NEURAL NETWORKS FOR MODELING SOURCE CODE EDITS

Rui Zhao

ABSTRACT

Programming languages are emerging as a challenging and interesting domain for machine learning. A core task, which has received significant attention in recent years, is building generative models of source code. However, to our knowledge, previous generative models have always been framed in terms of generating static snapshots of code. In this work, we instead treat source code as a dynamic object and tackle the problem of modeling the edits that software developers make to source code files. This requires extracting intent from previous edits and leveraging it to generate subsequent edits. We develop several neural networks and use synthetic data to test their ability to learn challenging edit patterns that require strong generalization. We then collect and train our models on a large-scale dataset of Google source code, consisting of millions of fine-grained edits from thousands of Python developers. From the modeling perspective, our main conclusion is that a new composition of attentional and pointer network components provides the best overall performance and scalability. From the application perspective, our results provide preliminary evidence of the feasibility of developing tools that learn to predict future edits.

编程语言正在成为机器学习的一个具有挑战性和趣味性的领域。构建源代码生成模型是近年来备受关注的一项核心任务。然而，据我们所知，以前的生成模型总是根据生成代码的静态快照来构建的。在这项工作中，我们将源代码视为一个动态对象，并处理软件开发人员对源代码文件进行编辑的建模问题。这需要从以前的编辑中提取意图，并利用它生成后续编辑。我们开发了几个神经网络，并使用合成数据来测试它们学习具有挑战性的编辑模式的能力，这些模式需要很强的泛化能力。然后，我们在一个包含大量谷歌源代码的数据集上收集并训练我们的模型，这个数据集由来自数千名Python开发人员的数百万个细粒度编辑组成。从建模的角度来看，我们的主要结论是注意力和指针网络组件的新组合提供了最佳的总体性能和可伸缩性。从应用的角度来看，我们的结果为开发能够预测未来编辑的工具的可行性提供了初步的证据。