**졸업 프로젝트 지정 과제 공고문**

|  |  |  |  |
| --- | --- | --- | --- |
| **과제명** | 자연어 처리(NLP) 모델 Adversarial Attack 분석・개선 | | |
| **과제 내용** | 자연어 처리 모델은 Text Classification, Natural Language Inference, Sentence Paraphrase Task 처럼 다양한 분야에서 활용되고 있다. (비고 1, 2 참조) 그리고 사람이 보기에는 자연스럽지만, 모델의 입력에 약간의 변화를 주어 모델이 오동작하는 입력을 생성하는 Adversarial Attack 분야도 활발히 연구되고 있다. 특히, 이러한 Adversarial Attack은 이미지 모델에 대해서 많이 연구가 되어져왔다.  최근 자연어 처리 모델 중 문장 분류기 모델을 공격하는 방법으로 CLARE(비고 3)가 있다. CLARE에는 입력을 분류하는 자연어 처리 모델(e.g. Bert Base)과 이러한 문장 분류기 모델이 오동작하는 입력을 변환하는 문장 변환 자연어 처리 모델(e.g. RoBERTa)이 존재한다. CLARE는 문장의 문맥을 고려하면서 Mask-and-Infill 알고리즘으로 Token(e.g. 단어의 단위)을 Replace, Insert, Merge 기법을 적용시켜 공격하는 방법이다. Mask-and-Infill 방법은 문장 변환 자연어 처리모델을 활용하여 입력의 특정 Token에 Mask를 씌우고, 해당 부분을 다른 대안의 Token으로 채워 공격을 수행한다.   * Replace : Mask된 부분에 오동작이 되도록 하는 Token을 대체 * Insert : Mask된 부분에 오동작이 되도록 추가 Token을 주입 * Merge : 인접한 두 단어를 Mask로 처리하고 이를 하나의 단어로 치환   예를 들면, 문장에 대한 긍정/부정을 평가하는 Task가 있고, “The Food has always been consistently good.”이라는 문장이 주어진다면, 문장 분류기 모델은 해당 문장을 긍정으로 평가한다. 하지만, 문장 변환 자연어 처리 모델을 통해서 문장이 “The Food has always been looking consistently good.”로 변환한다면 문장 분류기 모델은 해당 문장을 부정으로 평가한다.  본 프로젝트에서는 위와 관련된 논문과 소스코드를(비고 3) 분석하고, 이를 바탕으로 자연어 처리 모델에 대한 Adversarial Attack 방법을 개선하는 것을 목적으로 한다.   1. 기존의 Mask-and-Infill에 새로운 Operation 추가  * Delete : 주어진 문장에서 문맥적 혹은 문법적으로도 이상이 없는 일부 Token을 지워서 공격을 수행 * 이외에도 자유롭게 Operation 추가 가능  1. Gradient Ascent와 Explainable AI(XAI)와 같은 기법의 적용  * 기존의 방법처럼 모든 Token에 변화를 주기보다 입력의 어느 부분이 민감한지 파악하고, 특정 부분을 우선적으로 변환함으로써 정확하고 빠르게 공격을 수행 가능  1. 문장 변환 자연어 처리 모델의 Fine-Tuning  * CLARE 논문에서 문장 변환 자연어 처리 모델은 Pre-Trained된 모델을 사용함 * 문장 분류기 모델이 오동작하도록 하는 입력을 정확하게 생성하기 위해 문장 분류기 모델의 정보(e.g. Layer들의 Feature 정보, Gradient Ascent를 통한 입력의 Gradient 값)를 활용하여 문장 변환 자연어 처리 모델을 Fine-Tuning  1. 기존 이미지 기반의 Adversarial Attack 기법의 응용  * 기존의 이미지 기반의 Adversarial Attack 기법(e.g. FGSM, IGSM, C&W)을 자연어 처리 모델에 어떻게 적용할 수 있을지에 대한 수행 가능   필요로 하는 지식   * 자연어 처리 모델에 대한 이해 * DNN 모델의 Adversarial Attack에 대한 이해 (비고 4 참고) * Cosine Similarity 에 대한 이해 * PyTorch 활용 방법 | | |
| **평가 방법** |  | | |
| **비고** | 1. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://aclanthology.org/N19-1423.pdf)  2. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](https://openreview.net/attachment?id=SyxS0T4tvS&name=original_pdf)  3. [Contextualized Perturbation for Textual Adversarial Attack](https://arxiv.org/pdf/2009.07502.pdf) ( [소스 코드](https://github.com/cookielee77/CLARE) )  4. [Explaining and Harnessing Adversarial Examples](https://arxiv.org/pdf/1412.6572.pdf)  5. Pytorch 등 딥러닝 프레임워크에 대해 공부할 수 있는 자료 제공  6. Adversarial Attack에 대해 공부할 수 있는 자료 제공  7. 자연어 처리 모델에 대한 간단한 교육 제공(겨울방학기간 동안) | | |
| **프로젝트 유형** | A형:산업체 발주 주제 ( )  B형:산업체 참여 정부과제의 주제 ( )  C형: 산업체 수요 과제 ( ) | **참여 인원** |  |
| **지도교수** | 서지원 | **연락처** | 02-2220-2400 |

**한양대학교 공과대학 컴퓨터소프트웨어학부**