5ZM-06

# 深層学習によるソースコードの分類および判断根拠の可視化

三森 正椰 †
木更津工業高等専門学校 †

大枝 真一 ‡
木更津工業高等専門学校 ‡

## 1 はじめに

近年，初等・中等教育においてプログラミング教育が必修となるなど，その重要性が益々高まっている．プログラミング教育においては，学生の習熟度に応じた指導が不可欠であり，これを実現する上で，個々のコーディング能力を客観的に評価することが求められる．先行研究では，ソースコードの構造情報をグラフに変換し，深層学習を利用して能力の判定を行う手法が提案されており，高い精度で判定を行えることが示されている [1]．しかし，どの構造情報が判定に寄与しているのかは依然として不明確である．

そこで本研究では，ソースコードのグラフ表現を用いて，初級者と上級者のプログラムを分類するモデルを構築する．また，構築したモデルの判断根拠を可視化し，初級者・上級者のプログラムの構造的特徴を明らかにする．

## 2 提案手法

はじめに，ソースコードを抽象構文木 (AST) と呼ばれる木構造に変換し，この構造情報を Graph Convolutional Networks(GCN)[2] を用いて学習することで，ソースコードが初級者・上級者のどちらによって書かれたものであるかを予測するモデルを構築する (図 1)．

さらに，構築したモデルに対して Integrated Gradients(IG)[3] を適用し，各エッジ・ノードの寄与度を算出する．寄与度が正であれば赤系統，負であれば青系統で色付けを行い，モデルの判断根拠を可視化する．

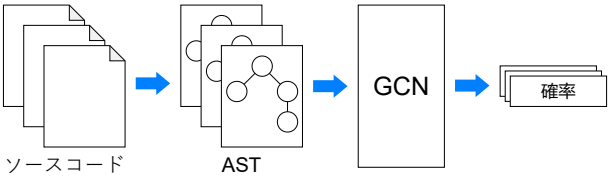



図 1: ソースコード分類モデル

## 3 実験

### 3.1 初級者・上級者の予測

本実験では，Codeforces[*1]の提出物で構成されるデータセットを使用する．このデータセットは，オンライン上で公開されている[*2]．

はじめに，C 言語のソースコードを提出者のレーティングに基づいて並び替え，下位 20% を初級者，上位 20% を上級者として分類する．分析対象となるソースコードは合計で 10,594 件で，このうち初級者に分類されるものは 4,708 件，上級者に分類されるものは 5,886 件である．

続いて，これらのソースコードを AST に変換し，GCN に入力することで，ソースコードの著者が初級者または上級者である確率を予測するモデルを構築する．評価指標としては，2 クラス分類問題における精度評価の指標である AUC(Area Under Curve) を用いる．

テストデータに対する予測結果の ROC 曲線を図 2 に示す．AUC は 0.902 となっており，過半数のソースコードを正確に分類できることを示している．したがって，AST から得られる情報は，ソースコードの著者が初級者または上級者であるかを予測するのに有効であると結論づけられる．

Deep Learning-based Source Code Classification and Visualization of Decision Rationales

† Seiya Mitsumori, National Institute of Technology, Kisarazu College
‡ Shinichi Oeda, National Institute of Technology, Kisarazu College

[*1] https://codeforces.com/
[*2] https://sites.google.com/site/miningprogcodeforces/

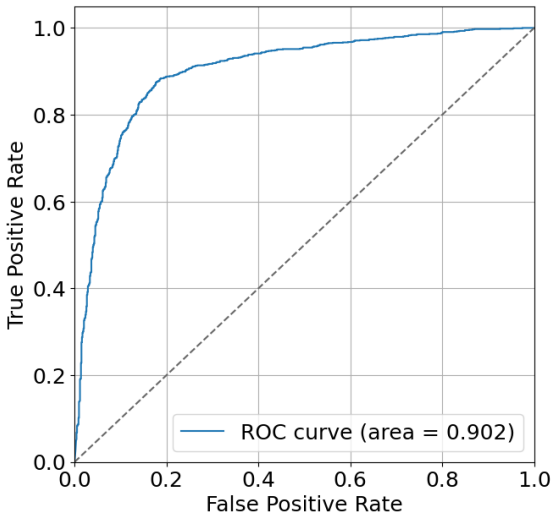



図 2: テストデータに対する予測結果

### 3.2 判断根拠の分析・可視化

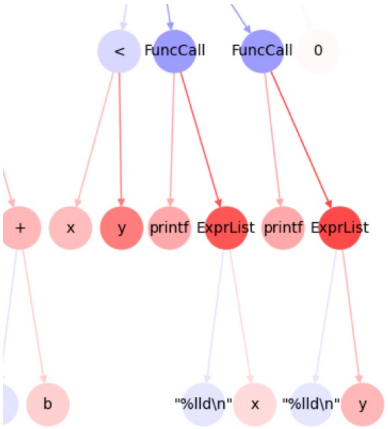
前節で構築した予測モデルに対して IG を適用して各ノード・エッジの寄与度を算出し，分析・可視化を行う．

まず，初級者の AST を対象に，正の寄与度が 0.7 以上のノードを集計した．その結果，関数に渡す引数を表す ExprList ノードが最頻出であった．上級者の AST に対しても同様の方法でノードの集計を行った結果，関数呼び出しを表す FuncCall ノードが最頻出であった．このことから，初級者は関数に渡す引数が不適切である可能性が高いこと，上級者は関数呼び出しを頻繁に行うことなどが考えられる．

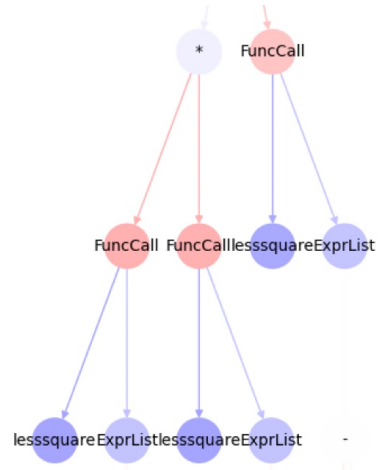
初級者の AST と上級者の AST について，判断根拠の可視化を行った結果をそれぞれ図 3a，図 3b に示す．ここでは，正の寄与度は赤，負の寄与度は青で表しており，濃い色であるほど寄与度が高いことを示している．

## 4 まとめ

本研究では，ソースコードを AST に変換し，GCN を用いて初級者・上級者の分類を行った．その結果，過半数のソースコードを正確に分類することができた．また，IG を適用して予測モデルの判断根拠を可視化し，予測値の増減に寄与する構造情報を明らかにした．今後は，初級者・上級者が書くソースコードの特徴をより細かく分析したい．




(a) 初級者の AST における寄与度の可視化




(b) 上級者の AST における寄与度の可視化

図 3: 判断根拠の可視化（一部抜粋）

## 謝辞

本研究は JSPS 科研費 JP19H01728, JP23K17604 の助成を受けたものです．

## 参考文献

[1] 松井智寛ほか. “ソースコードのグラフ表現を利用した深層学習によるコーディングの専門性の判定手法”. 情報処理学会研究報告, Vol.2022-SE-210, No.12, pp.1-8, 2022.

[2] TN Kipf, et al. “Semi-supervised classification with graph convolutional networks”. In *ICLR*, 2017.

[3] M Sundararajan, et al. “Axiomatic attribution for deep networks”. In *ICML*, Vol.70 of PMLR, pp.3319–3328, 2017.