



DeepLineDP: Towards a Deep Learning Approach for Line-Level Defect Prediction

Patrycja Kałużna (252864)

Jakub Walaszek (252897)

Michał Gałęcki (252869)

Spis treści:



- wybór tematu projektu
- infrastruktura badawcza online
- reprodukcja badań
- infrastruktura badawcza offline
- rozwój badań
- artykuł

Wybór tematu projektu

DeepLineDP: Towards a Deep Learning Approach for Line-Level Defect Prediction

Chanathip Pornprasit^{ID}, *Student Member, IEEE* and Chakkrit (Kla) Tantithamthavorn^{ID}, *Member, IEEE*

Abstract—Defect prediction is proposed to assist practitioners effectively prioritize limited Software Quality Assurance (SQA) resources on the most risky files that are likely to have post-release software defects. However, there exist two main limitations in prior studies: (1) the granularity levels of defect predictions are still coarse-grained and (2) the surrounding tokens and surrounding lines have not yet been fully utilized. In this paper, we perform a survey study to better understand how practitioners perform code inspection in modern code review process, and their perception on a line-level defect prediction. According to the responses from 36 practitioners, we found that 50% of them spent at least 10 minutes to more than one hour to review a single file, while 64% of them still perceived that code inspection activity is challenging to extremely challenging. In addition, 64% of the respondents perceived that a line-level defect prediction tool would potentially be helpful in identifying defective lines. Motivated by the practitioners' perspective, we present DeepLineDP, a deep learning approach to automatically learn the semantic properties of the surrounding tokens and lines in order to identify defective files and defective lines. Through a case study of 32 releases of 9 software projects, we find that the risk score of code tokens varies greatly depending on their location. Our DeepLineDP is 17%-37% more accurate than other file-level defect prediction approaches; is 47%-250% more cost-effective than other line-level defect prediction approaches; and achieves a reasonable performance when transferred to other software projects. These findings confirm that the surrounding tokens and surrounding lines should be considered to identify the fine-grained locations of defective files (i.e., defective lines).

Infrastruktura badawcza online

- korzysta ze środowiska Google Colaboratory
- pozwala na reprodukcję większości badań z wyjątkiem tzw. line-level baselines
- konfiguruje środowisko conda'y oraz dla języka R

<https://colab.research.google.com/drive/139uWve5H07uM0SIKZSuevsi-dEjWeK9P?usp=sharing>

Reprodukcja badań

- wytrenowano wszystkie modele wykorzystywane w ramach opracowywanego artykułu naukowego oraz dokonano predykcji przy ich pomocy, tj.:
 - model DeepLineDP
 - modele tzw. file-level baselines
 - modele tzw. line-level baselines
- na podstawie łącznie 32 release'ów pochodzących z 9 otwartoźródłowych projektów

Infrastruktura badawcza offline

- ze względu na fakt, że środowisko Google Colaboratory nie pozwala na reprodukcję tzw. line-level baselines zostanie stworzona infrastruktura badawcza offline w postaci skryptu konfigurującego środowisko i uruchamiającego reprodukcję badań

Rozwój badań

- pomysłem na rozwój badań jest poszerzenie zbioru danych wykorzystywanego do wytrenowania modelu DeepLineDP
- zbiór danych "Defectors" pochodzący z artykułu naukowego "Defectors: A Large, Diverse Python Dataset for Defect Prediction" jest obiecujący

Rozwój badań

- zbiór danych "Defectors" jest:
 - około 2 razy większy niż dotychczas największy zbiór danych wykorzystywany przy predykcji defektów,
 - lepiej zbalansowany niż inne zbiory danych (stosunek instancji zawierających defekty do instancji niezawierających ich wynosi około 1:1)
 - zawiera dane z 24 projektów pochodzących z 24 organizacji oraz 18 domen takich jak ML, automation i IoT
 - oparty o pythonowe projekty kiedy prawie wszystkie zbiory danych są oparte o projekty napisane w Javie
 - nowy - artykuł naukowy, z którego pochodzi, został opublikowany 8 marca 2023 roku

Artykuł



- rozpoczęto pisanie sekcji "Introduction"

<https://www.overleaf.com/project/6401cc2de33881644150cd5f>



Dziękujemy za uwagę!

Bibliografia

- Pornprasit C., Tantithamthavorn C., *DeepLineDP: Towards a Deep Learning Approach for Line-Level Defect Prediction*. IEEE Transactions on Software Engineering, tom 49, 2023.
- Mahbub P., Shuvo O., Rahman M. M., *Defectors: A Large, Diverse Python Dataset for Defect Prediction*. arXiv preprint arXiv:2303.04738, 2023.