知能技術が開発されている中で，化学や材料開発分野ではグラフ深層学習が効果的とされている．グラフデータおよびグラフ深層学習の特徴を下記にまとめる．

- グラフデータ：点と枝で表現されるデータ構造である．分子や結晶構造は点が原子，枝が原子間のつながりという形で自然にグラフとして表現できる．
- グラフ深層学習：グラフデータに対する深層学習である．物性予測などの多様なタスクにおいてグラフ深層学習が用いられており，既存の深層学習においてはグラフ深層学習が最高性能を達成している．

図1は結晶構造をグラフデータで表し，そのデータを深層学習に通すことで物性値や分類ラベルを予測する例を表している．この例のように，分子や結晶はグラフデータと表現できるため，特殊なデータの変形無しでそのまま深層学習の入力とすることができるため，データの事前処理方法などの知見がなくても使用することができる．

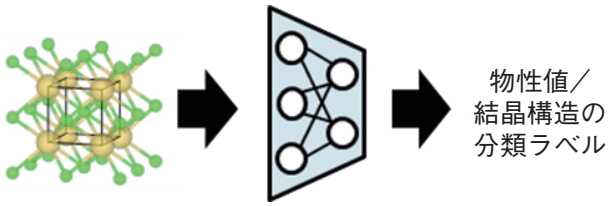

図1 グラフ深層学習から予測の例．

一方で，物質探索においては有機物・無機物など様々な物質を対象にし，さらに毒性の有無や形成エネルギーの予測など多様な予測タスクが用いられている．加えて，一般的には大量のデータが学習用に必要であるため，データの取得に時間が掛かる実験では少数のデータしかできない場合がある．そのため，対象の物質とタスクに対して高精度なモデルを構築できない可能性がある．

本研究の狙いは，グラフ深層学習による物性の予測を少数の実験データで高精度化させることである．そのために，広く公開されており収集が容易な物質やタスクのデータおよびシミュレーションデータを活用することで，対象の物質とタスク以外のデータによる事前学習モデルを構築する．この事前学習モデルに対して，少量の対象の物質とタスクに対する追加学習を行うことで，大量のデータ無しでグラフ深層学習の予測の高精度化を目的とする．

2．手法

事前学習モデルを構築する手順として下記の3ステップを行う．

1．学習用データの収集・整理：化学や材料分野におけるデータはmaterials projectやOQMDなど広くWebに公開されている．それらのデータを統一的なフォーマットに変更し，学習データを準備する．

2026年3月4日 受理
* 豊田理研スカラー
大阪大学大学院情報科学研究科マルチメディア工学専攻

2．グラフ深層学習モデルの検討：グラフ深層学習には様々なモデル構造が提案されており，その数はますます増加している．ここでは，多数のモデル構造からどの構造が事前学習モデルに適しているかを検討する．

3．事前学習モデルの構築：多様な物質とタスクに対して学習を行う事前学習モデルを構築する．本研究では連合学習を活用し，多様なデータに対応可能な事前モデルを作成する．連合学習は複数のクライアントが自身のデータをもつ状況を想定し，自身のデータでモデルを訓練し，それら訓練された複数のモデルをサーバが一つに統合する技術である．これにより，クライアントがどのようなデータをもつかわからない状況で，一つのグローバルなモデルを構築することができる．

3．結果

連合学習を用いた手法の性能を評価し，手法の妥当性を検証する．データとしてGraphs of materials project[1]を用い，band gapとformation energy per atomを予測するタスクとする．GNNモデルとして，GCN[2]，SAGE[3]，GINE[4]，EGNet[5]，M3GNet[6]を用いる．このうち，MEGnetとM3GNetは分子や結晶構造に対する物性予測を対象としたモデルである．連合学習手法として，FedAvg[7]とFedProx[8]の二つを用いる．FedAvgは一般的な連合学習技術であり各クライアントのモデルを平均化してグローバルなモデルを構築する．FedProxは各クライアントが保持するデータに偏りがあることを想定しグローバルなモデルを構築する．ここでは，クライアント数は5台とする．つまり，5台がデータをそれぞれ保持し，データそのものを共有することなくグローバルなモデルを構築する．評価指標として，MSE，RMSE，MAE，およびR2を用いる．前者3つは小さいほど精度がよく，R2は大きいほど精度がよいことを示す．

表1に実験結果を示す．この結果より，FedAvgとFedProx，band gapとformation energy per atomの全てにおいてM3GNetが最高精度を達成している．他のモデルでは，bandgapにおいてはGCNが精度が高く，formation energy per atomではGCNとSAGEの精度が高い．この結果より，M3GNetのような分子や結晶構造を想定したモデルがベースとして適していることがわかる．

表1 実験結果．

LABEL設定			band_gap				formation_energy_per_atom			
精度			MSE	RMSE	MAE	R2	MSE	RMSE	MAE	R2
連合学習 FedAvg	GCNConv	平均値	0.857	0.925	0.635	0.674	0.069	0.262	0.166	0.940
		標準偏差	±0.033	±0.017	±0.013	±0.013	±0.006	±0.011	±0.008	±0.005
	SAGEConv	平均値	1.006	1.002	0.690	0.617	0.071	0.267	0.166	0.938
		標準偏差	±0.032	±0.016	±0.015	±0.011	±0.004	±0.007	±0.004	±0.003
	GINEConv	平均値	1.211	1.099	0.830	0.540	0.189	0.433	0.299	0.836
		標準偏差	±0.095	±0.042	±0.039	±0.038	±0.033	±0.039	±0.025	±0.029
	MEGNet	平均値	1.032	1.016	0.728	0.610	0.135	0.366	0.267	0.883
		標準偏差	±0.043	±0.021	±0.017	±0.020	±0.027	±0.035	±0.042	±0.024
	M3GNet	平均値	0.640	0.799	0.527	0.757	0.028	0.167	0.099	0.975
		標準偏差	±0.019	±0.012	±0.014	±0.007	±0.002	±0.007	±0.005	±0.002
連合学習 FedProx	GCNConv	平均値	0.864	0.929	0.638	0.671	0.072	0.267	0.170	0.938
		標準偏差	±0.037	±0.019	±0.014	±0.013	±0.005	±0.009	±0.007	±0.004
	SAGEConv	平均値	1.011	1.005	0.693	0.616	0.072	0.268	0.166	0.937
		標準偏差	±0.055	±0.027	±0.017	±0.021	±0.008	±0.013	±0.008	±0.007
	GINEConv	平均値	1.183	1.087	0.819	0.550	0.186	0.428	0.295	0.839
		標準偏差	±0.044	±0.020	±0.022	±0.020	±0.034	±0.039	±0.027	±0.029
	MEGNet	平均値	1.046	1.023	0.734	0.602	0.125	0.352	0.245	0.892
		標準偏差	±0.034	±0.017	±0.010	±0.010	±0.015	±0.022	±0.018	±0.013
	M3GNet	平均値	0.633	0.795	0.522	0.760	0.037	0.187	0.108	0.968
		標準偏差	±0.032	±0.020	±0.019	±0.011	±0.010	±0.020	±0.005	±0.009

4．結論

本研究では，物性予測のための事前学習モデルの構築を目指し，データ収集，モデル検討，連合学習によるモデル構築を実施した．実験ではGraphs of materials projectを用いて実験を行い，精度の検証を行い，M3GNetの精度が高くベースモデルとして適していることがわかった．今後は，より多様なデータを用いた学習およびメタラーニングなどの学習技術を試す．

REFERENCES

1) Graphs of materials project: https://figshare.com/articles/dataset/Graphs_of_materials_project/7451351?file=15087992.
2) T. N. Kipf and M. Welling, *ICLR* (2017).
3) W. L. Hamilton, *et al*., *NeurIPS* (2017).
4) K. Xu, *et al*., *ICLR* (2019).
5) C. Chen, *et al*., *Chemistry of Materials*, **31.9** (2019) 3564-3572.
6) C. Chen and S. P. Ong, *Nature Computational Science,* **2.11** (2022) 718-728.
7) B. McMahan, *et al*., *Artificial intelligence and statistics,* (2017) 1273-1282.
8) T. Li, *et al*., *Proceedings of Machine Learning and Systems*, **2** (2020) 429-450.