AI는 미디어 경험을 어떻게 바꿀 것인가?

글.
전윤호 알티캐스트 AI사업TF

AI에 대한 관심이 식을 기미를 보이지 않는다. 이세돌을 이긴 알파고, 스피커를 넘어 다양한 제품으로 확산되고 있는 음성 비서, 자율 주행 자동차, 앨런 머스크 같은 사람들이 인류 멸종 걱정에 잠을 설치게 만드는 AI 등. 사실 여기서 언급한 예들은 서로 공통점이 그리 많지 않으나 연일 쏟아지는, 과장되고 곧잘 오도되는 AI 관련 뉴스를 접한 사람들은 어느 날 갑자기 만능의 AI 기술이 발명되어 조만간 인간이 할 수 있는 일을 대부분 AI가 대치하게 되지 않을까 걱정한다.

물론, 본지를 읽고 있는 독자라면 기술적 지식과 경험을 바탕으로 이러한 뉴스들을 걸러서 이해하고 나름의 전망을 하고 있겠지만, 어쨌거나 딥러닝과 같은 기술들이 최근에 빠르게 발전하면서 새로운 응용이 활발히 이루어지고 있는 것이 사실이므로 좀 더 차분하게 이러한 AI 관련 기술의 현 시점에서의 수준과 응용 사례를 미디어 산업을 중심으로 살펴보는 것도 의미가 있을 것이다.

요즘 AI에 대해 심도 있는 논의가 어려운 이유 중의 하나는, AI라고 할 때 각자 다른 것을 떠올리기 때문이다. 원래 AI의 정의는 모호한 것이다. 사전적으로는 기계에 의한 지능 또는 이를 연구하는 학문이라고 되어 있는데 여기서 모호한 것은 지능이다. 어느 정도 수준이어야 단순한 기계의 자동화된 기능이 아니라 "지능"이라고 할 수 있을까? 지능의 정의는 움직이는 표적(moving target)이다. 카메라로 숫자만 인식해도 AI의 영역으로 생각하던 때가 있었으나 지금은 모든 주차장에 그런 기능이 설치되어 있다. 언젠가는 자동차가 스스로 운전해도, "그건 그냥 자동 운전일 뿐"이라고 치부할 때가 올 것이다.

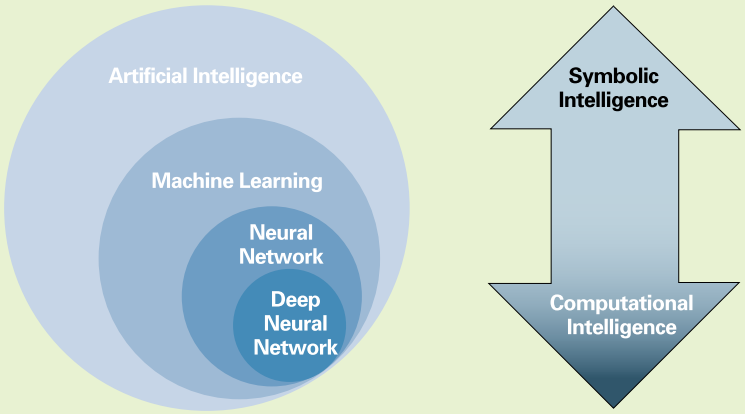

그림 1. AI, 머신러닝, DNN의 관계

머신러닝에 기반한 AI, 특히 최근 주목받는 딥러닝에 기반해야만 AI일까? 물론 그렇지 않다. AI는 앞서 말했듯이, 그때까지 기계엔 어렵고 인간만 할 수 있던 일을 새롭게 기계가 할 수 있게 되는 것인데, 기계가 주어진 데이터에 의하여 적절한 내부 파라미터나 구조를 스스로 찾도록 하는 기계학습(머신러닝)은 AI를 구현하기 위한 주요한 방법이지만 지능적인 기계를 구현하는 유일한 방법은 아닐뿐더러, 대부분의 머신러닝이 이를 학습시키는 사람의 경험과 직관으로 여러 하이퍼파라미터와 아키텍처를 결정해야 하기에 완전히 스스로 학습하는 것만도 아니다. 또 최근의 딥러닝과 같이 모든 것을 숫자로 바꿔 학습하고 추론하는 방식도 있으나 여전히 텍스트를 심볼릭하게 다루는 방식도 유효하다.

본 기고에서는 AI의 전반적인 트렌드나 특정한 기술을 다루기보다는 미디어의 생산과 소비에 이러한 AI 기술이 응용되고 있는 몇 몇 사례와 향후 전망에 대해 얘기하고자 한다.

콘텐츠의 인식

AI 관련 기술 중 최근에 가장 획기적으로 발달한 딥러닝 기술이 대표적으로 적용되는 분야가 이미지 인식인 만큼, 멀티미디어 콘텐츠에서 사람이나 사물을 인식하여 활용하기 위한 시도가 활발히 이뤄지고 있다.

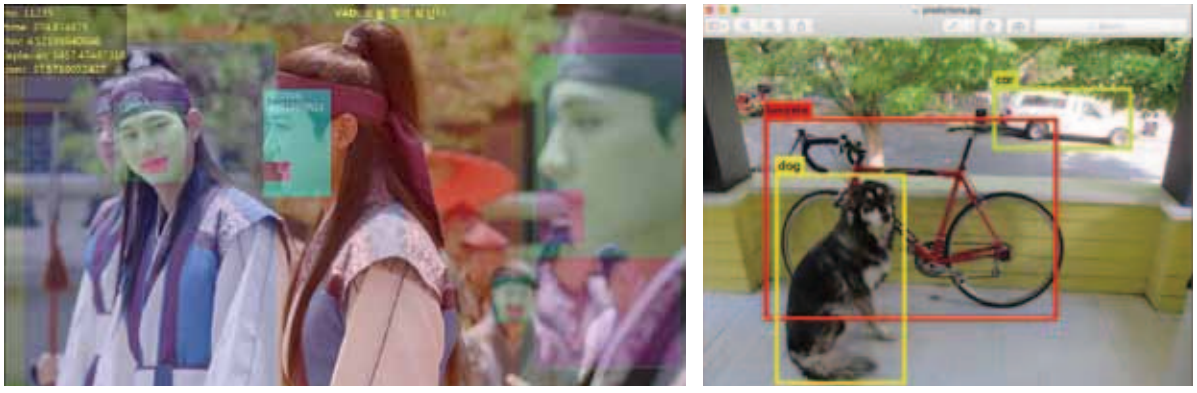

그림 2. 동영상에서 얼굴과 사물의 인식

물론 과거에도 얼굴의 탐지(detection)와 특정인을 인식(recognition)하는 기술은 있어왔고 여러 형태로 활용되었으나 ConvNet(Convolutional Neural Network)과 같이 이미지 인식에 탁월한 성능을 발휘하는 딥러닝의 오픈소스 구현, ImageNet(www.image-net.org)과 같은 큰 규모의 레이블된 이미지 데이터베이스 및 이를 이용하여 이미 학습까지 이뤄진 네트워크 데이터들이 공개됨에 따라 다양한 응용 시도가 이뤄지고 있으며 비디오로부터 이미지 인식, 자막과 오디오 분석 등을 통해 메타데이터를 추출하여 이벤트나 주제, 감정, 마이크로장르에 의한 검색, 추천 등의 서비스를 제공할 수 있다. 다만 이러한 이미지의 인식을 위해서는 대용량의 이미지 데이터베이스와 GPU 가속하드웨어를 이용하여 DNN(Deep Neural Network)을 학습시켜야 할 뿐만 아니라, 학습된 DNN을 이용하여 추론(inference)하기 위해서도 일반적인 임베디드 프로세서로 처리하기에는 버거운 프로세싱 성능과 전력이 필요하기 때문에 TV나 셋톱박스와 같은 소비자 디바이스에 이러한 기능이 직접 탑재되어 실시간으로 콘텐츠를 인식하는 것은 전용의 뉴로모픽(neuromorphic, 뇌신경모방) 하드웨어가 저렴하게 보급되는 수년 후에야 가능할 것으로 보이며, 아직은 서버에서 사전에 학습, 인식된 결과를 제공하는 형태로 서비스가 가능하고 픽셀(piksel), 아리스(Arris), 코미고(Comigo) 등의 회사들이 이와 같은 기술을 제품화하고 있다.

콘텐츠의 생성과 변형

멀티미디어 데이터를 인식하는 것 외에, 생성하거나 변형하는 용도로도 딥러닝 기술이 응용되고 있다. 이는 DNN이 데이터를 인식

전윤호
알티캐스트 AI사업TF

하기 위한 학습 과정에서 스스로 적절한 특징 검출기(feature detector)를 만들어낸다는 점을 이용한 것으로써 이러한 특징 검출기를 거꾸로 "있을 법한" 데이터의 생성에 활용함으로써, 비록 사람처럼 상위 수준의 컨셉을 이해하고 다양한 지식과 경험에 의해 창작을 하는 것은 아니지만 적어도 기존의 스타일 혹은 특징을 흉내 내고 재조합하는 형태의 새로운 콘텐츠를 만들어낼 수 있게 되었다.

그림 3. 스타일 트랜스퍼의 예 : 콘텐츠 소스 + 스타일 소스 = 생성된 이미지

대표적인 것이 스타일 트랜스퍼(StyleTransfer)라고 불리는 기술로써, 스타일 소스 이미지의 특징 분포를 따르지만 콘텐츠 소스 이미지와 동일한 콘텐츠를 가지는 결과 이미지를 생성하는 것이다. 이를 위하여 딥러닝 계열의 기술 중에서도 특히 최신의 GAN(Generative Adversarial Network)과 같은 기술들이 활용되고 있다.

그림 4. 최근의 StyleTransfer 예 (출처 : arxiv.org/pdf/1705.01088.pdf)

미디어 타입의 변환

이와 같은 DNN에 의한 인식과 변형 기술을 활용하면 좀 더 나아가서 미디어의 타입을 바꾸는 것도 가능해진다. [그림 5]는 알티캐스트에서 개발 중인 Picturize의 컨셉으로써, DNN 및 기타 미디어 인식 기술을 활용하여 동영상과 자막을 이미지와 말풍선 형태로 변환하고, 사람의 편집 과정을 거친 후에 필터를 적용하여 만화 형태로 변환하는 것이다.

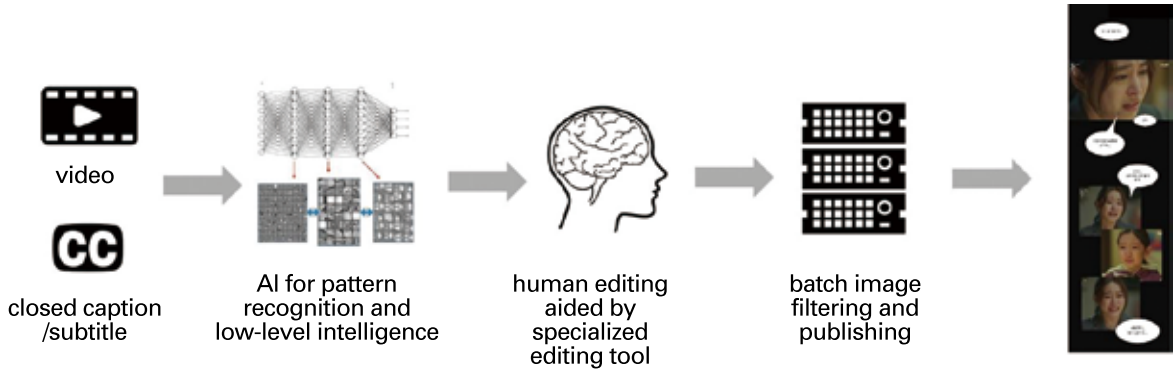

그림 5. Picturize 컨셉

이 과정에서 특히 기술적으로 어려운 문제는 동영상으로부터 적절한 프레임을 선택한 후 자막의 대사가 화면 상의(혹은 화면 밖의) 누가 한 말인지를 파악하여 적절한 말풍선을 생성하는 것이다. 이를 위하여 화면 전환, 얼굴의 감지와 인식, 입술의 움직임, 화자 인식 등의 기술을 적용하고 있으며 산출물의 품질을 향상시키기 위해 사람에 의한 편집과 만화 스타일을 위한 기존 방식 및 스타일 트랜스퍼 필터를 이용한다.

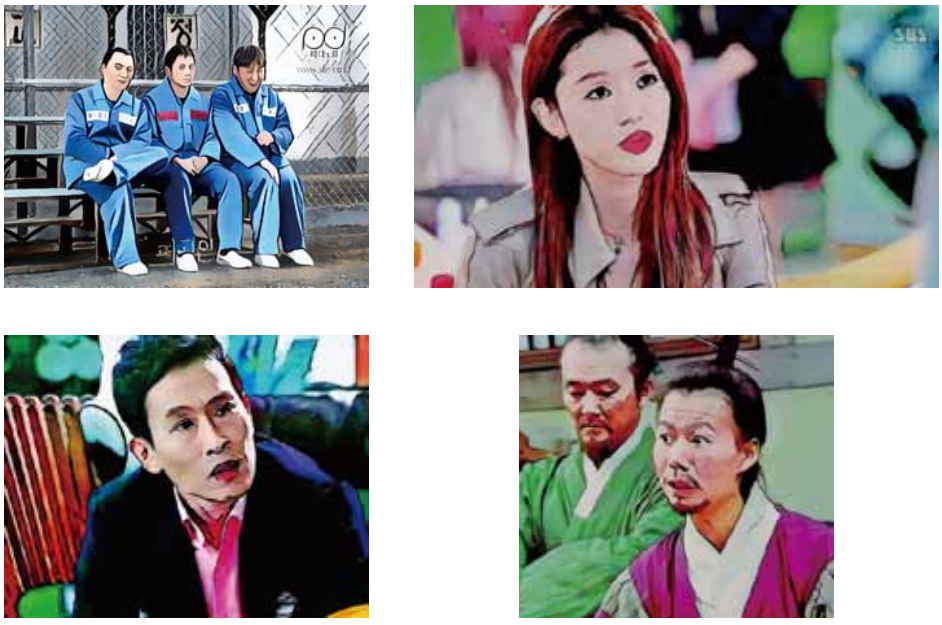

그림 6. 만화풍으로 변환하는 스타일 트랜스퍼의 예

유저 인터페이스를 위한 AI

AI 기술이 미디어 경험에 이용될 수 있는 또 하나의 분야는 TV 혹은 셋톱박스의 유저 인터페이스이다. 이미 AI 스피커가 TV 셋톱박스와 연동되거나 혹은 셋톱박스 내에 내장되어 TV 시청을 음성으로 제어할 수 있게 되었다. 이는 원거리 마이크(far-field microphone) 기술과 함께 딥러닝에 의하여 음성인식 성능이 향상되었고 [그림 7], 사용자가 구사하는 다양한 표현을 처리할 수 있는 자연어처리 기술이 적용되었기에 가능해진 것이다. TV를 통해 제공되는 서비스가 다양해지고 IoT 디바이스가 확산되면서 리모컨이나 전용앱을 대체하는 음성 인터페이스의 필요성은 점점 늘어날 것이다. 음성 명령과 질문을 인식하기 위하여 클라우드상의 서버로 보내지는 음성 데이터가 축적되고 다시 학습에 활용되면서 더욱 음성인식률을 높이는 효과를 가져오고 있다. 또한 클라우드에 의존할 수 없는 호출어 ('알렉사' 등 wake-up word) 인식에도 디바이스에서 실행되는 딥러닝이 활용되고 있다. 자연어처리 분야에서도 점차 데이터에 기반한 딥러닝이 기존 방식의 기술을 보완하여 성능을 향상시키고 있으나 아직 지식 Q&A 및 맥락을 이해하는 대화의 영역에서는 실용적으로 활용되지 못하고 있다.

하지만 음성인식률의 향상에도 불구, 실제로 AI 스피커를 사용해보면 아직도 불편한 점이 꽤 있고 각종 이용통계에서도 음악 재생 등 기본적인 기능 외의 활용률은 높지 않은 것으로 나타나고 있다(참고 : voicelabs.co/2017/01/15/the-2017-voice-report). 음성 입출력과 제한된 자연어처리 능력만으로는 사용자들에게 기존 기능의 사용법과 새로운 기능을 학습시키는 것이 용이하지 않고 여러 단계를 거쳐야 하는 예약, 구매 등의 서비스를 수용하기가 어렵기 때문이기도 하며, 또 주로 시각에 의한 비언어적 시그널을 활용하는 사람과 달리 청각 정보에만 의존하는 AI 스피커가 정황을 파악하는 데에도 한계가 있기 때문이다. 이에 TV 셋톱박스와 결합된 모델(KT 기가지니)이나 작은 디스플레이를 내장한 모델(아마존 Echo Show)과 같은 제품들이 등장하고 있으며, 카메라를 통해 사용자

전윤호
알티캐스트 AI사업TF

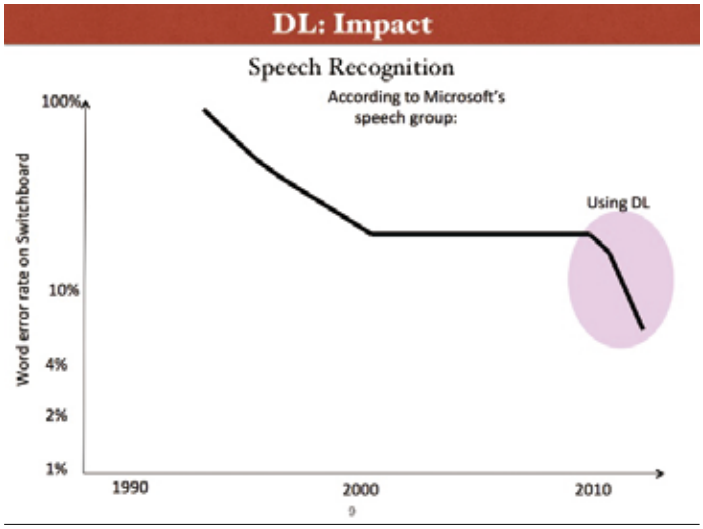

그림 7. 딥러닝에 의한 음성인식률 향상 / 출처 : Roelof Pieters, Deep Learning Intro for NLP Meetup Stockholm

의 얼굴이나 제스처를 인식하고자 하는 시도도 이뤄지고 있다. 전통적으로 사용자의 제스처를 인식하기 위해서는 마이크로소프트의 키넥트와 같은 depth 센서가 많이 쓰였고 최근 다양한 depth 센서의 성능과 가격이 개선되고 있기는 하나, TV-소파 간 거리에서 사용자의 작은 손놀림을 인식하기에는 해상도가 낮다. 최근에는 역시 딥러닝에 의해 하나의 2D 카메라로부터 제스처와 포즈를 인식하는 것이 가능해지고 있으며 (예 : www.cmu.edu/news/stories/archives/2017/july/computer-reads-body-language.html) 전술한 DNN 전용 하드웨어의 도입과 함께 실용화될 것으로 기대되고 있다.

멀티모달 UI

AI 스피커가 많은 관심을 받고 있으나 앞서 기술했듯이 단순히 음성 입력과 출력만으로는 복잡한 과제를 수행하거나 많은 정보를

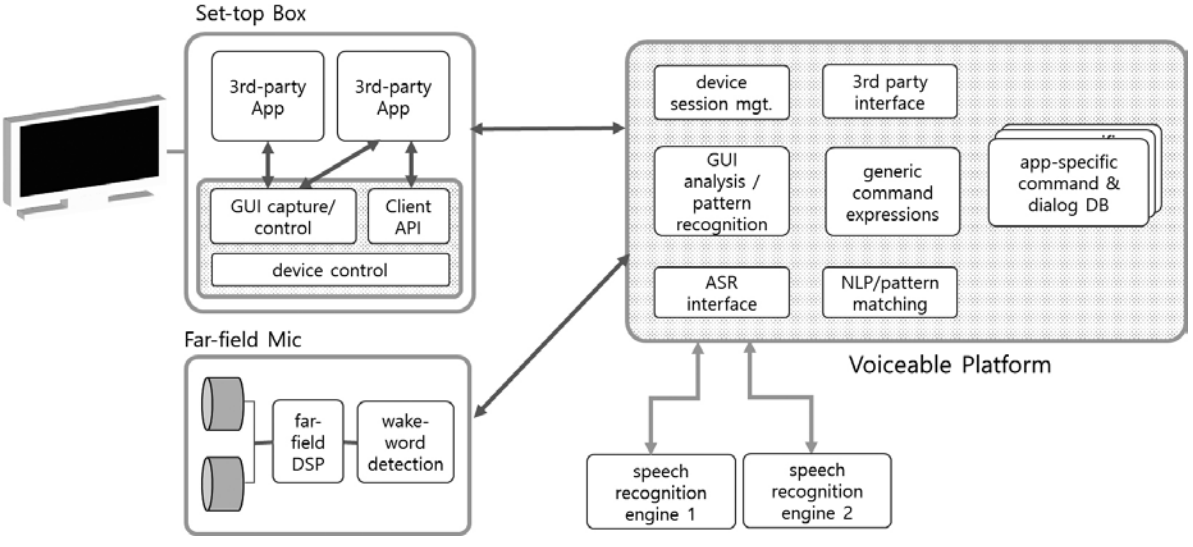

그림 8. Voiceable 아키텍처

전달받는 것이 어렵기 때문에 웹과 스마트폰의 뒤를 이어 다양한 서비스로의 관문(포털)을 제공하는 역할을 하기 위해서는 여러 입출력 장치를 상황과 목적에 따라 자유롭게 조합하여 사용할 필요가 있다. 마우스나 터치스크린이라는 물리적인 장치가 만들어졌다고 해서 다양한 GUI 기반의 서비스가 바로 가능해지는 것이 아닌 것처럼, 음성이나 제스처, 시선이나 포즈 인식의 요소 기술과 여러 형태의 출력 디바이스들이 손쉽게 다양한 서비스에 활용될 수 있으려면 이러한 멀티모달 UI를 쉽게 구현할 수 있고 일관된 방식으로 사용자에게 제공될 수 있는 새로운 API와 미들웨어의 표준화가 필수적이다. 이러한 표준은 단시일 내에 완성되지 않는다. 데스크톱 GUI나 웹의 표준, 터치스크린을 위한 API와 프레임웍은 수십 년에 걸쳐 진화해 온 것이며, 음성과 제스처 인식 등 AI 기술에 의해 가능해지는 입력 방식들이 기존의 GUI 패러다임에 수용되고 스피커나 IoT 장치로부터 서비스 로봇에 이르는 새로운 디바이스들을 지원하려면 오랜 시행착오와 새로운 개념이 필요할 것이다. 알티캐스트에서는 안드로이드와 웹기반 애플리케이션에 음성 입력을 손쉽게 추가할 수 있도록 해주는 Voiceable이라는 미들웨어를 개발 중이며 TV 셋톱박스에 우선 적용을 추진 중이다 [그림 8 참조].

마무리하며

최근 딥러닝 기술의 급속한 발전으로 AI에 대한 높은 관심과 기대가 촉발되었지만 사실 AI 스피커와 같이 상용화된 제품에 주로 사용되고 있는 음성인식, 자연어처리와 같은 기술은 기본적으로는 오래전부터 존재하던 것들이다. 하지만 이러한 기술의 성능이 향상되고 많은 리소스와 노력이 투입될 수 있는 여건이 조성되면 어느 순간 피처폰이 스마트폰으로 바뀌는 것과 같은 상태전이가 일어나게 되며 지금 여러 업체들이 AI 기술과 제품에 많은 노력을 들이는 것은 AI 기술의 수준에 대한 몰이해로 비현실적인 기대를 갖고 있다기보다는 이러한 생태계의 큰 변화가 다시 한번 일어날 수 있는 시점이 가까워지고 있다고 느끼기 때문일 것이다. AI 기술에 대한 막연한 기대나 두려움보다는 그 기술의 특성과 최신 동향을 이해하고 기존 사업 영역과 제품에 어떻게 활용될 수 있을지 가능성을 모색하는 현실적인 접근이 필요한 시점이다.