AuToeic : Auto-making exams of TOEIC part 1

팀장 : 엄소은

팀원 : 송규원, 이재우, 이성균

작성일자 : 2023.05.23.

1. AuToeic 소개

토익(TOEIC) LC Part 1은 음성으로 제공되는 선지들 중 주어진 사진을 가장 잘 묘사하는 문장을 선택하는 문제들로 이루어져 있다. 토익 시험은 월 2회 치루어지기 때문에 많은 수의 문제가 출제되어야 한다. 특히 사진이 포함된 해당 Part 문항들의 경우 사진 촬영에 드는 인력과 비용이 많이 들 것으로 예상된다. 하지만 문제 변별력은 주로 Part 3,4,7에서 이루어지고 Part 1은 비교적 쉬운 난이도의 문제들이기 때문에 이렇게 많은 인력과 비용을 쏟는 것은 비효율적이다. 따라서, 단어 키워드 혹은 문장을 입력하면 사진을 자동으로 생성하고 그에 맞는 정답 선지와 오답 선지들을 만드는 AuToeic 모델을 개발하였다. AuToeic이란, '토익' 문제를 '자동'으로 출제한다는 점에서 'Auto'와 'TOEIC'을 결합하여 만든 명칭이다.

본 모델은 CV (Computer Vision: 컴퓨터 비전)와 NLP (Natural Language Processing: 자연어 처리)를 결합한다. 컴퓨터 비전이란, 컴퓨터가 이미지 혹은 비디오와 같은 시각적 자료를 이해하고 학습하도록 하는 분야이다. 자연어 처리란, 컴퓨터가 인간의 언어를 이해하고 생성할 수 있도록 하는 분야이다. 따라서 사진을 생성하기 위해서는 컴퓨터 비전을, 선지를 생성하기 위해서는 자연어 처리 방법을 이용한다. 본 모델은 기존에 존재하는 모델의 일부를 가져와 사용하는 전이학습 방식을 사용한다. 이때 사용한 전이학습은 세 가지 단계로 구성되어 있다.

첫 번째 단계인 Stage 1에서는 'Stable Diffusion 1.5'라는 모델을 사용하여 문제를 출제하기 위해 필요한 실사에 가까운 흑백 사진을 생성한다. Stage 2에서는 BLIP 모델의 Capfilt을 사용하여 정답 선지 1개를 만든 후 난이도 조절을 위해 매력적인 오답 선지 1개, 그 외에 일반 오답 선지 2개를 생성한다. 이때, 사진 생성을 위해 여러 개의 이미지들과 그에 맞는 설명이 담긴 데이터셋을 확보하고 이를 모델에 학습시킨다. 실제 토익 Part 1 기출문제에 사용된 500개의 사진, 정답 선지, 오답 선지들도 확보하여 이 중 480개는 모델을 학습시키는 데에, 20개는 모델의

성능을 평가하는 데에 사용한다. 마지막 Stage 3에서는 CLIP의 contrastive learning과 zero-shot prediction을 통해 사진과 선지들 간의 적중률, 즉 정답률을 계산하여 문제의 적합성을 평가한다. 본 모델은 많은 수의 모의고사를 적은 비용으로 빠른 시간 내에 만들어내야 하는 사설 교육업체에서 유용하게 사용될 것으로 기대된다.

2. AuToeic에서 응용한 모델 3개의 특장점

1) Text to Image: **Stable Diffusion 1.5**

Compvis, Stability AI, LAION에서 만든 Stable Diffusion 모델은 사용자가 입력한 텍스트를 토대로 이미지를 만들어주는 모델이다. Latent Diffusion이라는 원리로 작동하는데, 이것은 1) 학습한 이미지 데이터셋에 무작위적인 Gaussian noise를 늘려가다가 2) 다시 줄여가면서 처음에 입력받은 텍스트에 상응하는 이미지를 생성하는 과정을 일컫는다. Stable Diffusion은 지금까지 공개된 데이터셋 중 가장 규모가 큰 LAION-5B 데이터셋으로 훈련되었고, 이미지 생성 모델 중 SOTA(State of the art, 현존하는 최고 수준 성능의 모델)를 달성했다. 더불어 Pixel-space diffusion model보다 메모리와 연상 요구량을 획기적으로 줄였다.

본 프로젝트에서는 Stable Diffusion 1.5 모델을 기반으로 사실적인 이미지들을 만들어주는 Dreamlike Photoreal 2.0 모델을 사용하였다. 이 모델은 기본 stable diffusion 1.5 모델에 초고화질 (768*768px) 이미지셋을 추가하여 훈련한 모델이다. 토익 Part 1의 사진은 실제 일상과 관련된 상황으로 구성되기 때문에 AuToeic은 이질적이지 않은 사진이 생성해야 한다. 따라서 직접 찍은 사진인지 만들어낸 사진인지 모를 정도의 이미지를 생성할 수 있는 Dreamlike Photoreal 2.0을 활용하였다.

2) Image to Text : **BLIP**

BLIP은 2022년 공개된 Vision Language Pretraining-model이다. 웹 크롤링으로 수집한, noise가 많은 이미지-텍스트 쌍을 데이터셋으로 학습한 모델인데, 이때 이미지에 상응하는 caption을 bootstrapping하여 노이즈가 많은 웹 데이터를 효과적으로 활용한다.

BLIP은 captioner + filter로 이루어져있는데, captioner는 caption을 합성하고, filter는 그것에서 noise를 줄이는 역할을 한다. BLIP 은 3가지 loss function 을 사용하여 학습을 진행하는데, 이것은 기존에 있는 Image to Text model 의 단점들을 보완한다. BLIP 의 가장 중요한 점은 captioner 과 filter 가 caption을 bootstrapping 하면서 성능을 높였고, 더욱 다양한 caption 을 생성한다는 점이다.

3) Evaluation : **CLIP**

OpenAI에서 2021년 1월 5일에 공개한 CLIP(Contrastive Language-Image Pre-training) 모델은 이미지에 가장 적절한 텍스트를 출력해주는 생성모델이다. BLIP과 유사하게도, CLIP은 Pre-training 과정에서 CLIP은 웹 크롤링을 통해 4억 개의 이미지와 연관 텍스트를 추출하여 거대한 데이터셋을 스스로 구축하였다는 점에서 효율적이다. 이미지 수집과 정답 label 생성에 많은 노력이 필요하지 않기 때문이다.

수집한 데이터를 특징 벡터로 변환한 후(Encoder), CLIP은 '대조 학습(contrastive learning)'이라는 방법으로 Pre-training 과정을 진행한다. 대조 학습이란 정답인 (이미지, 텍스트) 각 특징 벡터 간의 높은 코사인 유사도와 오답인 (이미지, 텍스트) 각 특징 벡터 간의 의 낮은 코사인 유사도를 모두 구해서 올바른 연결 관계를 학습하는 방법을 일컫는다. 이렇게 학습한 뒤, 새로운 이미지를 입력하고 정답/오답 텍스트를 여러 개 입력하여 정답을 골라내도록 했을 때 CLIP은 매우 높은 확률로 정답 텍스트를 선정해냈다. 많은 실험을 거친 결과, CLIP이 학습 과정에서 배우지 않은 일을 해내는 것(zero-shot learning)에 효과적이라는 사실이 밝혀졌다.

3. AuToeic의 가치 : 어떻게 토익 LC Part 1에 특화했나?

1) 경제성

토익(TOEIC) 시험의 출제기관은 미국 교육기업 ETS(Education Test Service)이다. 시험의 공정성을 위해 ETS는 정기시험에 한 번 출제된 문항을 이후 정기시험에 다시 출제하지 않는다. 우리나라에서는 1982년부터 YBM 한국 TOEIC 위원회가 주관하였고, 현재 2주마다 1회씩 전국 시험장에서 시험이 시행되며 대중적인 공인영어능력시험으로 자리잡았다.

토익 수험자는 청해 100문항(Part 1~4)과 독해 100문항(Part 5~7)을 120분간 해결해야 하는데, 그 중 Part 1(1번~6번, 총 6문항)은 성우가 들려주는 4개의 문장 선지 중 문제지에 주어진 상황 사진을 가장 적절하게 설명하는 선지를 고르는 유형이다. 난이도는 마지막 6번 문항을 제외하면 다소 낮은 편이며, 앞서 언급한 바와 같이 한 번 출제된 문항은 다시 출제되지 않으므로 ETS는 매번 새로운 상황의 사진을 촬영하고 선지를 출제해야 할 것으로 예상된다. AuToeic 모델은 이때 사진과 선지를 보다 적은 비용으로 빠르게 생성하는 데에 효과적이다. 모델이 스스로 사진을 생성하므로 저작권 문제에서 자유롭고, 문항의 정확도와 난이도의 객관성을 편리하게 보장할 수 있기 때문이다. 따라서 특히, 토익 수험교재를 출판하거나 정기시험 출제경향에 맞춰 교육 콘텐츠를 생성해야 하는 사설 교육업체(해커스어학연구소, YBM, 뤼이드AI 산타토익 등)에게 Part 1 출제에 들이는 비용을 줄이고, 변별력을 갖춰야 하는 Part 4 또는 Part 7의 출제에 더욱 집중할 수 있는 기회를 제공할 것으로 예상된다.

2) Fine Tuning 방법

본 프로젝트에서는 기존에 있는 모델을 그대로 사용하지 않고, 직접 모은 토익 이미지들과, 공개된 이미지-텍스트 데이터셋을 기반으로 더욱 토익 문제 출제에 적합하도록 Fine Tuning 을 하였다.

- **Stable Diffusion 1.5**

  우리가 실제로 수집한 토익 이미지-정답+오답 500개의 문제들을 사용하였다. 이미지-텍스트 데이터를 사용하여 Dreamlike Photoreal 2.0 을 Fine Tuning 했다.

- BLIP

  위의 모델과 마찬가지로 우리가 실제로 수집한 토익 이미지-정답+오답 500개의 문제들을 사용하였다. 실제 토익 문제 데이터를 8:2 의 비율로 train, test 으로 나누고, train set 으로 학습을 진행하였고 Test set 으로 평가하였다.

3) AuToeic만의 구조적 특징

- **3가지의 모델을 하나의 과정으로**

StableDIffusion1.5 모델은 입력된 키워드를 잘 표현하는 image를 생성하는데 장점을 보이고, BLIP 모델은 captioner와 filter를 사용하여 상당한 양의 데이터를 학습하여 image를 captioning 하는 것에 장점을 보이며, CLIP 모델은 제공되는 caption 중 image를 가장 잘 설명하는 caption을 선택하는 것에 장점을 보인다.

AuToeic은 토익 LC Part 1에 자주 등장하는 소재에 대해 Dreamlike Photoreal 2.0 모델을 활용하여 실제 토익에 출제되는 이미지와 유사한 이미지를 만들었으며, BLIP을 활용하여 위 과정을 통해 만들어진 이미지에 대한 적절한 선지를 만들어냈다. 더불어 AuToeic이 만들어낸 선지가 정말 정답으로서 역할을 할 수 있는가, 중복 정답은 존재하지 않는가에 대하여 CLIP 모델을 통해 선지의 가치를 평가하였다.

StableDIffusion1.5 모델을 사용하여 직접 찍은 듯한 이미지를 만드는 과정, BLIP 모델을 사용하여 image에 대한 captioning을 진행하는 과정은 각각 존재할 수 있으나 하나의 과정으로 묶어 토익 LC Part 1 문제 출제를 목적으로 하여 최적화 하는 과정은 이전에 존재하지 않던 새로운 시각과 방향성을 제시한다.

- **BLIP 과 CLIP 모델을 독창적으로 활용**

BLIP 모델을 통해 이미지를 captioning을 하는 과정은 이미 다른 프로젝트에서도 많이 활용되고 있다. 대부분의 프로젝트에서 이미지에 대해 옳은 caption을 생성하는 방향으로 BLIP 모델을 학습시킨다. 반면, Autoeic은 의도적으로 적당한(매력적인) 오답의 caption을 생성하도록 유도하는데에도 BLIP의 원리를 활용하였다는 점에서 다른 프로젝트와 구분된다. 정답 선지가 너무 자명하거나 오답 선지가 완벽하게 오답일 경우 문제의 가치가 떨어지기 때문에, 같거나 유사한 소재, 어구, 동사 등을 포함하는 매력적인 오답 선지가 필요하기 때문이다.

더불어 CLIP의 대조 학습(contrastive learning) 방법을 1개의 사진과 4개의 caption(선지) 간 연관성을 “평가”하는 데에 응용한 것은 AuToeic이 최초로 시도하는 것으로 파악된다. 생성한 사진과 정답/오답 caption 각각의 특징 벡터를 구한 후 계산한 상호 유사도는

caption이 사진을 얼마나 구체적으로 설명하는지에 대한 확률값이다. 따라서 유사도는 수험자들이 각 선지를 선택할 확률로 이해할 수 있으며, 문항의 정확도와 난이도의 평가 지표가 된다. 이렇게 문항을 평가하는 과정은 실제 수험자가 문항을 해결하는 과정과 상당히 유사하므로, 모델이 출제한 문항을 사람이 직접 정성적으로 평가할 필요가 없으며 문항 출제 자동화에 크게 기여하였다.

BLIP 모델을 통해 정답 선지 1개와 얼핏 들으면 정답으로 착각할 매력적인 오답 선지 3개를 생성하고, CLIP 모델을 통해 BLIP 모델이 생성한 정답 선지가 적절한지, 오답 선지 중 정답 선지로 볼 여지가 있는 선지가 있는지를 판단하는 과정을 자동화하는 것은 문항을 단순히 생성하는 것을 넘어서 문항으로서 가치를 지니고 있는지를 스스로 평가할 수 있도록 한다. 기존의 모델 내 일부 원리를 응용하여 토익 part 1에 특화되도록 했다는 점, 그리고 그러한 시도가 지금까지 발견되지 않았다는 점이 AuToeic 프로젝트만의 고유한 가치이다.

4. 실제 실행 결과

문제1.

입력 Text: Black and white photo, Some bicycles have been left unattained

생성 이미지:

실제 문제 선지 4개 (정답 1개 + 오답 3개)

	some bicycles are parked near a wall	a bike is leaning against a column.	the door is no longer on the hinges	a person sits on a bicycle and makes it move
image 1	99.16	0.09	0.00	0.75

문제2.

입력 Text: black and white photo, a man looking at a computer

생성 이미지

실제 문제 선지 4개 (정답 1개 + 오답 3개)

he's typing on a computer	the man is picking up a laptop	the man is typing on his laptop	the shirt is being fastened securely
20.48	15.45	63.97	0.04

5. 보완점

본 모델은 연세대학교 응용통계학과 소속 데이터사이언스 학회 'Data Science Lab' 의 'Hairy Potatoes' 조 모델링 프로젝트 결과물이다(2023.03.09~2023.04.06). 1달여 간의 짧은 기간 동안 진행한 프로젝트이기에 모델을 개선시킬 요소들이 많다. 따라서 추후에 특허로 발전 시키기 위해 보완할 수 있을 점들을 정리해 보았다. 특허로 발전할 만한 가능성이 보인다면, 아래 보완할 점들을 통해 모델의 효율성과 성능을 발전시킬 수 있을 것이다.

1) **더 많은 데이터 확보**: 모델의 성능 향상을 위해 더 많은 이미지-설명 데이터 및 기출문제 데이터를 확보하는 것이 중요하다. 이를 통해 모델이 더 다양한 키워드 혹은 문장에 해당하는 이미지를 생성하고 더 정확한 정답 선지 및 오답 선지들을 생성할 것으로 기대된다.

2) **이미지 유사도 검증**: 이미지 생성 단계에서 생성된 이미지가 실제 토익 이미지와 유사한지 검증하는 단계를 추가하는 것을 제안한다. 이미지 유사도 측정 기법을 활용하여 생성된 이미지와 실제 토익 이미지 간의 유사도를 평가하고, 유사도가 높은 경우에만 해당 이미지를 사용하는 방식으로 개선할 수 있다.

3) **생성된 정답 선지와 실제 정답 선지 유사도 비교**: 생성된 정답 선지와 실제 토익 정답 선지 간의 유사도를 비교하는 단계를 추가할 수 있다. 텍스트 유사도 모델을 사용하여 생성된 정답 선지들이 유사도가 높은 경우에만 문제를 활용하는 방식으로 개선할 수 있다.

4) 입**력한 키워드/문장과 생성된 정답 선지 유사도 비교**: 사진을 생성할 때 입력한 단어 키워드나 문장과 본 모델을 통해 생성된 정답 선지와의 유사도를 비교할 수 있다. 유사도가 높은 경우 사진 생성과 선지 생성 모두 잘 이루어졌다는 것을 증명할 수 있다.

5) **Negative Prompt 활용**: Negative Prompt는 사진을 생성할 때 들어가기를 원하지 않는 요소들을 명시해주는 것이다. Negative Prompt를 추가로 활용하여 생성된 이미지를 자연스럽게 만드는 것이 가능하다.

6) **선지 다양성 추가**: 현재 모델은 정답 선지와 매력적인 오답 선지를 제외한 나머지 두 개의 오답 선지를 생성하기 위해 동일한 데이터로 훈련시킨 모델을 사용한다. 이를 개선하기 위해 서로 다른 모델을 사용하여 두 개의 오답 선지에 차이를 주는 방법을 고려할 수 있다.

7) **선지 문법 구조 학습**: 모델의 마지막 단계에서 NLP 모델 하나를 추가로 사용할 수 있다. 실제 토익 선지들의 문법 구조를 학습하여 생성된 선지들을 동일한 문장 구조로 수정하는 모델을 도입할 수 있다. 이를 통해 문장 구조도 통일하고 문법적으로 올바른 선지를 생성할 수 있다.

8) **Text-audio**: 발음의 종류를 지정해주면 (ex. 영국식, 미국식, 등) 생성된 문제를 읽어주는 기능을 추가할 수 있다. 이를 통해 실제 토익 시험과 유사한 환경을 조성하고 학습자들의 실전 경험을 향상시킬 수 있다.

6. 참고문헌

<모델>

Stable Diffusion 1.5

- **논문**: High-Resolution Image Synthesis with Latent Diffusion Models - https://arxiv.org/abs/2112.10752
- **사용 모델** : Dreamlike-photoreal-2.0 - https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0

- **모델 License**

  This model is licesed under a modified CreativeML OpenRAIL-M license.

  - You are not allowed to host, finetune, or do inference with the model or its derivatives on websites/apps/etc. If you want to, please email us at contact@dreamlike.art
  - You are free to host the model card and files (Without any actual inference or finetuning) on both commercial and non-commercial websites/apps/etc. Please state the full model name (Dreamlike Photoreal 2.0) and include the license as well as a link to the model card (https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0)
  - You are free to use the outputs (images) of the model for commercial purposes in teams of 10 or less
  - You can't use the model to deliberately produce nor share illegal or harmful outputs or content
  - The authors claims no rights on the outputs you generate, you are free to use them and are accountable for their use which must not go against the provisions set in the license
  - You may re-distribute the weights. If you do, please be aware you have to include the same use restrictions as the ones in the license and share a copy of the modified CreativeML OpenRAIL-M to all your users (please read the license entirely and carefully) Please read the full license here: https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0/blob/main/LICENSE.md

BLIP

- **논문**: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation - https://arxiv.org/abs/2201.12086
- **사용 모델**: Salesforce/blip-image-captioning-large

CLIP

- **논문**: Learning Transferable Visual Models From Natural Language Supervision - https://arxiv.org/abs/2103.00020
- **사용 모델**: openai/clip-vit-base-patch32
- **Github** : https://github.com/openai/CLIP