(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년10월16일
(11) 등록번호 10-2717822
(24) 등록일자 2024년10월10일

(51) 국제특허분류(Int. Cl.)
G06F 8/75 (2018.01) *G06F 8/41* (2018.01)
G06N 20/20 (2019.01)
(52) CPC특허분류
G06F 8/75 (2013.01)
G06F 8/42 (2013.01)
(21) 출원번호 10-2022-0138456
(22) 출원일자 2022년10월25일
심사청구일자 2022년10월25일
(65) 공개번호 10-2024-0059681
(43) 공개일자 2024년05월08일
(56) 선행기술조사문헌
JP2016126727 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
한요섭
서울특별시 서대문구 연세로 50, 컴퓨터과학과 (신촌동, 연세대학교)
서현태
경기도 안양시 만안구 박달로 403, 101동 1502호 (박달동, 한일유앤아이아파트)
(뒷면에 계속)
(74) 대리인
특허법인우인

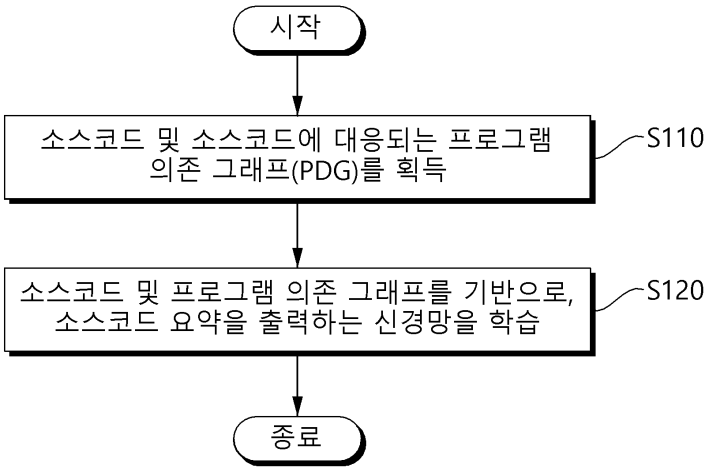
전체 청구항 수 : 총 8 항 심사관 : 채정목

(54) 발명의 명칭 **구조 정보를 이용한 인공지능 기반 소스코드 요약 방법, 이를 수행하는 장치 및 컴퓨터 프로그램**

(57) 요 약

본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법, 이를 수행하는 장치 및 컴퓨터 프로그램은, 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 소스코드를 요약함으로써, 프로그램 의존 그래프(program dependency graph, PDG)를 학습하는 인코더와 코드 시퀀스를 학습하는 인코더를 앙상블하여 구조적 특성과 코드의 의미를 함께 학습하는 신경망 모델을 제안하여 코드 요약 성능을 향상시킬 수 있다.

대 표 도 - 도2



(52) CPC특허분류
G06F 8/44 (2013.01)
G06N 20/20 (2021.08)

(72) 발명자
손지경
서울특별시 관악구 승방6길 6, 401호(남현동)

한중혁
서울특별시 마포구 토정로18길 11, 102동 1604호(현석동, 래미안웰스트림)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711159332
과제번호 2020R1A4A3079947
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 기초연구실지원사업
연구과제명 휴먼-AI 협업 프로그래밍 플랫폼 기술 연구실(3/3)
기 여 율 1/2
과제수행기관명 연세대학교 산학협력단
연구기간 2022.03.01 ~ 2023.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711152718
과제번호 2020-0-01361-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송연구개발사업
연구과제명 인공지능대학원지원사업[1단계] (3/5)
기 여 율 1/2
과제수행기관명 연세대학교 산학협력단
연구기간 2022.01.01 ~ 2022.12.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

소스코드 및 상기 소스코드에 대응되는 프로그램 의존 그래프(program dependency graph, PDG)를 획득하는 단계; 및

상기 소스코드 및 상기 프로그램 의존 그래프(PDG)를 기반으로, 소스코드 요약을 출력으로 하는 신경망을 학습하는 단계;를 포함하고,

상기 신경망은, 상기 소스코드를 기반으로 코드 시퀀스를 학습하는 제1 인코더; 상기 프로그램 의존 그래프(PDG)를 기반으로 구조적 특징을 학습하는 제2 인코더; 및 상기 제1 인코더의 출력과 상기 제2 인코더의 출력 앙상블을 입력받아 상기 소스코드에 대한 상기 소스코드 요약을 생성하는 디코더;를 포함하고,

상기 제1 인코더는, 상기 소스코드의 토큰 임베딩을 입력받고, 상기 소스코드에 대한 토큰 수준의 잠재 표현을 출력하는, 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에서,

상기 제2 인코더는,

상기 프로그램 의존 그래프(PDG)를 입력받고, 상기 소스코드에 대한 구문 수준의 잠재 표현을 출력하는,

구조 정보를 이용한 인공지능 기반 소스코드 요약 방법.

청구항 5

제4항에서,

상기 제2 인코더는,

상기 프로그램 의존 그래프(PDG)의 그래프 간선의 어텐션(attention)을 사용하는,

구조 정보를 이용한 인공지능 기반 소스코드 요약 방법.

청구항 6

제5항에서,

상기 디코더는,

상기 제1 인코더의 출력인 토큰 수준의 잠재 표현과 상기 제2 인코더의 출력인 구문 수준의 잠재 표현이 연결된 값을 입력받고, 자연어 요약문으로 이루어진 상기 소스코드 요약을 출력하는,

구조 정보를 이용한 인공지능 기반 소스코드 요약 방법.

청구항 7

제1항 및 제4항 내지 제6항 중 어느 한 항에 기재된 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법을 컴퓨터에서 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램.

청구항 8

소스코드의 구조 정보를 이용하여 인공지능을 기반으로 상기 소스코드를 요약하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및

상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 상기 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 상기 소스코드를 요약하기 위한 동작을 수행하는 하나 이상의 프로세서;

를 포함하며,

상기 프로세서는,

상기 소스코드 및 상기 소스코드에 대응되는 프로그램 의존 그래프(program dependency graph, PDG)를 획득하고,

상기 소스코드 및 상기 프로그램 의존 그래프(PDG)를 기반으로, 소스코드 요약을 출력으로 하는 신경망을 학습하고,

상기 신경망은, 상기 소스코드를 기반으로 코드 시퀀스를 학습하는 제1 인코더; 상기 프로그램 의존 그래프(PDG)를 기반으로 구조적 특징을 학습하는 제2 인코더; 및 상기 제1 인코더의 출력과 상기 제2 인코더의 출력 앙상블을 입력받아 상기 소스코드에 대한 상기 소스코드 요약을 생성하는 디코더;를 포함하고,

상기 제1 인코더는, 상기 소스코드의 토큰 임베딩을 입력받고, 상기 소스코드에 대한 토큰 수준의 잠재 표현을 출력하는, 구조 정보를 이용한 인공지능 기반 소스코드 요약 장치.

청구항 9

삭제

청구항 10

삭제

청구항 11

제8항에서,

상기 제2 인코더는,

상기 프로그램 의존 그래프(PDG)를 입력받고, 상기 소스코드에 대한 구문 수준의 잠재 표현을 출력하는,

구조 정보를 이용한 인공지능 기반 소스코드 요약 장치.

청구항 12

제11항에서,

상기 디코더는,

상기 제1 인코더의 출력인 토큰 수준의 잠재 표현과 상기 제2 인코더의 출력인 구문 수준의 잠재 표현이 연결된 값을 입력받고, 자연어 요약문으로 이루어진 상기 소스코드 요약을 출력하는,

구조 정보를 이용한 인공지능 기반 소스코드 요약 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법, 이를 수행하는 장치 및 컴퓨터 프로그램에 관한 것으로서, 더욱 상세하게는 소스코드를 요약하는, 방법, 장치 및 컴퓨터 프로그램에 관한 것이다.

배 경 기 술

[0002] 소스코드에 대한 요약문을 생성하는 딥러닝 연구는 최근에 시퀀스 형태의 정보뿐만 아니라 코드의 구조적 특성도 반영하여 학습하려는 경향이 있다. 이를 위해, 추상 구문 트리(abstract syntax tree, AST) 혹은 제어 흐름 그래프(control flow graph, CFG) 등의 구조로 코드를 변환해 이를 학습에 사용한다.

[0003] 특히, 추상 구문 트리(AST)를 많이 사용하는데, mAST+GCN, CAST, SiT 등의 딥러닝 모델이 기존 시퀀스만을 처리하는 연구보다 성능 향상을 이끌어 냈다. 하지만, 추상 구문 트리(AST)는 긴 코드에 대해 지나치게 복잡한 그래프 구조를 형성하기 때문에 이러한 경우, 요약문 생성 성능이 많이 떨어지는 문제가 있다. 또한, 그래프 구조를 이용하지만 기존 기술은 이에 적합한 학습 모델 구조에 대한 연구는 상대적으로 부족한 상황이다.

발명의 내용

해결하려는 과제

[0004] 본 발명이 이루고자 하는 목적은, 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 소스코드를 요약하는, 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법, 이를 수행하는 장치 및 컴퓨터 프로그램을 제공하는 데 있다.

[0005] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제의 해결 수단

[0006] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법은, 소스코드 및 상기 소스코드에 대응되는 프로그램 의존 그래프(program dependency graph, PDG)를 획득하는 단계; 및 상기 소스코드 및 상기 프로그램 의존 그래프(PDG)를 기반으로, 소스코드 요약을 출력으로 하는 신경망을 학습하는 단계;를 포함한다.

[0007] 여기서, 상기 신경망은, 상기 소스코드를 기반으로 코드 시퀀스를 학습하는 제1 인코더; 상기 프로그램 의존 그래프(PDG)를 기반으로 구조적 특징을 학습하는 제2 인코더; 및 상기 제1 인코더의 출력과 상기 제2 인코더의 출력 앙상블을 입력받아 상기 소스코드에 대한 상기 소스코드 요약을 생성하는 디코더;를 포함할 수 있다.

[0008] 여기서, 상기 제1 인코더는, 상기 소스코드의 토큰 임베딩을 입력받고, 상기 소스코드에 대한 토큰 수준의 잠재 표현을 출력할 수 있다.

[0009] 여기서, 상기 제2 인코더는, 상기 프로그램 의존 그래프(PDG)를 입력받고, 상기 소스코드에 대한 구문 수준의 잠재 표현을 출력할 수 있다.

[0010] 여기서, 상기 제2 인코더는, 상기 프로그램 의존 그래프(PDG)의 그래프 간선의 어텐션(attention)을 사용할 수 있다.

[0011] 여기서, 상기 디코더는, 상기 제1 인코더의 출력인 토큰 수준의 잠재 표현과 상기 제2 인코더의 출력인 구문 수준의 잠재 표현이 연결된 값을 입력받고, 자연어 요약문으로 이루어진 상기 소스코드 요약을 출력할 수 있다.

[0013] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 컴퓨터 프로그램은 컴퓨터 판독 가능한 저장 매체에 저장되어 상기한 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법 중 어느 하나를 컴퓨터에서 실행시킨다.

[0015] 상기의 기술적 과제를 달성하기 위한 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 장치는, 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 상기 소스코드를 요약하기 위한 하나 이상의 프로그램을 저장하는 메모리; 및 상기 메모리에 저장된 상기 하나 이상의 프로그램에 따라 상기 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 상기 소스코드를 요약하기 위한 동작을 수행하는 하나 이상의 프로세서;를 포함하며, 상기 프로세서는, 상기 소스코드 및 상기 소스코드에 대응되는 프로그램 의존 그래프(program dependency graph, PDG)를 획득하고, 상기 소스코드 및 상기 프로그램 의존 그래프(PDG)를 기반으로, 소스코드 요약을 출력으로 하는 신경망을 학습한다.

[0016] 여기서, 상기 신경망은, 상기 소스코드를 기반으로 코드 시퀀스를 학습하는 제1 인코더; 상기 프로그램 의존 그

래프(PDG)를 기반으로 구조적 특징을 학습하는 제2 인코더; 및 상기 제1 인코더의 출력과 상기 제2 인코더의 출력 앙상블을 입력받아 상기 소스코드에 대한 상기 소스코드 요약을 생성하는 디코더;를 포함할 수 있다.

[0017] 여기서, 상기 제1 인코더는, 상기 소스코드의 토큰 임베딩을 입력받고, 상기 소스코드에 대한 토큰 수준의 잠재 표현을 출력할 수 있다.

[0018] 여기서, 상기 제2 인코더는, 상기 프로그램 의존 그래프(PDG)를 입력받고, 상기 소스코드에 대한 구문 수준의 잠재 표현을 출력할 수 있다.

[0019] 여기서, 상기 디코더는, 상기 제1 인코더의 출력인 토큰 수준의 잠재 표현과 상기 제2 인코더의 출력인 구문 수준의 잠재 표현이 연결된 값을 입력받고, 자연어 요약문으로 이루어진 상기 소스코드 요약을 출력할 수 있다.

발명의 효과

[0020] 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법, 이를 수행하는 장치 및 컴퓨터 프로그램에 의하면, 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 소스코드를 요약함으로써, 프로그램 의존 그래프(program dependency graph, PDG)를 학습하는 인코더와 코드 시퀀스를 학습하는 인코더를 앙상블하여 구조적 특성과 코드의 의미를 함께 학습하는 신경망 모델을 제안하여 코드 요약 성능을 향상시킬 수 있다.

[0021] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0022] 도 1은 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 장치를 설명하기 위한 블록도이다.

도 2는 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법을 설명하기 위한 흐름도이다.

도 3은 본 발명의 바람직한 실시예에 따른 신경망의 구조를 설명하기 위한 흐름도이다.

도 4는 본 발명의 바람직한 실시예에 따른 신경망의 학습 과정의 일례를 설명하기 위한 도면이다.

도 5는 본 발명의 바람직한 실시예에 따른 프로그램 의존 그래프의 일례를 설명하기 위한 도면이다.

도 6은 본 발명의 실시예에 따른 소스코드의 토큰 유형의 일례를 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0023] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 게시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 게시가 완전하도록 하고, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다.

[0024] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0025] 본 명세서에서 "제1", "제2" 등의 용어는 하나의 구성 요소를 다른 구성 요소로부터 구별하기 위한 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다. 예컨대, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.

[0026] 본 명세서에서 각 단계들에 있어 식별부호(예컨대, a, b, c 등)는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 단계들은 명기된 순서와 동일하게 일어날 수도 있고 실질적

으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.

[0027] 본 명세서에서, "가진다", "가질 수 있다", "포함한다" 또는 "포함할 수 있다" 등의 표현은 해당 특징(예컨대, 수치, 기능, 동작, 또는 부품 등의 구성 요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.

[0030] 이하에서 첨부한 도면을 참조하여 본 발명에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법, 이를 수행하는 장치 및 컴퓨터 프로그램의 바람직한 실시예에 대해 상세하게 설명한다.

[0032] 먼저, 도 1을 참조하여 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 장치에 대하여 설명한다.

[0033] 도 1은 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 장치를 설명하기 위한 블록도이다.

[0034] 도 1을 참조하면, 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 장치(이하 '소스코드 요약 장치'라 한다)(100)는 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 소스코드를 요약할 수 있다.

[0035] 즉, 소스코드 요약 장치(100)는 프로그램 의존 그래프(program dependency graph, PDG)를 학습하는 인코더와 코드 시퀀스를 학습하는 인코더를 앙상블하여 구조적 특성과 코드의 의미를 함께 학습하는 신경망 모델을 제안하여 코드 요약 성능을 향상시킬 수 있다.

[0037] 이를 위해, 소스코드 요약 장치(100)는 하나 이상의 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130) 및 통신 버스(150)를 포함할 수 있다.

[0038] 프로세서(110)는 소스코드 요약 장치(100)가 동작하도록 제어할 수 있다. 예컨대, 프로세서(110)는 컴퓨터 판독 가능한 저장 매체(130)에 저장된 하나 이상의 프로그램(131)을 실행할 수 있다. 하나 이상의 프로그램(131)은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 컴퓨터 실행 가능 명령어는 프로세서(110)에 의해 실행되는 경우 소스코드 요약 장치(100)로 하여금 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 소스코드를 요약하기 위한 동작을 수행하도록 구성될 수 있다.

[0039] 컴퓨터 판독 가능한 저장 매체(130)는 소스코드의 구조 정보를 이용하여 인공지능을 기반으로 소스코드를 요약하기 위한 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장 매체(130)에 저장된 프로그램(131)은 프로세서(110)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능한 저장 매체(130)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 소스코드 요약 장치(100)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0040] 통신 버스(150)는 프로세서(110), 컴퓨터 판독 가능한 저장 매체(130)를 포함하여 소스코드 요약 장치(100)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0041] 소스코드 요약 장치(100)는 또한 하나 이상의 입출력 장치를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(170) 및 하나 이상의 통신 인터페이스(190)를 포함할 수 있다. 입출력 인터페이스(170) 및 통신 인터페이스(190)는 통신 버스(150)에 연결된다. 입출력 장치(도시하지 않음)는 입출력 인터페이스(170)를 통해 소스코드 요약 장치(100)의 다른 컴포넌트들에 연결될 수 있다.

[0044] 그러면, 도 2 및 도 3을 참조하여 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법에 대하여 설명한다.

[0045] 도 2는 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법을 설명하기 위한 흐름도이고, 도 3은 본 발명의 바람직한 실시예에 따른 신경망의 구조를 설명하기 위한 흐름도이다.

[0046] 도 2를 참조하면, 소스코드 요약 장치(100)의 프로세서(110)는 소스코드 및 소스코드에 대응되는 프로그램 의존 그래프(PDG)를 획득할 수 있다(S110).

[0048] 이후, 프로세서(110)는 소스코드 및 소스코드 의존 그래프(PDG)를 기반으로, 소스코드 요약을 출력하는 신경망을 학습할 수 있다(S120).

[0049] 여기서, 신경망은 도 3에 도시된 바와 같이, 제1 인코더, 제2 인코더, 및 디코더를 포함할 수 있다.

[0050] 제1 인코더는 소스코드를 기반으로 코드 시퀀스를 학습할 수 있다.

[0051] 즉, 제1 인코더는 소스코드의 토큰 임베딩을 입력받고, 소스코드에 대한 토큰 수준의 잠재 표현을 출력할 수 있다.

[0052] 제2 인코더는 제1 인코더에 입력되는 소스코드에 대응되는 프로그램 의존 그래프(PDG)를 기반으로 구조적 특징을 학습할 수 있다.

[0053] 즉, 제2 인코더는 프로그램 의존 그래프(PDG)를 입력받고, 소스코드에 대한 구문 수준의 잠재 표현을 출력할 수 있다. 이때, 제2 인코더는 프로그램 의존 그래프(PDG)의 그래프 간선의 어텐션(attention)을 사용할 수 있다.

[0054] 디코더는 제1 인코더의 출력과 제2 인코더의 출력 앙상블을 입력받아 소스코드에 대한 소스코드 요약을 생성할 수 있다.

[0055] 즉, 디코더는 제1 인코더의 출력인 토큰 수준의 잠재 표현과 제2 인코더의 출력인 구문 수준의 잠재 표현이 연결된 값인 연결된 잠재 표현을 입력받고, 자연어 요약문으로 이루어진 소스코드 요약을 출력할 수 있다.

[0056] 다시 설명하면, 프로세서(110)는 복수개의 훈련 데이터로 이루어진 학습 데이터를 기반으로, 소스코드와 이에 대응되는 프로그램 의존 그래프(PDG)가 입력되면 해당 소스코드에 대한 자연어 요약문인 소스코드 요약을 출력하는 신경망을 학습할 수 있다. 여기서, 훈련 데이터는 소스코드, 해당 소스코드에 대응되는 프로그램 의존 그래프(PDG) 및 해당 소스코드에 대응되는 정답 요약을 포함할 수 있다. 즉, 훈련 데이터의 소스코드와 프로그램 의존 그래프(PDG)는 입력 값으로 이용하고, 훈련 데이터의 정답 요약은 출력 값으로 이용하여, 훈련 데이터를 토대로 신경망을 반복적으로 학습할 수 있다. 신경망의 학습이 완료된 이후, 프로세서(110)는 대상 소스코드와 이에 대응되는 프로그램 의존 그래프(PDG)가 입렵되면, 학습되어 구축된 신경망을 이용하여 대상 소스코드에 대한 소스코드 요약을 획득할 수 있다.

[0059] 그러면, 도 4 내지 도 6을 참조하여 본 발명의 바람직한 실시예에 따른 구조 정보를 이용한 인공지능 기반 소스코드 요약 방법에 대하여 보다 자세하게 설명한다.

[0060] 도 4는 본 발명의 바람직한 실시예에 따른 신경망의 학습 과정의 일례를 설명하기 위한 도면이고, 도 5는 본 발명의 바람직한 실시예에 따른 프로그램 의존 그래프의 일례를 설명하기 위한 도면이며, 도 6은 본 발명의 실시예에 따른 소스코드의 토큰 유형의 일례를 설명하기 위한 도면이다.

[0061] 본 발명에 따른 소스코드 요약 장치(100)는 추상 구문 트리(abstract syntax tree, AST)의 코드 길이에 따른 복잡성을 극복하기 위해 제어 흐름과 데이터 흐름을 고려한 프로그램 의존 그래프(PDG)를 사용하고 그래프 구조를 학습하는 노드 인코더(즉, 제2 인코더)를 포함하는 신경망에 대한 것이다.

[0062] 즉, 본 발명은 JAVA와 C 언어를 대상으로 프로그램의 요약문을 제공하는 딥러닝 기법을 제안한다. 본 발명에서는 프로그램을 표현하는 소스코드의 구조적 특성을 주목해 프로그램 의존 그래프(PDG)를 사용하고 이를 학습할 수 있는 노드 인코더(즉, 제2 인코더)를 제안한다. 본 발명은 모델의 효과적인 학습을 위해 소스코드 시퀀스를 학습하는 인코더(즉, 제1 인코더)와 노드 인코더(즉, 제2 인코더)를 앙상블하여 자연어 요약문을 제공하는 딥러닝 모델을 제안한다.

[0063] 도 4를 참조하면, 본 발명은 소스코드를 입력받아 자연어 요약문을 출력을 돕는 트렌스포머 기반의 구조 정보 학습 모듈이다. 기존 소스코드 요약 자동 모델들은 토큰 수준의 소스코드 정보를 학습하였으나, 본 발명이 제안하는 모듈은 구문 수준의 구조 정보를 학습하는 노드 인코더 모듈(즉, 제2 인코더)이 추가된다. 입력된 소스코드는 1차원의 벡터로 변환되어 인코더(즉, 제1 인코더)와 노드 인코더(즉, 제2 인코더)에 입력되며 각각 소스코드 토큰 수준의 잠재 표현과 구문 수준의 잠재 표현을 출력한다. 디코더는 두 잠재 표현의 출력의 입력받아 소스코드와의 관계를 학습한 후 요약문을 출력한다.

[0065] 1. 소스코드 분리 및 프로그램 의존 그래프(PDG) 정보 생성

[0066] 소스코드의 데이터 의존과 제어 의존에 따라 도 5에 도시된 바와 같은 프로그램 의존 그래프(PDG)를 생성한다. 여기서, 데이터 의존은 이전에 사용된 데이터가 다른 변수에 영향을 끼치는 것을 의미한다. 그리고, 제어 의존은 제어문 노드가 다른 변수에 영향을 끼치는 것을 의미한다. 소스코드 내에서 제어문이란 주어진 조건에 따라 수행 명령문을 결정하는 기능을 의미한다.

[0067] 도 5에 도시된 바와 같이, 주어진 소스코드에 대한 프로그램 의존 그래프(PDG)는 생성된 노드를 기준으로 소스코드를 분리한 '노드 정보'를 생성한다. 그리고, 분리된 노드 간 '간선 정보'를 생성한다. 간선 정보는 이후

에 구문 인코더 레이어(즉, 제2 인코더)에서 연결 노드간 어텐션(attention)을 수행하기 위해 사용된다.

[0069] 2. 입력 프로그램 소스코드와 요약 자연어 토큰의 어휘 사전 생성

[0070] 입력 프로그램 소스코드는 도 6에 도시된 바와 같이, 변수, 변수 유형, 키워드, 특수 문자, 함수, 리터럴 및 변수와 같은 다양한 종류의 토큰으로 구성된다. 이중에서 변수 유형, 특수 문자 등은 서로 다른 소스코드에서 공용으로 사용되는 어휘이다. 하지만, 변수는 프로그래머의 지정 어휘에 따라 달라진다. 따라서, 어휘의 전체 수는 어휘의 최대 빈도수를 기준으로 제한한다. 소스코드에서 어휘를 추출하는 기준은 노드 인코더 모듈(즉, 제2 인코더)과 앙상블될 기존 모델(즉, 제1 인코더)의 방식을 따른다.

[0072] 3, 기존 모델 인코더(즉, 제1 인코더)의 토큰 인코더 수행

[0073] 기존 모델 인코더(즉, 제1 인코더)의 소스코드 인코딩 방식들은 다르지만 토큰 임베딩을 입력받아 토큰 수준의 잠재 표현을 출력한다. 소스코드 자동 요약에서 임베딩이란 사람이 쓰는 자연어를 기계가 이해할 수 있는 숫자 형태인 벡터로 바꾼 결과를 의미한다. 잠재 표현이란 딥러닝에서 은닉층의 출력을 의미한다.

[0074] 기존 모델 인코더(즉, 제1 인코더)의 입력과 출력을 위한 표기는 다음과 같다. 소스코드의 길이가 m일 때 토큰 $t=(t_1, t_2, \ldots, t_m)$에 대해 t의 토큰 임베딩을 $t_e=(t_{1e}, t_{2e}, \ldots, t_{me})$로 표기한다. 그리고, 기존 토큰 인코더(즉, 제1 인코더)로부터 학습된 토큰 잠재 표현을 $c_e=(c_{1e}, c_{2e}, \ldots, c_{me})$로 표기한다. 최종적으로 출력된 토큰 잠재 표현은 노드 인코더(즉, 제2 인코더)의 잠재 표현 출력과 앙상블된다.

[0076] 4. 노드 인코더 모듈(즉, 제2 인코더)의 노드 풀러 수행

[0077] 노드 풀러는 과정 3의 토큰 임베딩을 입력받아 프로그램 의존 그래프(PDG)의 노드 잠재 표현(1)을 출력하는 함수이다. 노드 풀러란 노드에 포함되는 토큰들을 통해 노드를 나타내는 잠재 표현(1)을 얻는다는 뜻이다. 노드 잠재 표현(1)을 얻기 위해 토큰 임베딩 t_e와 과정 1의 노드 정보 MASK가 주어진다. 아래의 [수학식 1]은 노드 풀러의 수식이다. Node는 프로그램 의존 그래프(PDG)의 구문 수준의 노드를 의미하며, $Node_e$는 노드의 함수를 통해 출력될 잠재 표현(1)을 의미한다.

[0078] 먼저, MASK는 소스코드를 입력받는다. 만약 소스코드 내 토큰이 임베딩하고자 하는 노드에 포함돼 있다면 1 값을, 그렇지 않다면 0 값을 갖는다. MASK는 소스코드의 토큰에 대하여 Node에 대한 포함 여부를 가지고 있기에 m의 차원을 가진다.

수학식 1

[0079]
$$MASK = \begin{cases} 1 \;\; if \;\; t_j \in Node, for \quad j = 1, \ldots, m \\ 0 \;\; otherwise \end{cases}$$

[0080] 다음으로, 아래의 [수학식 2]와 같이, MASK를 통해 식별된 노드 내의 토큰은 훈련 가능한 가중치 W와 연산된 후 비선형 활성화 함수 ReLU를 거쳐 노드의 잠재 표현(1)을 생성한다.

수학식 2

[0081]
$$Node_e = ReLU((MASK \cdot t_e) \cdot \; W)$$

[0083] 5. 노드 인코더 모듈(즉, 제2 인코더)의 구문 인코더

[0084] 구문 인코더(즉, 제2 인코더)는 과정 4의 출력인 노드 잠재 표현(1)을 입력으로 받아 프로그램 의존 그래프(PDG)의 구조 정보를 학습한 노드 잠재 표현(2)를 출력하는 과정이다. 구문 인코더(즉, 제2 인코더)는 트랜스포머의 인코더를 기반으로 한다. 이때, 구문 인코더(즉, 제2 인코더)는 트랜스포머의 셀프-풀-어텐션(self-full-attention) 대신 프로그램 의존 그래프(PDG)의 그래프 간선의 어텐션(attention)을 사용한다. 아래의 [수

학식 3]은 구문 인코더의 간선 언텐션(attention) 수식이다. Key(K_e), Query(Q_e), Value(V_e)는 과정 4의 출력인 노드 잠재 표현(1)들의 모음이다. E_d는 프로그램 의존 그래프(PDG)의 데이터 의존 간선이며, E_c는 프로그램 의존 그래프(PDG)의 제어 의존 간선이다. E는 데이터와 흐름 의존 정보를 모두 포함한 간선이다.

수학식 3

$$Attention(Q_e, K_e, V_e) = softmax\left(\frac{E * Q_e K^{\top}_e}{\sqrt{d_k}}\right) V_e$$

$$E^{|Node| \times |Node|} = E_d + E_c$$

[0085]

[0087] 6. 디코더

[0088] 디코더는 기존 모델의 디코더 구조를 따른다. 하지만, 기존 모델 인코더(즉, 제1 인코더)의 출력인 C_e뿐만 구문 인코더(즉, 제2 인코더)의 출력인 노드 잠재 표현(2)들의 모음을 입력받는다. C_e와 노드 잠재 표현(2)는 연결되어 디코더에 입력된다.

[0091] 본 실시예들에 따른 동작은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능한 저장 매체에 기록될 수 있다. 컴퓨터 판독 가능한 저장 매체는 실행을 위해 프로세서에 명령어를 제공하는데 참여한 임의의 매체를 나타낸다. 컴퓨터 판독 가능한 저장 매체는 프로그램 명령, 데이터 파일, 데이터 구조 또는 이들의 조합을 포함할 수 있다. 예컨대, 자기 매체, 광기록 매체, 메모리 등이 있을 수 있다. 컴퓨터 프로그램은 네트워크로 연결된 컴퓨터 시스템 상에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다. 본 실시예를 구현하기 위한 기능적인(Functional) 프로그램, 코드, 및 코드 세그먼트들은 본 실시예가 속하는 기술 분야의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

[0092] 본 실시예들은 본 실시예의 기술 사상을 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0093] 100 : 소스코드 요약 장치,

110 : 프로세서,

130 : 컴퓨터 판독 가능한 저장 매체,

131 : 프로그램,

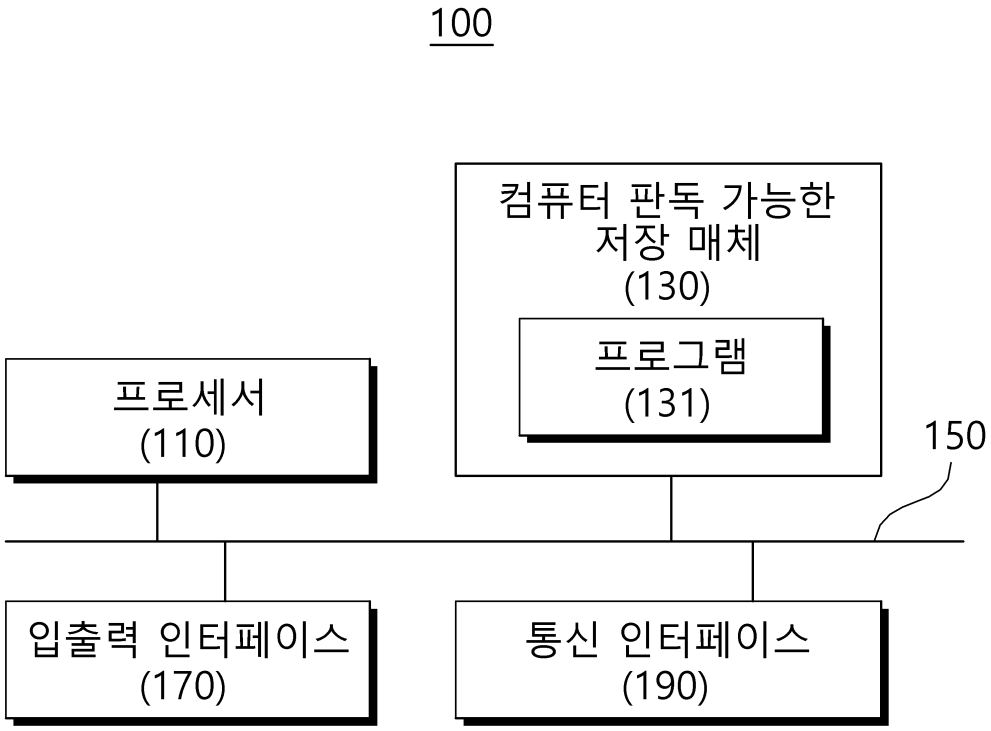
150 : 통신 버스,

170 : 입출력 인터페이스,

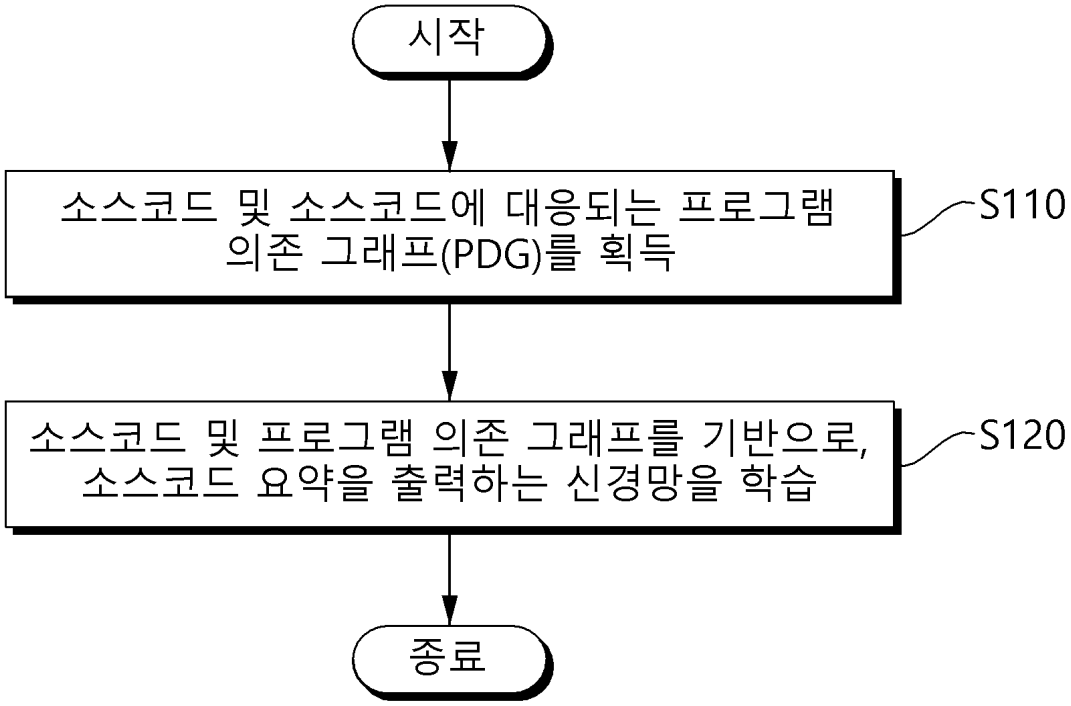
190 : 통신 인터페이스

도면

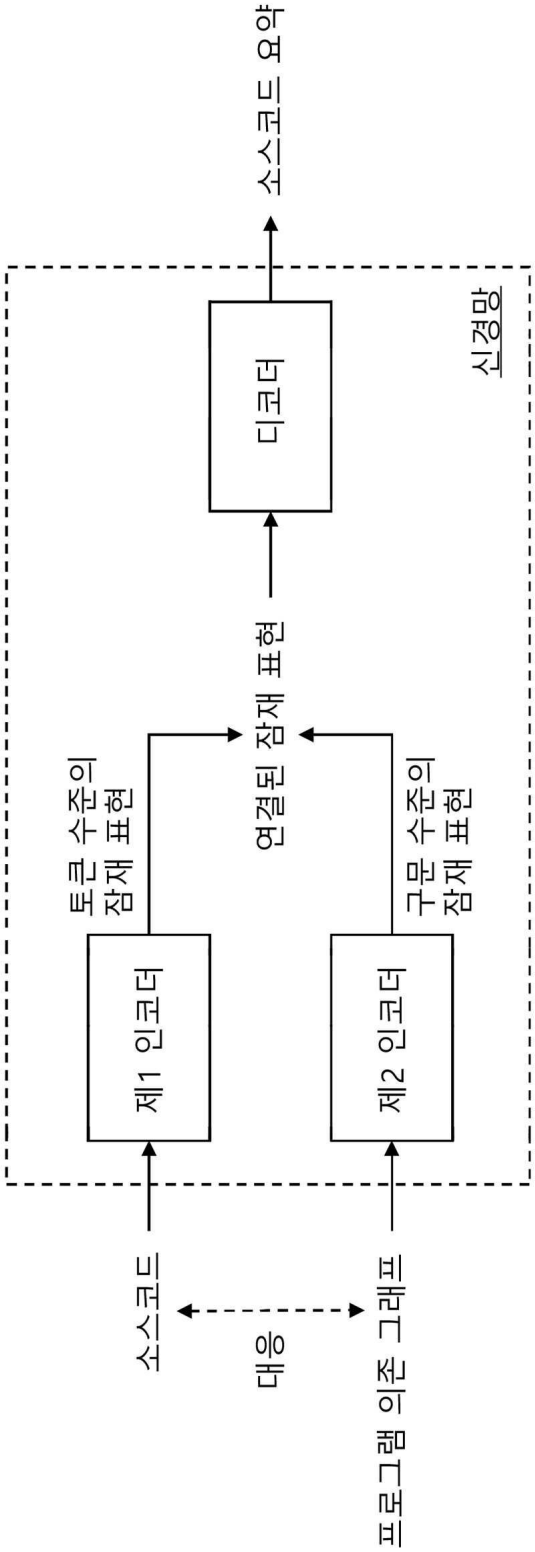
도면1


100
컴퓨터 판독 가능한
저장 매체
(130)
프로그램
(131)
프로세서
(110)
150
입출력 인터페이스
(170)
통신 인터페이스
(190)

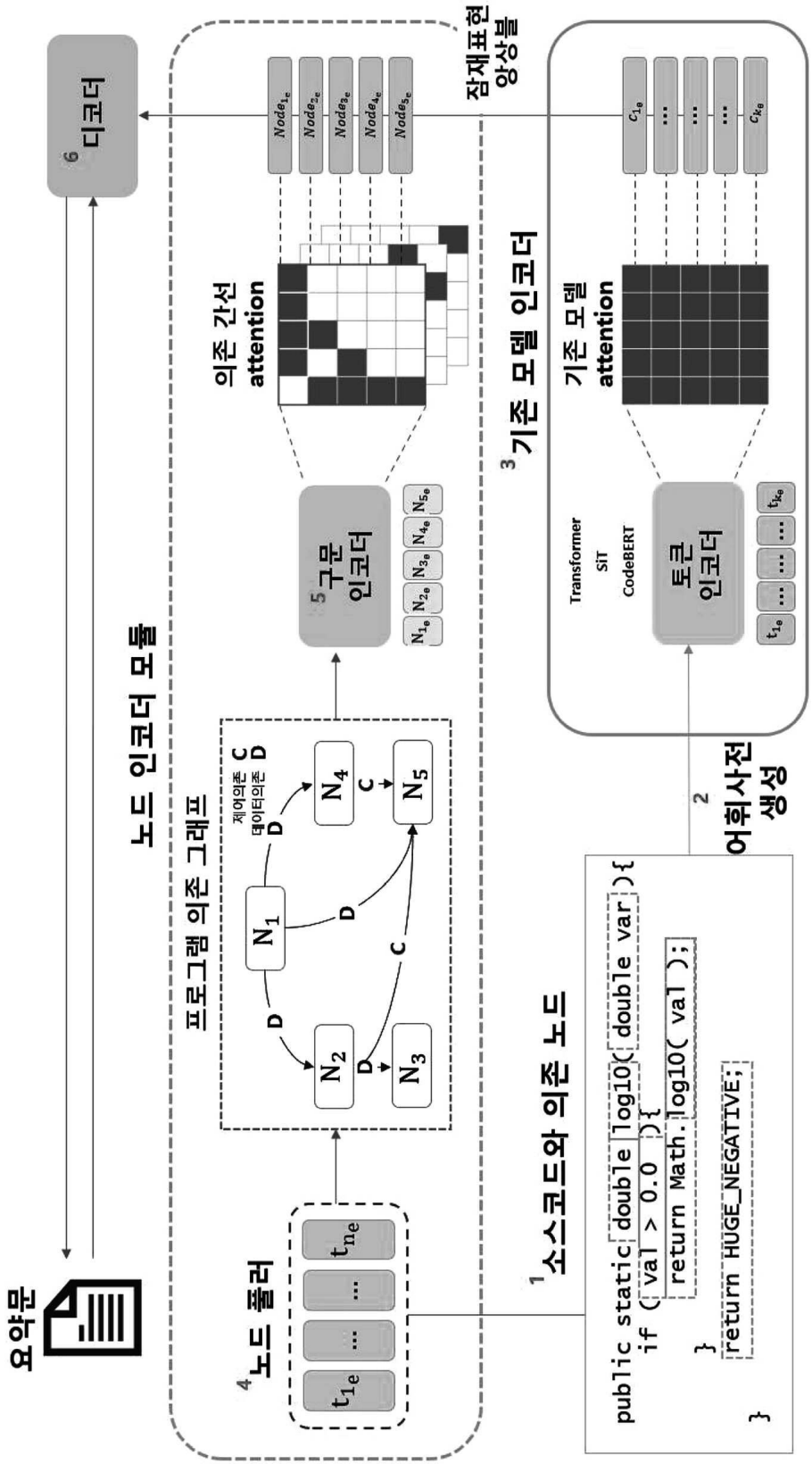
도면2


시작
소스코드 및 소스코드에 대응되는 프로그램
의존 그래프(PDG)를 획득
S110
소스코드 및 프로그램 의존 그래프를 기반으로,
소스코드 요약을 출력하는 신경망을 학습
S120
종료

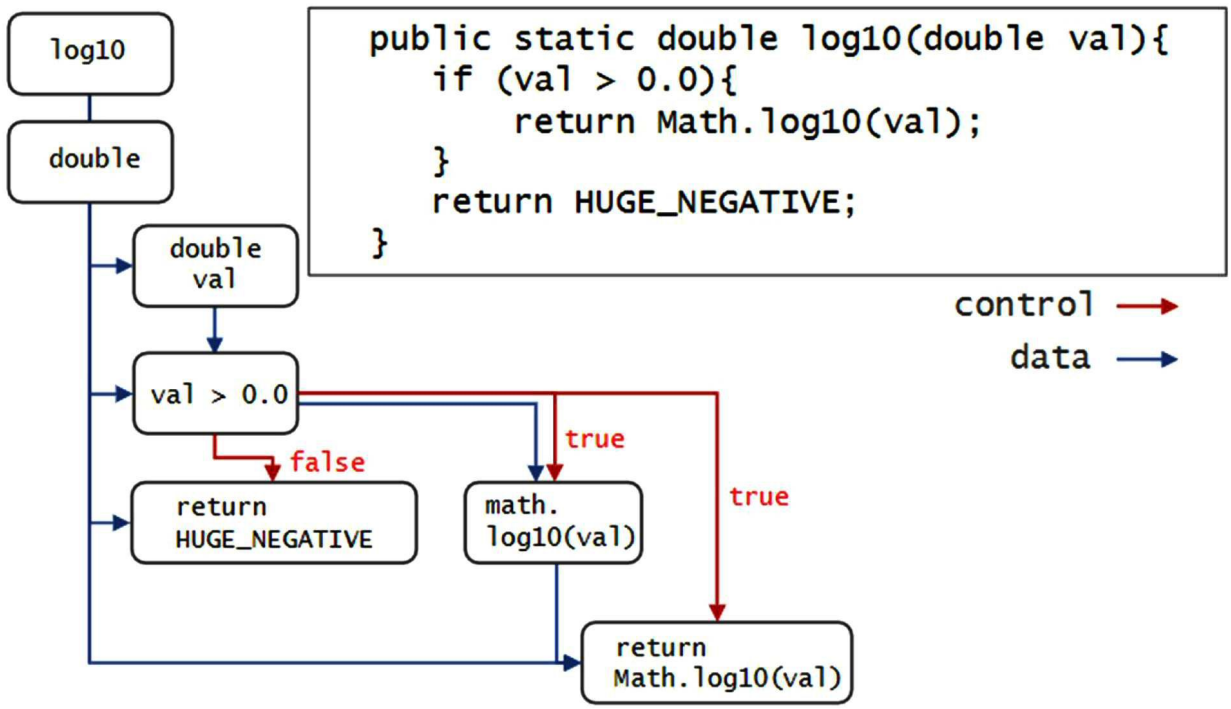
도면3


소스코드 요약
신경망
디코더
연결된 잠재 표현
토큰 수준의 잠재 표현
구문 수준의 잠재 표현
제1 인코더
제2 인코더
소스코드
대응
프로그램 의존 그래프

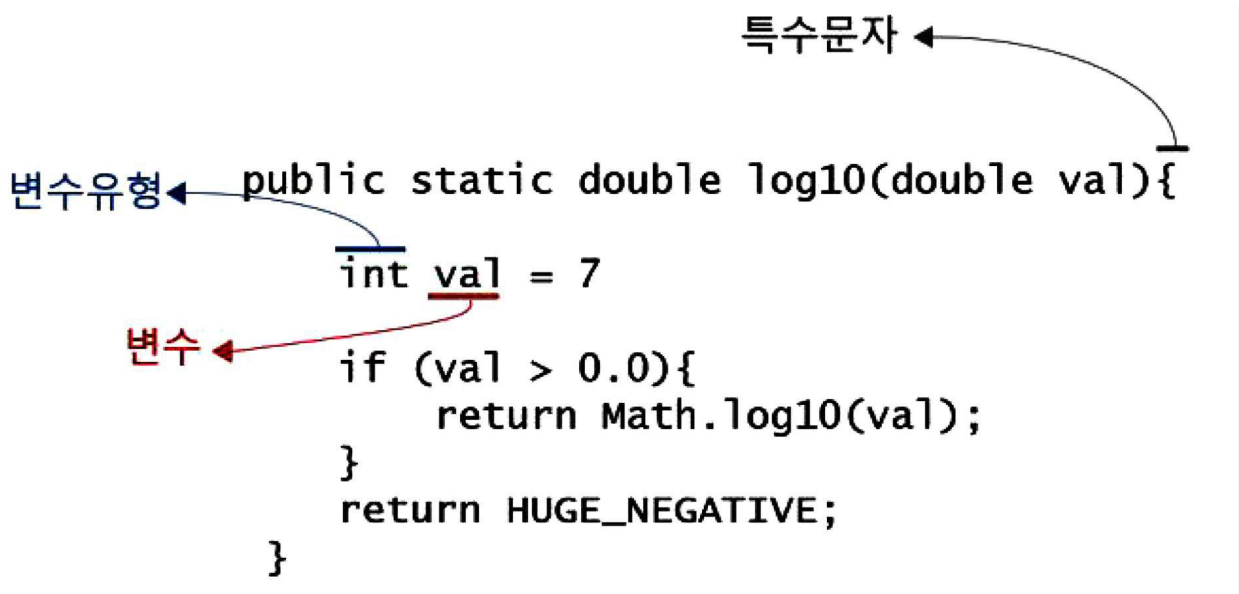
도면4


요약문
노드 인코더 모듈
6 디코더
프로그램 의존 그래프
제어의존 C
데이터의존 D
N1
N2
N3
N4
N5
4 노드 풀러
5 구문 인코더
의존 간선 attention
잠재표현 앙상블
3 기존 모델 인코더
Transformer
SiT
CodeBERT
토큰 인코더
기존 모델 attention
1 소스코드와 의존 노드
2 어휘사전 생성
public static double log10(double var){
if (val > 0.0){
return Math.log10(val);
}
return HUGE_NEGATIVE;
}

도면5


log10
double
double
val
val > 0.0
false
return
HUGE_NEGATIVE
true
math.
log10(val)
true
return
Math.log10(val)
public static double log10(double val){
if (val > 0.0){
return Math.log10(val);
}
return HUGE_NEGATIVE;
}
control
data

도면6


특수문자
변수유형
public static double log10(double val){
int val = 7
변수
if (val > 0.0){
return Math.log10(val);
}
return HUGE_NEGATIVE;
}