Chapter 01

딥러닝 기반 행동 인식을 위한 비디오 처리기술 연구 동향

송순용_한국전자통신연구원 선임연구원

행동 인식 기술은 비디오 클립에 녹화된 대상 객체가 어떠한 범주의 행위를 취하는지 어디로 이동하는지 파악하는 기술이다. 딥러닝 기술을 통해 행동 인식이 가능한데, 여러 벤치마크를 통해 범주를 분류하는 측면에서 60~80% 대의 정확도를 보이고 있음을 확인할 수 있다. 본 고에서는 행동 인식을 위한 딥러닝 기술의 개발 동향을 살펴보고자 한다. 수년간 연구된 결과들은 CNN 기반과 Transformer 기반으로 구분할 수 있는데, 개념도를 제시하거나 주요 특징들을 요약하는 형식으로 연구 동향을 분석 · 정리하였다.

I. 서론

현대의 인공지능과 컴퓨터 비전 기술은 4차 산업혁명 시대를 이끌어 가는 대표적인 원천 기술 중 하나로 로봇, 헬스케어 등과 같은 실생활에서 쉽게 접할 수 있는 차세대 기술 분야들과의 융합을 주도하고 있다. 여기서 인공지능은 순전파 및 역전파를 통해 주어진 데이터의 분포를 학습하여 어떠한 태스크의 결과를 출력한다. Convolutional Neural Network(CNN), Vision Transformer와 같은 아키텍처가 개발되면서 이미지, 비디오, 사운드와 같이 우리 현실과 가까운 정보를 데이터화 하여 인공지능을 적용할 수 있게 되었다. 컴퓨터 비전의 근간이 되는 이미지 처리 기술은 CNN 아키텍처를 통해 다양한 태스크를 처리하는데, 분류 및 객체인식 태스크의 경우 여러 연구자들의 노력을 통해 성능이 향상되고 있다.

비디오 데이터는 시간 순서로 나열된 이미지 데이터의 집합으로 구성되어 있다. 따라서 비디오 처리 기술은 기술적으로 이미지 처리 기술과 매우 유사하게 데이터를 처리한다. 그러

* 본 내용은 송순용 선임연구원(☎ 042-860-1737, soony@etri.re.kr)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

***본 연구 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [22ZR1100, 자율적으로 연결 · 제어 · 진화하는 초연결 지능화 기술 연구]

나 공간적인 정보만 담고 있는 이미지 데이터와 달리 비디오 데이터에는 시간의 흐름이 반영되어 있어서 시간-공간적인 정보가 포함되어 있다. 즉, 비디오 데이터를 통해 객체 혹은 상태의 동적인 특성 정보를 얻어내는 행동 인식과 같은 태스크를 수행할 수 있다는 장점이 있다. 행동 인식은 비디오 클립에 녹화된 대상 객체가 어떠한 범주의 행동을 취하는지, 혹은 어디로 이동하는지 등과 같이 행동 또는 이동 동선을 분류 및 예측하는 태스크를 다루게 된다.

비디오 기반의 행동 인식 기술은 크게 특징점의 직접적인 사용 유무에 따라 기술을 분류할 수 있다. 사람과 같이 특징점이 명확하게 드러나는 경우 특징점을 직접적으로 인식하고 특징점 간의 관계성을 통해 자세를 파악한 뒤 시간 특징을 인식하도록 신경망을 구축할 수 있다. 행동 인식을 위한 특징점으로 스켈레톤을 예로 들 수 있다. OpenPose[1]에서 스켈레톤 인식을 위해 사람의 관절과 뼈대를 차례대로 인식한다. 관절의 관계성을 통해 연결 관계를 복원하는 방식으로 뼈대를 인식한다. 이를 테면 팔목과 팔꿈치, 무릎과 발목은 서로 연결되어 있으므로, 신경망이 전술한 관절을 인식하였다면 자연스럽게 뼈대를 추정할 수 있다. 현재는 손가락 관절을 식별할 수 있는 수준에 도달해 있으며, 신체를 골고루 사용하는 수화(sign language) 인식이 가능할 정도의 성능 결과를 보여준다. 사람의 특징점을 제공하는 데이터셋이 다수 존재하여 이러한 데이터를 통해 실세계에서 사용이 가능하다. 일단, 특징점을 인식하면 그래프 혹은 고전적인 머신러닝 등 다양한 기법을 통해 행동 인식에 적용할 수 있다. 그러나 사람 이외의 객체에 대해서는 특징점 연구가 거의 없는 편이어서, 아직까지는 특징점 기반의 방식은 인식 대상이 사람이어야 행동을 인식할 수 있다는 명확한 한계가 있다.

반면, 특징점을 사용하지 않는 행동 인식 기술은 비디오를 구성하는 이미지 혹은 옵티컬 플로 같이 주어진 데이터의 범위 안에서 얻을 수 있는 정보만을 사용한다. 시간적인 정보를 인공신경망에 반영하기 위해 여러 가지 접근법을 사용하는데, 본 고의 II절에서 CNN 기반 접근법의 연구 동향을 설명하고, III절에서 Vision transformer 기반 접근법의 연구 동향을 설명한다. 마지막으로 IV절에서 본 고의 결론을 제시한다.

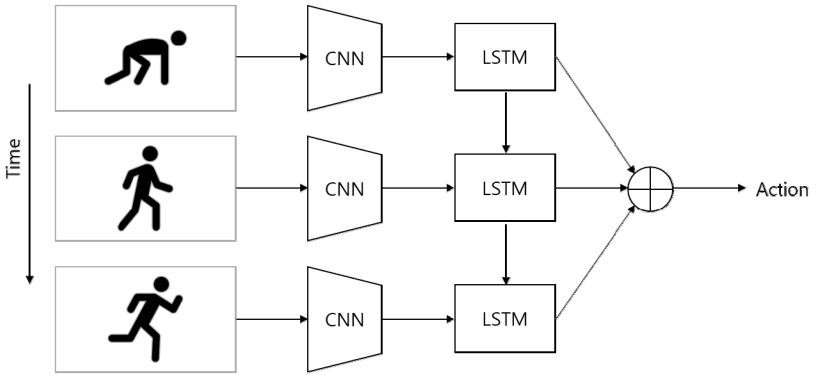
II. CNN 기반 행동 인식 기술

CNN 기반의 접근 방식은 두 가지의 단계로 행동을 결정한다. 첫 번째 단계에서 공간적인 특징을 추출한다. 이 때 CNN 블록을 사용하게 된다. 비디오 데이터는 이미지 데이터를 시간

적으로 나열한 형태이기 때문에 CNN 블록은 비디오 클립을 구성하는 여러 이미지에 대한 압축된 특징을 나열하게 된다. 두 번째 단계에서 시변 특징을 추출하기 위해 압축된 특징을 결합하는데, 흔히 알려진 방식으로는 자연어 처리에서 순열 데이터 처리를 위해 사용하는 RNN(Recurrent Neural Network) 구조 중 하나인 LSTM(Long Short-Term Memory) 블록을 사용하는 방식과 3D CNN을 사용하는 방식이 있다. 또한, 데이터의 입력 방식에 따라 이미지만을 사용하는 단일 입력 처리 방식과 이미지와 옵티컬 플로를 동시에 사용하는 멀티모달 처리 방식으로 구분 지을 수 있다. 본 고에서는 [6]에 조사된 자료를 토대로 CNN 기반의 행동 인식 기술을 다음과 같이 요약 정리한다.

1. CNN + LSTM

본 기술은 “Long-term Recurrent Convolutional Networks for Visual Recognition and Description”이라는 제목으로 CVPR 2015년에 발표된 연구 결과이다[2]. 인공신경망의 아키텍처는 [그림 1]과 같이 설계되어 있다. 비디오 클립을 구성하는 이미지들은 개별적으로 CNN 블록을 통해 단일 이미지에 대한 latent vector로 변환된다. Latent Vector의 시퀀스는 LSTM 블록을 통과하면서 시퀀스의 특징을 학습하게 된다. LSTM의 출력은 비디오



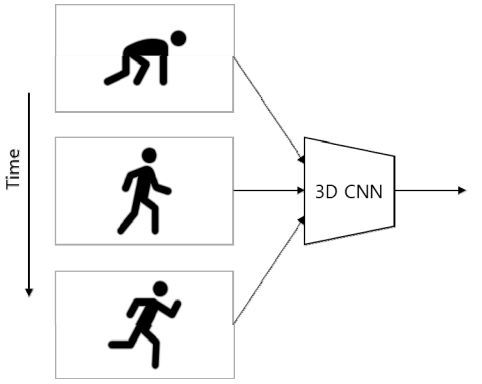
〈자료〉 Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell, “Long-Term Recurrent Convolutional Networks for Visual Recognition and Description,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015, pp.2625-2634. Joao Carreira, Andrew Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp.6299-6308. 재구성

[그림 1] CNN+LSTM 개념도

클립의 공간적/시간적 특징을 함축하게 되며, 함축된 정보를 통해 행동 범주를 예측한다. [6]에 정리된 시험 결과에 따르면 비교적 예측 정확도가 좋은 것으로 판단되지만, 모든 이미지에 대해 CNN 블록을 통과하고 이를 LSTM 블록에 다시 통과시키는 절차로 인해 인공신경망의 계산량 복잡도가 높아지는 단점이 있다. Latent Vector에 행동 구분이 가능한 정보가 포함되어 있어야 하기 때문에 본 모델의 성능은 CNN 블록의 성능에 비례하게 된다.

2. 3D CNN(C3D)

본 기술은 “Learning Spatiotemporal Features with 3D Convolutional Networks”라는 제목으로 ICCV2015 학회에서 발표된 연구 결과이다[3]. [그림 2]와 같이 이미지 처리를 위한 2D CNN에서 차원을 하나 늘린 구조로, 늘어난 차원을 이용하여 시변 데이터를 처리한다. [6]에 정리된 시험 결과에서는 가장 성능이 낮은 것으로 나타났지만, CNN 블록만으로 행동 인식의 수행이 가능하며 데이터를 빠르게 처리하는 장점이 있다.

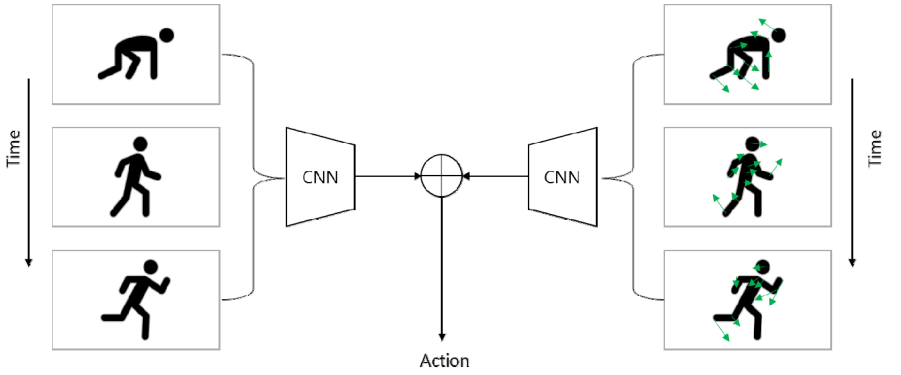


〈자료〉 Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, “Learning Spatiotemporal Features With 3D Convolutional Networks,” Proceedings of the IEEE International Conference on Computer Vision(ICCV), 2015, pp.4489-4497.
Joao Carreira, Andrew Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp.6299-6308. 재구성

[그림 1] CNN+LSTM 개념도

3. Two-Stream CNN

본 기술은 “Two-Stream Convolutional Networks for Action Recognition in Videos”



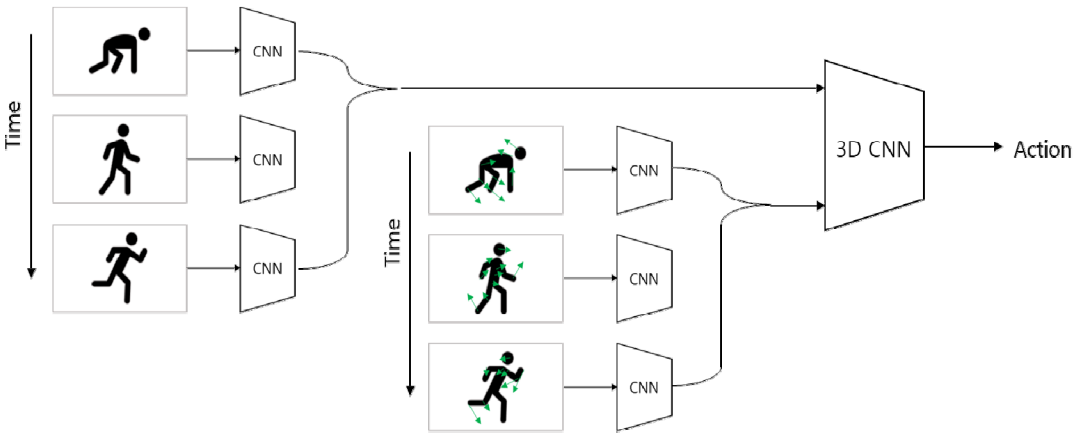
〈자료〉 Karen Simonyan, Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," Advances in neural information processing systems, 2014, 27.
Joao Carreira, Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp.6299-6308. 재구성

[그림 3] Two-Stream CNN 개념도

라는 제목으로 NeurIPS2014 학회에서 발표된 연구 결과이다([그림 3] 참조)[4]. 전술한 두 기술은 RGB 이미지만을 다루고 있었던 반면, 본 기술은 옵티컬 플로라는 새로운 데이터를 추가적으로 사용하였다. RGB 이미지는 Spatial stream CNN 블록을 통해 처리하고 옵티컬 플로는 3개의 그레이 이미지를 모아 3채널 이미지로 변한한 뒤 Temporal stream CNN 블록으로 처리한다. 처리된 결과를 모아 예측 결과를 출력한다. 논문에서는 Spatial stream CNN과 Temporal stream CNN의 아키텍처가 서로 동일하게 나타나 있다. 시공간 정보를 개별적으로 처리한다는 점에서 CNN+LSTM과 비슷하지만, 벤치마크 상 성능은 본 기술이 나은 것으로 알려져 있다.

4. 3D-Fused Two-Stream

본 기술은 "Convolutional Two-Stream Network Fusion for Video Action Recognition"이라는 제목으로 CVPR2016 학회에서 발표된 연구 결과이다[5]. [그림 4]와 같이 Two-stream의 마지막 단계에서 시공간 정보를 퓨전하는 것은 [그림 3]과 유사하지만 [그림 3]의 구조에서는 단순 결합하는 형태를 보인다. 본 기술은 3D CNN과 3D 풀링(pooling)을 통해 시공간 정보의 결합(joint) 분포로 데이터를 처리할 수 있어서 더 나은 예측결과를 기대할 수 있다. 실제로 전술한 기술들에 비해 나은 성능을 보인다.

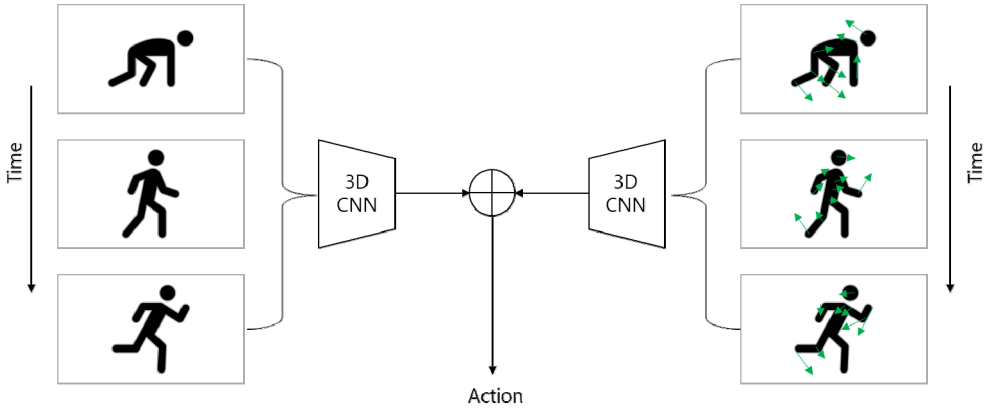


〈자료〉 Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.1933-1941.
Joao Carreira, Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp.6299-6308. 재구성

[그림 4] Two-Stream CNN 개념도

5. Two-Stream 3D CNN(I3D)

본 기술은 "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset"



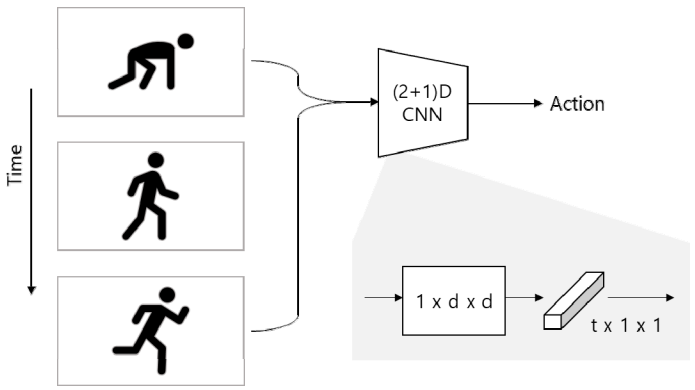
〈자료〉 Joao Carreira, Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp.6299-6308. 재구성

[그림 5] Two-Stream 3D CNN 개념도

이라는 제목으로 CVPR2017 학회에서 발표된 연구 결과이다[6]. [그림 5]와 같이 [그림 4]의 Two-stream 구조의 CNN 블록이 3D CNN 블록으로 대체된 아키텍처로 구성된다. 즉, spatial stream과 temporal stream 모두 3D CNN 블록을 통해 처리되고 처리된 결과를 마지막에 결합하여 예측결과를 출력한다. 이러한 구조를 통해 RGB와 옵티컬 플로의 변화를 모두 감지할 수 있어 예측결과 향상에 도움을 준다.

6. R2plus1D

본 기술은 “A Closer Look at Spatiotemporal Convolutions for Action Recognition” 이라는 제목으로 CVPR2018 학회에서 발표된 연구 결과이다[7]. [그림 6]과 같이 (2+1)D CNN 구조를 제안하였다. 3차원을 직접적으로 처리하는 3D CNN과 달리 2D CNN으로 공간 데이터를 처리하고 1D CNN으로 시간 데이터를 처리하는 방식으로 분해하여 시공간 데이터를 처리한다. 이러한 구조를 적층하여 (2+1)D CNN 블록을 제작한다. 2D CNN을 사용하기 때문에 이미지 처리 영역에서 사용하였던 모델을 사용할 수 있다는 장점이 있다.

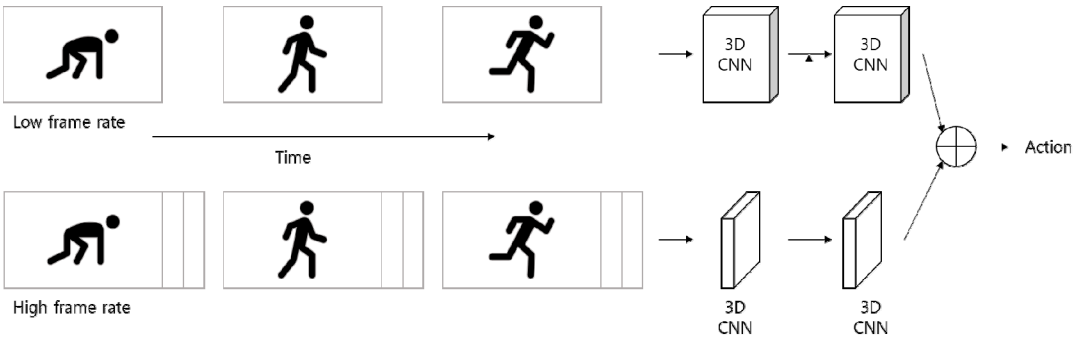


〈자료〉 Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, Manohar Paluri, “A Closer Look at Spatiotemporal Convolutions for Action Recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018, pp.6450-6459. 재구성

[그림 6] Two-Stream CNN 개념도

7. SlowFast

본 기술은 “Convolutional Two-Stream Network Fusion for Video Action Recognition”



〈자료〉 Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.1933-1941. 재구성

[그림 7] Two-Stream 3D CNN 개념도

이라는 제목으로 ICCV2019 학회에서 발표된 연구 결과이다[8]. [그림 7]과 같이 비디오 클립을 low frame rate와 high frame rate로 나누어 처리한다. Low frame rate에서는 공간적인 변화가 두드러지기 때문에 3D CNN의 temporal 축을 줄이고 채널 축을 깊게 설정하여 spatial stream을 처리하도록 유도한다. 반면, high frame rate에서는 시간적인 정보를 많이 포함하기 때문에 3D CNN의 temporal 축을 늘리고 채널 축을 얕게 설정하여 temporal stream을 처리하도록 유도한다. 두 개의 stream 간 dimension이 일치하는 구간에서 파라미터를 공유하여 시공간 정보를 결합한다.

8. X3D

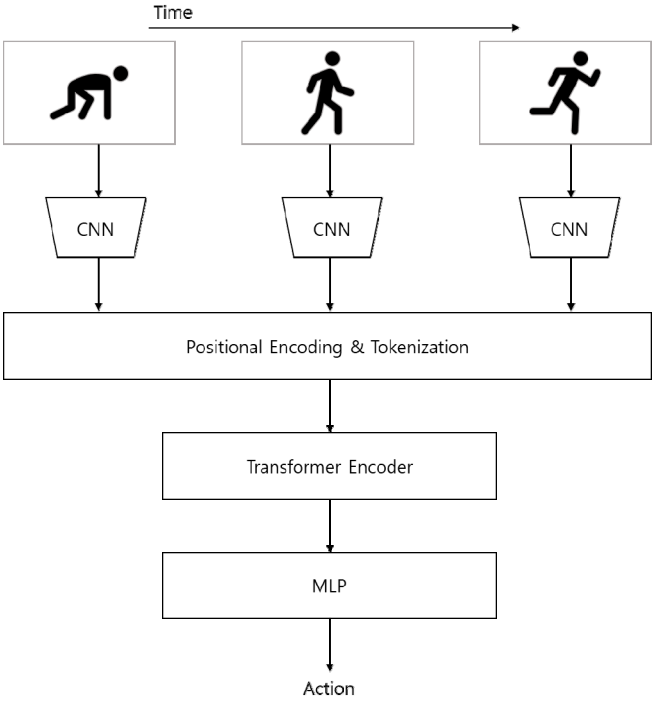
본 기술은 "X3D: Expanding Architectures for Efficient Video Recognition"이라는 제목으로 CVPR2020 학회에서 발표된 연구 결과이다[9]. 2D CNN을 기반으로 Temporal Duration, Frame rate, Spatial resolution, Width, Bottleneck width, Depth로 구성된 6개의 차원으로 확장하는 방식을 취한다. Forward expansion 과정으로 모델을 확장하고, backward contraction 과정에서 설계자가 정해 놓은 자원 사용량에 도달하였는지 확인한다. 본 고는 낮은 복잡도의 연산량을 갖는 모델을 찾는데 초점을 맞추었다.

III. Video Transformer 기반의 행동 인식 기술

Video Transformer는 CNN 기반의 기술과 유사한 방식으로 기술의 발전이 이루어졌다. CNN 기반의 기술은 공간적인 특성을 얻은 뒤 시간적인 특성을 얻는 방법과 시공간 특성을 동시에 얻는 방법으로 구분 지을 수 있는데, video transformer도 다음과 같은 흐름으로 연구가 진행되었다.

1. Video Transformer Network(VTN)

본 기술은 “Video Transformer Network”라는 제목으로 ICCV2021 학회 워크샵에서 발표된 연구 결과이다[10]. [그림 8]과 같은 아키텍처를 갖는데, CNN+LSTM과 같이 주요 특징을 2D CNN을 통해 얻고 latent vector의 sequence를 토큰화 한다. 각 토큰에 순서



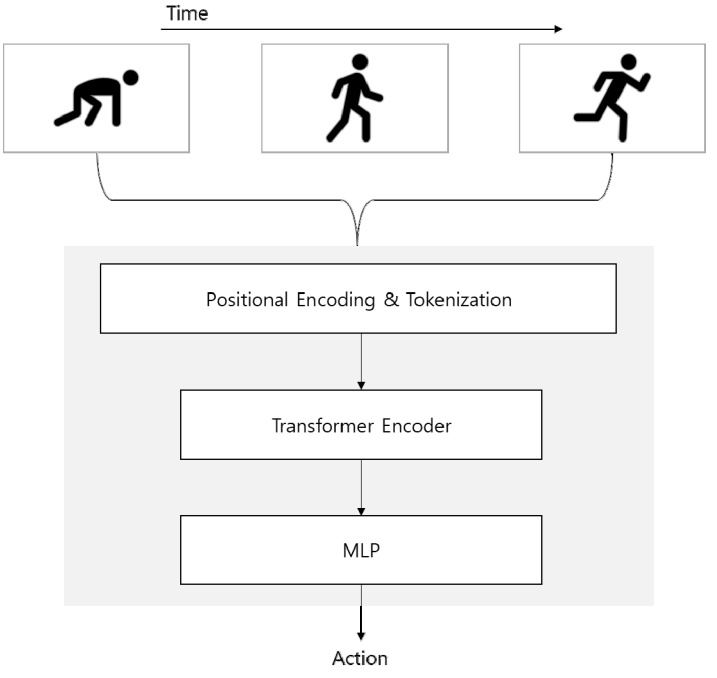
〈자료〉 Daniel Neimark, Omri Bar, Maya Zohar, Dotan Asselmann, “Video Transformer Network,” Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV) Workshops, 2021, pp.3163-3172. 재구성

[그림 8] Video Transformer Network 개념도

정보를 부가하기 위해 positional encoding한 뒤 transformer encoder에 입력한다. 마지막으로 MLP head를 통해 예측값을 얻어낸다.

2. ViViT

본 기술은 “ViViT: A Video Vision Transformer”라는 제목으로 ICCV2021 학회에서 발표된 연구 결과이다[11]. [그림 9]에서 회색 음영으로 표현된 부분은 ViT를 적용한 부분으로 비디오를 구성하는 모든 이미지를 패치로 나누고, 모든 패치에 대해 positional encoding 및 토큰화 한다. [그림 8]과 유사한 구조이지만, 트랜스포머 구조만 사용한다는 차이점이 있다.



〈자료〉 Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, Cordelia Schmid, “ViViT: A Video Vision Transformer,” Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), 2021, pp.6836-6846. 재구성

[그림 9] ViViT 개념도

3. TimeSFormer

본 기술은 “Is Space-Time Attention All You Need for Video Understanding?”이라는 제목으로 ICML2021 학회에서 발표된 연구 결과이다[12]. Transformer encoder에서 5가지 유형의 self-attention 블록을 제안하였고 각각을 테스트 하였다. Visualization 및 벤치마크를 통해 divided space-time attention을 사용하여 성능 개선의 효과를 입증하였다. 나머지 블록은 ViViT과 비슷한 방법으로 구성된다.

4. MViT

본 기술은 현재까지 두 번째 개선 버전이 발표되었다. 첫 번째는 “Multiscale Vision Transformers”라는 제목으로 ICCV2021 학회에서 발표되었다[13]. 첫 번째 버전의 MViT는 기존 트랜스포머의 multi-head attention 블록 대신 multi-head pooling attention 블록을 사용하여 다양한 크기의 시공간 데이터를 처리한다. 이러한 구조를 통해 트랜스포머를 CNN처럼 입력단 근처에서 low-level feature를 얻고 출력단 근처에서 high-level feature를 얻을 수 있게 한다. 본 논문의 방법은 다른 트랜스포머와 달리 전이학습을 사용하지 않더라도 충분히 좋은 성능을 보여준다는데 장점이 있다.

두 번째 버전의 MViT는 “MViTv2: Improved Multiscale Vision Transformers for Classification and Detection”이라는 제목으로 CVPR2022 학회에서 발표되었다[14]. 두 번째 버전에서는 이미지 분류 및 객체 인식 태스크 수행이 가능하도록 트랜스포머의 구조를 변경하였다. 성능 개선을 위해 decomposed relative position embedding과 residual pooling connection을 추가하였고, 이러한 개선사항을 통해 우수한 성능을 달성하였다.

5. Video Swin Transformer

본 기술은 “Video Swin Transformer”라는 제목으로 CVPR2022 학회에서 발표된 연구 결과이다[15]. 본 기술은 Swin Transformer를 기반으로 제작된 것으로 3D Shifted window를 통해 다양한 크기의 픽셀 영역 및 시간적 특징을 추출한다. 성능 관점에서 MViTv2와 유사하게 우수한 성능을 달성하였다.

Ⅳ. 결론

개발자의 관점에서 CNN은 트랜스포머에 비해 오랜 시간 동안 알려져 있으며, 대부분의 딥러닝 SDK를 통해 원하는 기능을 편리하게 구현할 수 있다. 하지만 트랜스포머의 self-attention은 뛰어난 특징 추출 성능을 가지고 있고 이러한 장점에 의해 최근 활발하게 연구되고 있다. 2021년부터 행동 인식과 관련하여 중요한 지표가 될 만한 트랜스포머 관련 기술들이 개발되었으며, 2022년에는 행동 인식에 적합한 구조의 연구 결과가 발표되었다. 현재까지는 성능과 연산 복잡도는 trade-off 관계인 것으로 나타나지만, 차츰 복잡도가 개선되어 실생활에 적용 가능할 정도가 될 것으로 전망한다. 행동 인식은 향후 다양한 응용이 가능하며 이에 따른 가치 창출이 가능할 것으로 예상되는 만큼, 발전 추이와 연구 동향에 대해 지속적으로 관심을 기울여야 할 것으로 판단한다.

참고문헌

[1] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp.7291-7299.

[2] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015, pp.2625-2634.

[3] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri, "Learning Spatiotemporal Features With 3D Convolutional Networks," Proceedings of the IEEE International Conference on Computer Vision(ICCV), 2015, pp.4489-4497.

[4] Karen Simonyan, Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," Advances in neural information processing systems, 2014, 27.

[5] Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016, pp.1933-1941.

[6] Joao Carreira, Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp.6299-6308.

[7] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, Manohar Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," Proceedings of the IEEE Conference

on Computer Vision and Pattern Recognition(CVPR), 2018, pp.6450-6459.
[8] Christoph Feichtenhofer, Axel Pinz, Andrew Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016, pp.1933-1941.
[9] Christoph Feichtenhofer, "X3D: Expanding Architectures for Efficient Video Recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp.203-213.
[10] Daniel Neimark, Omri Bar, Maya Zohar, Dotan Asselmann, "Video Transformer Network," Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV) Workshops, 2021, pp.3163-3172.
[11] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, Cordelia Schmid, "ViViT: A Video Vision Transformer," Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), 2021, pp.6836-6846.
[12] BERTASIUS, Gedas; WANG, Heng; TORRESANI, Lorenzo. "Is space-time attention all you need for video understanding?," In: ICML. 2021. p.4.
[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, Christoph Feichtenhofer, "Multiscale Vision Transformers," Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), 2021, pp.6824-6835.
[14] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, Christoph Feichtenhofer, "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2022, pp.4804-4814.
[15] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, Han Hu, "Video Swin Transformer," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2022, pp.3202-3211.

Chapter 02

식품산업의 푸드테크 빅데이터/AI 기술 동향

김준호_SK주식회사 매니저

I. 식품산업과 푸드테크

식품산업은 인류가 살아가는데 가장 근본적인 의식주 산업이며, 이 중에서도 우리 생활에 가장 밀접한 산업 분야이다. 식품산업은 축산물 가공, 음료, 유가공 등 다양한 업종을 아우르는 Mega Market으로([그림1] 참조)[1], 인류의 발전과 함께 항상 동시대의 신기술을 반영해 발전해 왔다. 인류의 역사와 함께 발전해 온 식품산업은 최근 푸드테크(Food Technology)라는 디지털 혁신 기술로 세간의 큰 관심을 받고 있다. 푸드테크 란 식품(food)과 기술(technology)의 합성어로, 식품산업과 관련 산업에 인공지능(AI), 사물인터넷(IoT), 정보통신기술(ICT) 등 4차 산업기술을 적용하여 이전보다 발전된 형태의 산업과 부가가치를 창출하는 기술을 의미하며 단순히 개량/생산의 발전에만 국한되지 않고, 식품의 생산에서부터 유통, 판매 등 전 과정을 포괄하여 적용 가능 한 개념이다[2]. 본 고에서는 국내외 식품산업에서 펼쳐지고 있는 빅데이터/AI 기술 관점에서 주요 기술 동향과 향후 발전 방향에 대해 살펴보겠다.

II. 푸드테크 시장 및 산업구조

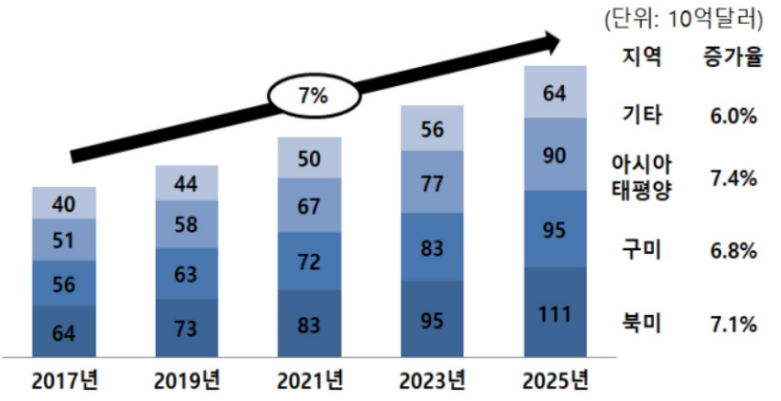
2016년 다보스 포럼에서 4차 산업혁명이 소개된 이후, 전 산업영역에서 다양한 디지털 기술과 혁신이 추진되었다. 특히, 통신, 유통, 콘텐츠, 금융 분야에서 AI와 빅데이터, 클라우드는 이제 시장에 보편화된 기술로 전파되었고 우리의 삶 속에서 디지털 기술을 밀접하게

* 본 내용은 김준호 매니저(☎ 02-6400-4714, maverick@sk.com)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

〈자료〉 메리츠증권, 2020년 하반기 전망시리즈 Meritz 음식료 "Q의 시대", 2020. 6.18. p.14.

[그림 1] 음식료 주요 섹터별 시장규모 맵

체감해 볼 수 있게 되었다. 이러한 전방위적인 디지털 기술 확산에도 불구하고 식품 분야는 상대적으로 디지털 혁신이 비교적 더디게 진행되는 산업 중 하나였다. 식품산업의 디지털 혁신 키워드인 '푸드테크'는 2019년을 기점으로 전 세계적으로 주목받게 되었는데, 식품제조에서부터, 유통, 배달(food delivery), 대체식품, 로봇 분야에서 디지털 기술 발전이 급격



〈자료〉 한국농촌경제연구원, 식품산업의 푸드테크 적용 실태와 과제, 2019. 10. p.24.

[그림 2] 푸드테크 시장규모 전망

하게 진행되고 있다. 글로벌 푸드테크 시장은 2017년부터 2025년까지 연평균 7% 성장할 것으로 전망되며, 2025년에는 시장규모가 3,600억 달러(한화 약 513조 원)에 이를 것으로 전망된다([그림 2] 참조)[3].

푸드테크 산업은 크게 물류유통(Logistics & Retail Service), 온디멘드 서비스(On demand Service), 정보 콘텐츠(Information & Content), 인프라 테크(Infra Tech) 영역으로 구분할 수 있다([그림 3] 참조)[4]. 국내에서는 마켓컬리, 오아시스마켓을 필두로 한 물류유통 분야와 배달의 민족, 요기요와 같은 음식배달 서비스를 중심으로 한 온디멘드 서비

〈자료〉 리테일매거진, 푸드테크 어디까지 왔을까, 2021. 3. 17.

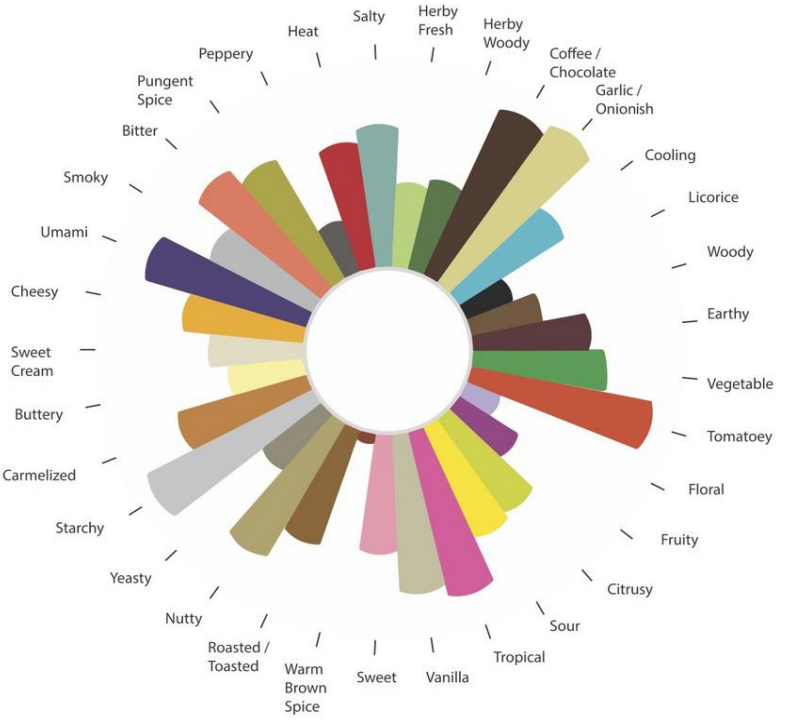
[그림 3] 국내 푸드테크 사업 구분

스 영역이 가장 디지털화가 활발히 진행되고 있다. 또한, 최근 대체식품(대체육), 로봇, 스마트팜 분야의 사업 경제성이 확보되면서 인프라 테크 영역도 투자 시장의 주목을 받고 있다. 향후 빅데이터/AI 기술 활용 관점에서는 그 활용도가 가장 큰 영역인 정보 콘텐츠 서비스 영역(information contents service)이 각광 받을 것으로 전망된다.

III. 푸드테크 디지털 기술 및 산업 동향

1. 해외 푸드테크 디지털 기술 적용 사례

해외에서는 일찍이 선진국을 중심으로 다양한 푸드테크 기술이 선보이고 있다. 이 중 빅데이터 기술을 적용한 대표적인 사례는 미국의 맥코믹(Mccormick)사의 디지털 플랫폼 사례



〈자료〉 Balimore Business Journal, McCormick spinoff specializing in 'flavor fingerprinting' nabs investment from SAP, 2016. 8. 26

[그림 4] 멕코믹(McComick)의 FlavorPrint 분류표

를 들 수 있다. 맥코믹은 세계적으로 유명한 향신료 회사로, 맥코믹의 CSO인 하메드 파라디는 조리법 패턴을 학습하여 새로운 요리메뉴를 제안한다는 IBM의 인공지능 '세프왓슨(Chef Watson)의 내용을 듣고 IBM과 5년 간의 파트너십 계약을 맺어 맥코믹의 머신러닝 기반의 맛(taste) 플랫폼을 개발하였다. 맥코믹은 오랜 기간 축적된 향신료 사업 데이터와 분석 역량을 기반으로 식품의 맛과 향에 대한 빅데이터 베이스를 구축하였으며, 이용자의 개인별 식습관과 맛의 기회를 접목시켜 레시피와 식품을 AI로 추천하는 디지털 서비스를 선보였다.

이 빅데이터 사업은 본래 B2C 사용자를 대상으로 한 서비스로 출발하였으나 요리정보를 필요로 하는 레스토랑, 식자재 유통회사와의 B2B로의 사업으로도 확대되어 2014년 비벤다(Vivanda)라는 사명으로 분사하여 독립적인 사업을 추진하고 있다. 비벤다는 플레이버 프린트(FlavorPrint)로 명명되는 데이터 맛에 대한 빅데이터 시스템을 구축하여 현재까지 다양한 빅데이터 사업을 전개하고 있다([그림 4] 참조)[5]. 맥코믹은 1980년대부터 축적해온 실험 데이터에 머신러닝을 도입하여 신제품 개발의 시간을 70% 이상 절감하였으며, 제품 구매율 또한 크게 증진하는 성과를 거두었다.

요리를 하는 공간인 주방에 IoT, AI, 빅데이터 기술을 적용한 스마트 키친(smart kitchen)도 빠르게 상업화가 진행되고 있다. 특히, 스마트 키친은 가전업체가 사업을 주도하고 있는데 삼성전자와 LG전자, 보슈의 홈커넥트 서비스가 그 대표적인 예이다.

삼성전자의 스마트 조리기구인 큐커는 식음료 업체들과의 제휴로 제품의 코드를 모바일로 인식하면 바로 조리기구에서 조리시간을 설정하는 기능을 선보여 2022년에 10만 대 이상의 판매고를 보이고 있으며, LG전자는 자사의 가전 IoT 플랫폼인 씽큐(ThinQ)를 통해 모바일로 주방가전을 제어할 수 있는 기술을 선보여 가전업계의 큰 호응을 얻고 있다. 해외에서는 보슈(Bosche)가 2017년 스마트 키친사업에 진출하여 스마트 주방가전 및 주방 데이터를 기반으로 한 스마트 키친사업을 선도하고 있다. 빅테크 기업인 아마존과 구글은 AI 스피커인 아마존 알렉사(Alexa), 구글 어시스턴트를 통해 음성 AI 기반의 레시피 콘텐츠 서비스를 선보이며 스마트 키친 분야 진출에 도전하고 있다. 아마존과 구글이 선보인 음성 AI는 AI 컨버세이션(conversation) API와 STT(Speech To Text)/TTS(Text to Speech) API를 활용한 기술이며, 주방에서 레시피를 음성으로 듣거나, 음성으로 주방가전을 통제할 수 있는 기능을 제공함으로 요리할 때 불편한 두 손을 자유롭게 하는 사용자 편익을 제공하고 있다.

2. 국내 푸드테크 디지털 기술 적용 사례

가. 이커머스 분야

국내 푸드테크 분야에서 가장 주목받고 있는 섹터는 신선식품을 온라인으로 판매하는 이커머스 분야이다. 푸드 커머스 기업의 대표주자인 마켓컬리는 자체 개발한 데이터 수집 분석 시스템(데이터 물어다 주는 멍멍이)을 개발하여 소비자의 주문을 예측하고 있다. 마켓컬리는 자체 빅데이터 분석 시스템을 통해 신선식품 폐기율을 1% 미만으로 유지해 왔다고 밝힌 바 있다[6]. 일반적으로 대형마트 식품 폐기율은 3% 내외, 슈퍼마켓은 7~8% 수준을 감안한다면 빅데이터/AI 기술을 이용하여 유통 물류 분야에서 괄목할 만한 성과를 이루어낸 것이다. 마켓컬리는 본 빅데이터 분석 시스템을 통해 실시간 데이터 모니터링 및 빅데이터 분석을 하고 있으며 시간대별, 지역별 주문 현황 및 향후 판매량까지 예측하고 있다. 또한, 재고정보를 실시간으로 추적하여 소비자에게 상품 프로모션을 제안하기도 하며, 수집된 데이터 분석 결과는 물류팀에 전달되어 분류, 포장, 배송 등 전 영역에 대응하여 유통의 핵심인 물류 시스템 전반의 효율화를 이루어내고 있다.

나. 정보 콘텐츠 분야

쿠켓은 요리정보와 밀키트 생산 판매를 하는 스타트업으로 커머스 사이트 및 SNS 채널 이용자의 웹(web)/앱(app) 로그(log)를 추적 분석하여 신제품 개발과 마케팅 전략에 활용하고 있는 푸드테크 기업이다. 웹과 앱에서는 사용자의 다양한 행동 데이터의 발자국(log)을 수집·분석할 수 있으며([표 1] 참조)[7], 온라인 로그 데이터를 수집 분석하는 것은 소비자의

[표 1] 웹(Web), 앱(App) 로그 추적 가능 데이터 유형

구분	추적 및 분석 가능 데이터
누가	방문자 IP주소, 로그인/비로그인 여부, 회원ID
언제	접속시간, 전체 체류시간, 페이지별 체류시간, 이탈시간 등
어디서	사이트 방문경로(유입경로), 이탈경로(유출경로), 클릭 및 이동 페이지 등
무엇을	검색어, 조회 페이지, 조회상품, 관심상품, 결제상품 등
어떻게	재방문 횟수, 장바구니 담기여부, 결제여부, 결제정보, 반품정보, 추천/좋아요 등 댓글 여부
왜	유입 키워드 및 유입사이트, 검색어 등

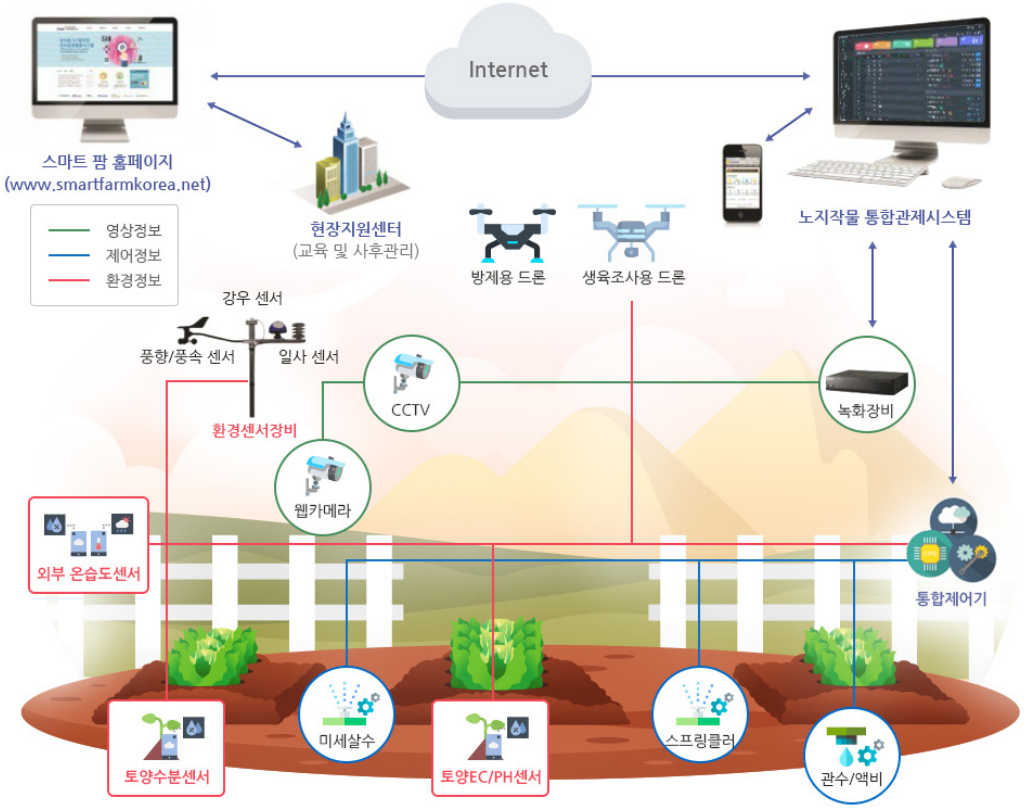
〈자료〉 윤미정, "빅데이터는 어떻게 마케팅의 무기가 되는가", 클라우드나인, 2020, p.117.

잠재된 의식과 행동을 객관적으로 파악할 수 있기에 소비자의 표출되지 않은 욕구(needs)와 불편함(pain point)을 소비자에게 직접 물어보는 서베이보다 더 객관적이고 정확히 파악할 수 있다는 이점이 있다.

쿠켓은 이러한 웹/앱 기반의 데이터 로그를 수집·분석하여 새로운 밀키트 메뉴를 개발하고 있으며, 이는 타 푸드테크 기업과는 차별화된 디지털 기술의 활용 사례라 볼 수 있다.

다. 스마트팜

전세계적인 기후 변화와 식량 위기로 스마트팜 분야의 디지털 혁신도 본격적으로 추진되고 있다. 스마트팜 분야는 농업과 직결된 테크 영역으로, 정부 차원의 연구 개발이 가장 적극적으로 추진되는 분야이기도 하다. 스마트팜 기술이 적용되는 분야는 크게 드론을 이용한



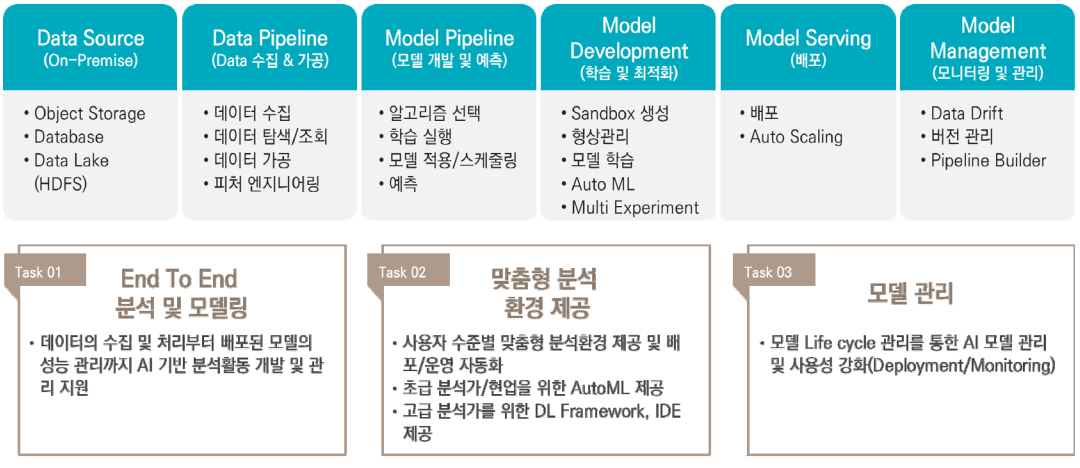
〈자료〉 스마트팜코리아 홈페이지, 스마트팜 안내, 노지분야 스마트 채소밭 구성도

[그림 5] 스마트 채소밭 구성도

농작물의 생육/병해충 등을 탐지하고 해당 데이터를 기반으로 한 농작물 관리를 하는 분야([그림 5] 참조)[8], 센서를 이용한 병해충 판별 AI 개발, IoT 수집 데이터와 기후 변화 데이터를 융합 분석한 AI 기반 수확량 예측 기술 분야가 대표적이다. 농림축산식품부는 2021년 11월 농림식품기술기획평가원과 함께 지능형 농장 연구개발(R&D) 사업 전 과정의 데이터를 모아 공유하는 “스마트팜 R&D 빅데이터 플랫폼”을 구축한 바 있다. 해당 플랫폼은 스마트팜 연구자들의 데이터 기반 R&D 결과를 민간에 공유하는 플랫폼으로, 정부에서 민간에게 스마트팜 데이터 및 기술을 전파하는 가장 핵심적인 플랫폼이라 할 수 있다. 이러한 국가차원에서 개발된 R&D 데이터를 기반으로 민간과 학계에서는 농생물 자원의 DB를 통합 분석한 신품종 개발 연구가 활발히 진행되고 있다. ICT 기술이 접목된 스마트팜의 기술 보급이 확대되면, 농업 분야의 노동 및 에너지 효율화에 기여하게 되며, 동시에 탄소 배출 절감 등 환경적인 성과도 기대할 수 있다. 스마트팜 디지털 기술의 확대는 국가적 차원의 농업 종사자의 삶의 질 개선, 농촌지역 발전에도 크게 기여할 것으로 전망된다.

라. 종합식품제조기업

2019년을 전후로 국내 종합식품제조기업도 AI/빅데이터 기술 개발을 본격적으로 추진하고 있다. 디지털 기술과 데이터를 확보한 푸드테크 기업의 시장 침투가 빠르게 진행되고 있기에 전통적 식품제조기업은 데이터 기반의 디지털 혁신을 추구하고자 데이터 수집/분석/



〈자료〉 SK주식회사, Accuinsight+ 기업용 홍보자료, 2021.

[그림 6] 빅데이터 통합분석 플랫폼 구조 예시

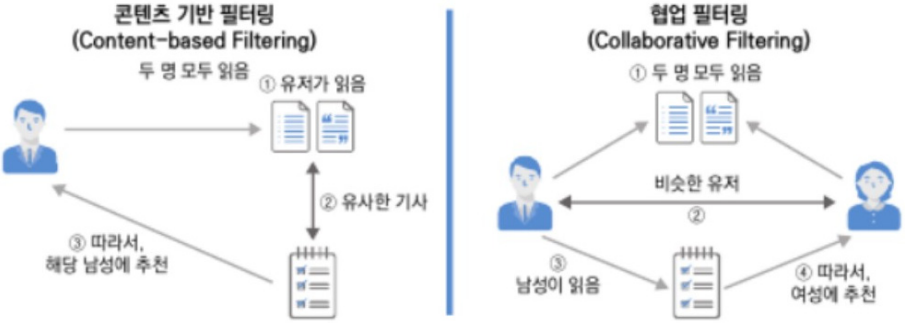
처리 및 모델링/배포/모델관리가 가능한 빅데이터 통합 분석 플랫폼을 앞다투어 개발 도입하고 있다([그림 6] 참조)[9].

종합식품제조기업 중 가장 앞서 투자를 하고 있는 곳은 CJ그룹으로, 2022년 AI센터를 출범하여 고객행동 양식 등 데이터 분석 인프라와 생산에서부터 유통, 물류까지 데이터와 AI를 기반으로 한 통합 인프라를 구축하고 있다. 풀무원도 2021년 통합DT플랫폼 구축을 진행하여 DCX(고객경험), SRM(공급자관리), SCM(공급망관리), DSF(생산품질), CDA(고객통합데이터분석) 등 전 비즈니스 영역의 디지털 전환을 모색하고 있다. 라면업계의 선두기업인 농심은 2022년 빅데이터/AI 기술을 기반으로 한 차세대정보시스템 구축에 나섰으며, B2C 고객접점 및 데이터 확보를 위해 디지털 전담조직을 확대하고 B2C 디지털 마케팅을 강화해 나가고 있다.

3. 빅데이터, AI의 핵심기술: 초개인화와 추천 기능

푸드테크의 B2C 사업에서 가장 많이 활용되고 있는 빅데이터/AI 기술은 초개인화(hyper personalization)와 AI의 제품/서비스 추천 기능(AI curation)이다. 각 기업이 적용하고 있는 추천 알고리즘은 대부분 공개되어 있지 않으나, 일반적으로 알려진 추천 알고리즘의 필터링 기법은 크게 콘텐츠 기반 필터링(content-based filtering)과 협업 필터링(collaborative filtering)으로 나뉜다([그림 7] 참조)[10].

첫 번째로 콘텐츠 기반 필터링은 콘텐츠 정보를 기반으로 다른 유사 콘텐츠를 추천하는 방식이다. 음식의 경우라면 음식의 맛, 영양소, 식재료, 열량 정보를 데이터화 하고 상품이라



〈자료〉 과학기술정보통신부, 일상속 과학 이야기, 유튜브와 넷플릭스의 추천 알고리즘, 2020. 3. 24.

[그림 7] 추천 알고리즘 필터링 기법

면 상품의 상세 정보를 데이터화하여 선택된 정보와 상품을 기준으로 유사한 정보와 상품을 추천하는 방식이다. 콘텐츠 기반 필터링은 콘텐츠 자체를 분석하는 방식이기 때문에 서비스 초기 사용자와 사용자의 행동 데이터 크기가 작더라도 추천 시스템을 작동시킬 수 있다는 장점이 있다. 반면, 콘텐츠 정보를 모두 함축하여 적용하기는 다소 어렵기 때문에 알고리즘 측면에서 이용자의 성향을 상세하게 분석하고 파악하기 어렵다는 한계를 지니기도 한다.

두 번째는 협업 필터링이다. 협업 필터링은 다수의 사용자로부터 획득한 행동, 기호 정보에 따라 사용자의 관심사를 자동으로 예측해 주는 기법이다. 예를 들면, 한 레시피 사이트에서 유사한 이용 패턴을 보인 그룹을 하나의 프로파일링 그룹으로 묶어 그와 유사한 패턴의 행동을 보이는 사람에게 동일한 레시피 정보나 밀키트 제품을 추천해 주는 방식이다. 협업 필터링의 경우 사용자의 다양한 행동 데이터를 기반으로 그 유사성을 찾아 추천하기에 정확도와 신뢰도 측면에서 효과적인 부분이 많으나, 서비스를 운영한지 얼마 안 되어 사용자의 데이터가 없는 경우 효과적인 적용이 어려우며, 데이터 양이 증가할 경우 처리 속도의 문제가 발생하는 점, 유사성만을 근거로 추천하기 때문에 편향된 정보를 제공하게 될 우려가 있다는 단점이 있다. 고도화된 추천 알고리즘으로 유명한 넷플릭스의 경우 상기한 필터링의 한계를 극복하고자 앙상블 체계(ensemble system)를 적용하고 있으며, 현재 웹이나 앱 서비스의 다수는 앙상블 체계를 적용하고 있다. 푸드테크 분야에서는 이러한 AI 추천 알고리즘을 적용한 요리정보 추천, 맛과 기호, 섭취 이력의 빅데이터 분석을 통한 메뉴 추천 등 다양한 시도가 진행되고 있다.

4. 푸드테크의 미래: 주방OS

AI/빅데이터, IoT, 클라우드 등 디지털 기술이 총체적으로 적용된 푸드테크의 미래로 주방OS의 개념이 주목받고 있다. 주방OS라는 용어는 2016년 미국에서 개최된 “스마트키친 서밋(SKS)”에서부터 사용되기 시작했으며, 주방가전의 IoT, 모바일 플랫폼 활용이 보편화되면서 본격적으로 거론되기 시작하였다[11]. 주방OS는 주방에서 요리할 때 발생하는 모든 데이터를 수집·분석하고, 주방가전과 디바이스를 통제하는 스마트 키친의 통합 플랫폼이라 할 수 있다. 요리라는 행위를 위해서는 식재료의 구매, 레시피 검색, 주방가전을 통한 조리라는 일련의 행위가 일어나는데, 현재는 각 단계의 정보를 제공하는 디바이스와 데이터가 분절되어 있다. 하지만 최근 빅데이터와 AI에 대한 관심과 기술이 발전함에 따라 주방에서 발생

하는 모든 데이터를 수집하고 분석하여 최적의 요리 환경을 만들고자 하는 시도가 진행되고 있으며 이를 주방OS가 지향하는 미래라 정의할 수 있다. 앞으로 누가 주방OS에 대한 기술을 먼저 개발하고, 통합된 데이터를 확보할 것인가에 따라 시장의 판도가 좌우될 것으로 예상된다.

IV. 맺음말

앞서 살펴본 바와 같이 빅데이터에 대한 수집, 적재, 분석 기술과 AI 기술이 식품산업의 패러다임을 바꾸게 될 핵심 동인이 될 것이다. 최근 국내에서도 푸드테크 관련 학계와 협회의 움직임이 본격화되고 국가적 차원의 푸드 마스터 데이터 R&D도 진행되고 있는 점을 미루어 볼 때, 향후 5년 이내 빅데이터와 AI를 필두로 한 푸드테크 기술은 가장 혁신적인 기술로 우리 삶에 다가올 것으로 전망된다. 또한, 식품은 인간의 건강과 의료 분야에 밀접한 연관성을 가지고 있기에 메디푸드(Medical Food)의 형태로도 연구되어 진화 발전해 나갈 것이다. 식품은 인간의 생존에 가장 근간이 되는 영역이므로, 푸드테크 기술은 우리의 삶과 인류의 미래를 책임질 산업으로 지속 발전해 나갈 것이다.

참고문헌

[1] 메리츠증권, 2020년 하반기 전망시리즈 Meritz 음식료 “Q의 시대”, 2020. 6. 18. p.14.
[2] 푸드테크와 식품산업, 식품산업통계정보시스템(atFIS), 식품시장 뉴스레터 2021. 8. 2주, p.2.
[3] 한국농촌경제연구원, 식품산업의 푸드테크 적용 실태와 과제, 2019. 10. p.24.
[4] 리테일매거진, 푸드테크 어디까지 왔을까, 2021. 3. 17.
[5] Balimore Business Journal, McCormick spinoff specializing in ‘flavor fingerprinting’ nabs investment from SAP, 2016. 8. 26.
[6] IT조선, 마켓컬리 폐기율 1% 미만 비결은 빅데이터, 2022. 3. 11.
[7] 윤미정, “빅데이터는 어떻게 마케팅의 무기가 되는가”, 클라우드나인, 2020. p.117.
[8] 스마트팜코리아 홈페이지, 스마트팜 안내, 노지분야 스마트 채소밭 구성도
[9] SK주식회사, Accuinsight+ 기업용 홍보자료, 2021.
[10] 과학기술정보통신부, 일상속 과학 이야기 “유튜브와 넷플릭스의 추천 알고리즘”, 2020. 3. 24.
[11] 다나카 히로이카 외 공저, Kmac, “푸드테크 혁명”, 2021. p.160.