

퍼징의 변형 (Mutation) 기술 인공지능에 활용하기

이동재

20240613

Abstract

확산 (Diffusion) 모델은 자가 회귀 모델과 함께 생성형 인공지능의 핵심 기술 중 하나이다. 이미지를 생성하기 위해 처음 개발되었지만, 자연어, 코드를 생성하는데도 활용되고 있다. 본 글에서는 Kapur가 제시한 확산 모델 기반 프로그램을 편집 기술을 소개하고, 그 활용 방안에 대해 다룬다.

퍼징의 변형은 무작위성 입력을 만들어내는데 필수적이다. 보편적인 프로그램을 대상으로 한 퍼징에서는 비트 단위의 변형을 통해 새로운 입력을 만들어낸다. 그러나 컴파일러와 같은 프로그램을 퍼징하기 위해서는 입력 문법에 맞는 입력을 만들어내는 기술이 필요하다. 이러한 상황에서는 구문 트리를 변형하는 방식을 통해 무작위성 입력을 생성할 수 있다.

무작위성 입력 변형은 확산 모델의 잡음 과정 (Noising process)과 유사하다. 확산 모델은 전파 과정에서 입력에 잡음을 조금씩 추가해 가우시안 분포를 따르도록 만든다. 그리고, 모델은 잡음을 제거하는 역과정 (Backward process)을 학습한다. 무작위성 입력 변형 과정을 입력에 노이즈를 더하는 과정으로 간주한다면, 모델은 입력 변형의 역과정을 학습할 수 있다.

무작위성 변형의 역과정을 학습할 수 있다면 프로그램 합성, 퍼징 등 다양한 작업에 활용할 수 있을 것이다. 프로그램 합성은 탐색을 통해 프로그램을 생성하기 때문에 큰 프로그램을 만드는데 시간이 오래걸린다. 확산 모델을 활용하면 적당히 완성된 프로그램을 확산 모델로 편집하여 빠르게 프로그램을 합성하는 모델을 구축할 수 있을 것이다.

퍼징에서는 컴파일러 등 특정 형식의 입력을 만족하는 프로그램을 생성하는데 활용할 수 있을 것이다. 결함에 도달하는 입력에 무작위성 입력 변형을 적용하고, 그 역과정을 모델이 학습할 수 있다면 효율적인 변형 도구를 만들 수 있을 것이다.

자가 회귀 모델과 달리 확산 모델은 주어진 입력을 직접적으로 수정하는 것이 가능하다. 이러한 장점을 살려 프로그램을 편집하고, 재구성하는데 활용할 수 있을 것으로 기대된다.