|  |
| --- |
| **SW프로젝트 제안서** |

|  |  |  |
| --- | --- | --- |
| **프로젝트명** | 딥러닝 기반 스펙트럼 번역 | |
| **프로젝트**  **유형** | 지정 주제 (O)  자유 주제 ( ) | 졸업 작품 ( )  졸업 논문 (O) |
| **프로젝트**  **요약** | 딥러닝 단백체 분야에서, 스펙트럼을 펩티드로 분석하는 문제를 언어 번역 문제로 전환하여 해결하고자 한다.  이에 최근 언어번역에서 기본이 되는 Transformer를 구현, 학습하며, 딥러닝 모델을 Transfer learning을 이용하여 스펙트럼 검증을 진행한다. 이를 이용한 어플리케이션 개발과, 결과물은 논문을 작성하여 발표하고자 한다. | |
| **R&D 산출물** | SW (), HW (), 특허 (), 논문 (O), 프로그램등록 () | |
| **지도교수** | 박 희 진 | |
| **예상기간** | 2022.02 ~ 2022.10 | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **전공** | **학번** | **학년** | **이름** | **연락처** |
| 컴퓨터소프트웨어학부 | 2017029889 | 4 | 심승현 | [imazine97@hanyang.ac.kr](mailto:kor0108silvia@gmail.com)  010-4895-3528 |
| 컴퓨터소프트웨어학부 | 2019017974 | 4 | 정희원 | [haley980406@naver.com](mailto:haley980406@naver.com)  010-2692-7220 |

**목 차**

**1. 프로젝트 배경 및 목표**

**2. 프로젝트 주요 내용**

**3. 추진 계획**

**4. 결론**

**5. 참고 문헌**

**1. 프로젝트 배경 및 목표**

단백체 분야에서 스펙트럼은 펩티드를 나타내는 데이터이다. 스펙트럼을 펩티드로 분석하는 문제를 언어 번역 문제로 전환하여 해결하는 방법을 모색하고자 한다. 이를 위해 딥러닝 모델을 사용하고자 한다.

Standard NN, CNN, RNN에 이어서 Transformer 라는 신경망 번역 모델이 구글에 의해 소개되었고, 스펙트럼을 2차원 벡터로 저장하는 방법이 발달함에 따라 이를 분석하는 문제를 해결하고자 한다.

매달 지도교수님과의 세미나를 통해 진행 상황을 점검한다. 구체적으로는 Keras 라이브러리를 학습하며, Transformer 를 학습/구현하고, 이를 이용하여 어플리케이션을 개발한다.

어플리케이션 개발과 더불어 프로젝트의 최종 결과물을 국내/외 학회에 발표하며 프로젝트를 마무리 할 예정이다.

**2. 프로젝트 내용**

프로젝트에는 Python을 이용하여, Keras API를 주로 사용할 예정이고, Transformer을 구현 및 학습하여 스펙트럼 검증을 진행할 예정이다.

이 프로젝트에서는 대표적인 신경망 라이브러리 중 하나인 Keras를 사용하게 된다. Keras 는 프랑수아 숄레가 2015년, 연구 프로젝트의 일환으로 개발한 딥러닝 API이다. Apache MXNet, Core ML, Javascript, Typescript, PlaidML 등에서 실행할 수 있으며, 텐서플로에서는 tf.keras 를 번들로 포함시켜 이용할 수 있다. 이를 통해 많은 종류의 신경망을 만들고, 훈련 및 평가, 실행할 수 있다.

two dimensional gel electrophoresis 에서, 단백질을 등전위점과 분자량의 2차원으로 분리하여 2D-gel (스펙트럼) 을 얻는다. 펩티드는 문자열이므로 2차원 벡터를 문자열로 번역하는 문제로 전환하여 접근한다.

이에 Transformer 모델을 사용한다. Transformer 모델은 2017년 구글에 의해 소개된 딥러닝 모델으로, Attention mechanism 과 몇몇 임베딩 층, 정규화 층 등의 구성 요소들만을 사용해 Neural Machine Translation 문제에서 최고 수준 성능을 크게 향상시킨 모델이다. 특히 이 모델은 병렬화화 학습 속도에 있어서 큰 이점이 있어 최근 많은 프로젝트에서 사용하고 있다.

이 프로젝트에서 스펙트럼을 분석하는 문제를 언어 번역 문제로 전환하며 진행하므로, Transfer Learning (전이 학습) 방식을 사용하여 진행한다. 전이 학습이란, 특정한 환경에 맞춰 개발된 머신 러닝 모델 및 데이터를 다르지만 비슷한 분야에 적용하여 학습시키는 방식이다. 전이 학습을 사용하였을 때는 기존의 모델을 이용한다는 점에서 시간을 절약할 수 있다는 장점을 가진다.

**3. 프로젝트 추진 계획**

매달 세미나를 통해 모델 구현 및 학습 결과를 발표할 계획이며, 1학기인 3, 4, 5월에는 Keras 및 Transformer 모델에 대한 학습 위주로 시행하며, 6,7,8월은 Transformer을 구현하고, 어플리케이션 개발에 중점을 둔다. 2학기가 시작되는 9월부터는 과제의 결과물을 논문으로 작성하여, 국/내외 학회에 발표하는 것을 목표로 둔다.

프로젝트는 지도교수이신 박희진 교수님, 프로젝트를 도와 주실 이상정 조교님과 팀원 2명 ( 컴퓨터소프트웨어학부 심승현, 정희원 ) 으로 이루어져 있다.

이에 필요한 학습을 각자 분담하여, 정기적인 회의를 통해 서로에게 공유하며 효율적인 학습을 도모할 계획이다.

**4. 결론**

현재 코로나 시대를 맞이하며 이를 극복하려는 과정에서, 코로나 바이러스의 단백질 구조와 발생 가능한 돌연변이 등을 미리 컴퓨터가 예측하여 화제가 되고 있다. 이에 따라 현재 딥러닝의 생물학에의 응용에 많은 이목이 집중되고 있는데, 산업의 최전선에 있는 분야에 직접 뛰어들어 역량을 향상시킨다는 의미가 있을 것이고, 성장해나가는 개발자로서 도전적인 과제를 해결하며 학습 의욕을 향상시키고자 한다.

화제가 되고 있는 딥러닝 분야에 관한 깊이 있는 이해를 기대하며, 바이오인포매틱스 등 분야에의 응용을 직접 경험하여 프로그래밍 역량 향상이 기대된다. 또한 학부 생활을 하면서 흔치 않은 협업 기회를 얻은 만큼, 협업 능력을 향상시킬 수 있는 기회가 될 것이다.

**5. 참고 문헌**

* Cs231n, Transfer Learning, https://cs231n.github.io/transfer-learning/, 2022년 3월 9일
* 권경훈, 김경환. 탄뎀 질량 스펙트럼을 이용한 단백질 검색 알고리듬 개발, 한국과학재단, 2004년 10월 20일
* 오렐리앙 제롱 지음, 박해선 옮김, 핸즈온 머신러닝 2판, 한빛미디어, 2020년 5월 4일