뉴스 영상 생성 AI

김선무[1], 이승준[1], 이정원[1], 박지혜[2]
[1]울산대학교 IT융합학부, [2]현대엘리베이터
tjsan199813@gmail.com, tmdwns778@naver.com,
gardeny98@naver.com, nata85@naver.com

News Image Generation AI

Seon-moo Kim[1], Jeong-won Lee[1], Seung-jun Lee[1], Ji Hye Park[2]
[1]Dept. of Computer Science, University of Ulsan, [2]Hyundai Elevator

요 약

뉴스와 같은 정확한 정보를 제공하는 영상을 제작하는 과정은 많은 자원과 시간이 소요된다. 작성된 기사를 이용하더라도 영상 기반의 뉴스를 제작하는 것은 인적, 시간적인 자원의 투여가 불가피하다. 뉴스를 송출하기 위해 소요되는 시간을 줄이기에 현실적으로 어렵다. 따라서 우리는 이러한 문제를 해결하고 빠른 뉴스 영상 제공이 가능한 "뉴스 영상 생성 AI"를 개발하기로 하였다.

1. 서론

최근 AI의 중요성은 나날이 증가하고 있다. AI를 활용한 업무의 자동화는 인적 노동력과 시간, 비용의 절감 효과를 거둘 수 있고 이는 생산성 향상과 아울러 업무 효율의 향상을 기대할 수 있다. 인공지능 및 영상처리 기술의 발전과 함께 이를 활용한 다양한 영상 컨텐츠를 생성하는 분야가 대두되고 있다. '뉴스 영상 생성 AI'는 빠르고 신뢰성이 있는 텍스트 기반으로 출판된 뉴스를 바탕으로 원하는 인물의 목소리를 포함하여 시청자에게 정확한 정보를 전달할 수 있는 영상 매체를 생성함으로서 정보 전달을 목적으로 하는 다양한 분야에 활용할 수 있다.

본 논문은 뉴스 본문을 입력하여 음성으로 변환하는 TTS (Text-To-Speach) 기술을 사용하여 본문을 특정 인물의 목소리로 읽게 하고, 뉴스 본문의 중요 키워드를 추출하여 키워드에 대한 이미지 검색으로 뉴스 영상에 들어갈 관련 이미지들을 수집, 영상 합성 기술을 사용하여 생성한 음성과 이미지를 합성하여 최종 뉴스 영상을 생성하는 프로그램을 소개한다.

2. 적용 기술

2.1. Glow TTS & Hifi Gan

음성합성 모델은 자기회귀적 음성합성 모델과 비자기회기적 음성학성 모델 두 가지로 분류할 수 있다. 비자기회귀적 음성합성 모델은 자기회귀적 모델의 느린 음성합성 속도를 개선하기 위해 만들어진 모델로, Glow-TTS [1] 가 대표적인 모델이다. 이런 음성 합성 모델은 입력받은 텍스로부터 멜 스펙트로그램을 생성하는 역할을 한다. Glow-TTS는 자기휘귀적 음성 합성 모델인 Tacotron2에 비해 약 15배 더 빠르게 음성합성을 하고 품질도 비슷하다는 장점이 있다. 음성합성 모델로부터 생성된 멜 스펙트로그램을 후속처리를 통해 음성으로 최종 변환을 해야하는데 이런 역할을 수행하는 모델을 Vocoder라고 부른다. HiFi GAN은 GAN을 이용한 보코더이다. 기존 GAN을 사용하면 샘플링과 메모리 사용 효율의 개선으로 인한 빠른 음성합성 속도의 장점을 가지고 있었지만, 음성합성 품질은 자기회귀 모델에 비해 떨어지는 단점이 있었다. 하지만 HiFi-GAN은 자기회귀모델에 필적하는 높은 품질과 자기회귀모델에 비해 빠른 속도를 기대할 수 있다는 장점이 있다.

2.2. Utagger

형태소 분석이란 자연 언어 분석의 첫 단계로서 단어(한국어의 경우 어절)를 구성하는 각각의 형태소들을 인식하고 불규칙 활용이나 축약, 탈락 현상이 일어난 경우 원형을 복원하는 과정을 말한다. U Tagger(유태거) [2] 는 한국어 형태소 분석기과 동형이의어 분별을 동시에 수행하며, 품사태그는 세종

테그셋을 사용한다. 동형이의어 번호 체계는 세종을 기준으로 하며, 대체로 국립국어원의 표준국어대사전과 일치한다. 기본적으로는 세종말뭉치를 학습하여 작동하며, 다른 도메인에 대한 특화 기능을 "사용자 말뭉치"라는 기술로 제공하고 있다."사용자 말뭉치"를 이용하여 신조어, 용언의 활용형, 인접 두 어절 간의 새로운 문맥을 실시간으로 학습한다.

2.3. TextRank

TextRank 알고리즘 [3] 은 그래프 기반 랭킹 알고리즘으로 PageRank에 기반 하여 문서 내의 중요 키워드 추출을 할 수 있습니다. 문서를 그래프로 표현하여 Edge 간의 관계를 이용하여 Vertex의 중요도를 계산하는 알고리즘이다.

TextRank는 단어 또는 구문의 집합을 최종 출력으로 하는 Keyword Extraction 방식과 중요 문장을 출력으로 하는 Sentence Extraction 방식이 있다. 본 논문에서는 Keyword Extraction 방식을 사용하여 중요 키워드를 추출하였다.

2.4. Google Custom Search API

Google 이미지 검색을 지원하는 OpenAPI 이다. 검색하려는 검색어를 쿼리로 API를 호출하면 결과를 JSON 형태로 반환하여 검색어에 대한 이미지와 기타 검색 결과를 얻을 수 있다. Restful API, Python 라이브러리 등 다양한 방법으로 접근할 수 있다.

2.5. OpenCV

OpenCV(Open Source Computer Vision)은 실시간 컴퓨터 비전을 목적으로 한 프로그래밍 라이브러리다. 원비전 뿐만 아니라 일반 영상 처리 분야에서도 많이 사용되고 있으며, 카메라 애플리케이션에서도 OpenCV가 사용되기도 한다. 이 라이브러리는 윈도우, 리눅스 등에서 사용 가능한 크로스 플랫폼이며 오픈소스 BSD 허가서 하에서 무료로 사용할 수 있다. C++와 Python에서 지원을 하지만 본 프로젝트에서는 Python 버전을 사용하였다.

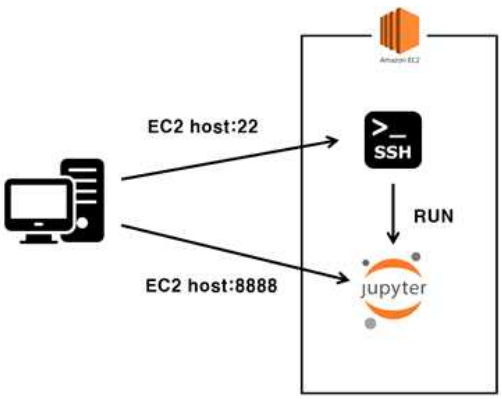
2.6. FFmpeg

FFmpeg (www.ffmpeg.org) 은 비디오, 오디오, 이미지를 쉽게 인코딩 (Encoding), 디코딩 (Decoding), 먹싱 (Muxing), 디먹싱 (Demuxing) 을 할 수 있도록 도움을 주는 멀티미디어 프레임워크이다. 인코딩이란 우리가 문서의 용량을 줄이기 위하여 zip 프로그램을 사용해서 문서를 압축하는 것처럼 동영상이나 이미지의 용량을 줄이기 위해서 압축하는 과정을 의미한다. 디코딩이란 zip으로 압축된 워드 문서를 보기 위해서 먼저 zip 프로그램을 압축을 해제해야 하는 것처럼, 압축된 동영상을 재생하기 위하여 압축을 해제하는 과정을 디코딩이라고 부른다. 엔지니어링 분야에서 먹싱 (Muxing)이라는 단어는 여러 입력을 하나로 합치는 과정을 의미하고 디먹싱 (Demuxing)이라는 과정은 하나로 합쳐진 입력을 다시 여러 출력으로 만드는 것을 의미한다.

3. 구현

3.1. 시스템 구성

시스템 구성은 그림 1과 같이 AWS에서 지원하는 EC2 가상환경 컴퓨팅 환경을 사용하였다. EC2 인스턴스는 Ubuntu OS를 사용하고 4vCPU (NVIDIA Tesla M60), 30.5GB의 메모리 성능을 가지고 있다. EC2에 주피터 노트북 환경을 구축하여 로컬 컴퓨터에서 접속하여 개발, 테스트를 진행했다.

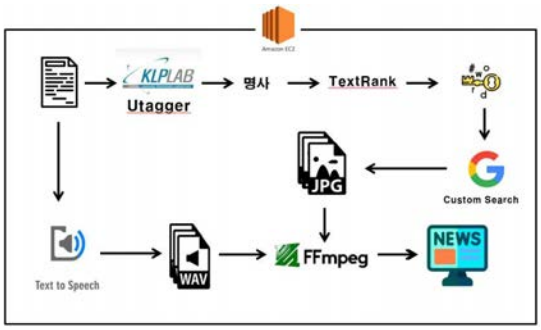


(그림 1) 하드웨어 구성

3.2. 시스템 흐름

서비스는 입력으로 뉴스 본문을 입력받아 최종적으로 뉴스 영상을 출력한다. 최초 입력인 뉴스 본문을 Glow TTS와 Hifi GAN 기술을 이용하여 음성 파일로 변환 시킨다. 입력된 뉴스 본문에 대해 Utagger(형태소 분석기)를 사용하여 명사를 추출하고, 추출된 명사에 대해 TextRank 알고리즘을 이용하여 중요한 키워드를 추출한다. 이 키워드에 대해 Google Custom Search API로 이미지 검색을 수행하여 뉴스 내용을 대표할 수 있는 이미지들을 수집한다. 최종적으로 FFmpeg와 OpenCV를 사용하여 생성한 음성과 수집한 이미지 파일을 합쳐 최종 출력인 뉴

스 영상을 생성하여 출력한다.



(그림 2) 프로세스 흐름도

3.3. 음성 학습

Glow TTS와 Hifi Gan의 이용은 음성을 생성하기 위해 특정 인물의 목소리를 학습 시켜야 한다. 본 논문에서는 KSS(Korean Single Speaker Speech) 데이터셋, 가수 배철수 (배철수의 음악캠프), 아나운서 손석희 (JTBC 앵커룸) 의 목소리를 학습 시켜 해당 목소리로 기사를 음성으로 변환할 수 있도록 하였다. 데이터 수집 과정에서 화자의 음성을 한 문장씩 끊어 샘플링하고, 해당 음성이 말하는 내용을 레이블로 제공해야 한다. 레이블링 과정에서 STT (Speech To Text) 모델을 사용하여 음성을 텍스트로 변환시켜 사용하였다.

4. 결론

본 작품의 기대효과는 세 가지로 말할 수 있다. 첫 번째는 다양한 영상 콘텐츠를 만들 수 있다는 것이다. 뉴스나 기사만이 아닌 다양한 정보 전달을 위한 유튜브 영상, 책 등을 읽어주는 비디오 북의 영역까지 확장하여 영상 생성이 가능하다. 두 번째 기대효과는 뉴스 속보처럼 영상 정보를 빠른 시간 내에 불특정 다수의 많은 사람에게 알려야 하는 상황에 '뉴스 영상 생성 AI'는 상황과 시간에 관계없이 빠르게 대응할 수 있다는 것이다. 마지막으로 영상 생성 과정의 많은 시간과 비용을 절약할 수 있다. '뉴스 영상 생성 AI'는 기존의 뉴스 생성 방식과 달리 스튜디오와 카메라 등이 필요 없으므로 물리적인 시간, 비용 측면에서 굉장한 경쟁력이 있을 것이다.

본 논문에서 '뉴스 영상 생성 AI'는 KSS, 배철수, 손석희의 음성을 이용하고 한국어 서비스만 제공하고 있다. 그러나 서비스에 적용하는 목소리와 언어를 더욱 다양하게 제공하고 이를 사용자가 선택할 수 있도록 하면 더욱 폭 넓은 서비스가 가능하다. 또한 뉴스 스크립트에서 뽑아낸 키워드로 Google Custom Search API를 이용해 수집한 이미지 자료만을 이용하여 뉴스를 생성하는 것이 현재의 한계이나 추후 영상자료를 추가하여 뉴스를 생성할 수 있도록 확장할 수 있다. 부가적으로 현재 "뉴스 영상 생성 AI'는 CLI로 실행되는 프로토타입으로 영상 생성 서비스를 제공하고 있어 접근성이 떨어진다. 이를 Web App이나 Mobile App으로 제작하면 사용자의 접근성을 높일 수 있다.

※ 본 프로젝트는 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] Kim, Jaehyeon, et al. "Glow-tts: A generative flow for text-to-speech via monotonic alignment search." Advances in Neural Information Processing Systems 33 (2020): 8067-8077.

[2] 신준철, and 옥철영. "기분석 부분 어절 사전을 활용한 한국어 형태소 분석기." 정보과학회논문지: 소프트웨어 및 응용 39.5 (2012): 415-424.

[3] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.