온디바이스 미디어 제작을 위한 생성형 인공지능 경량화 기술 현황 및 전망

□ 박종희, 곽종훈 / 한국전자기술연구원

요 약

본 기고문은 디지털 미디어 제작 환경이 인공지능(AI)을 활용한 자동화로 전환되는 추세에서, 특히 생성형 AI 기술이 온디바이스 환경에서 실시간으로 미디어 콘텐츠를 생성하는 기술적 도전 과제를 다룬다. 온디바이스에서 AI 모델을 효과적으로 구현하기 위해 CNN 기반 경량화 기술인 Pruning(가지치기), Quantization(양자화), Knowledge Distillation(지식 증류)을 살펴보고, 해당 기술이 생성형 AI 기술에 접목되거나 확장되는 기술에 대해서 설명한다. 해당 기술을 기반으로 미디어 제작을 위한 온디바이스 생성형 AI 서비스와 전망을 통해 향후 기술 발전 방향에 대해 논한다.

Ⅰ. 서 론

디지털 미디어 환경이 급격히 변화하면서, 콘텐츠 제작 방식도 전통적인 수작업 중심에서 인공지능(AI)을 활용한 자동화 방식으로 전환되고 있다. 스마트폰, 태블릿, 스마트 TV, AR/VR 기기와 같은 디지털 디바이스에서 AI를 활용한 이미지 및 영상 처리, 음성 인식, 콘텐츠 추천 시스템 등이 보편화되고 있다. 특히, 생성형 인공지능(Generative AI) 기술이 발전하면서 이미지 생성, 영상 합성, 3D 콘텐츠 생성 등 다양한 형태의 미디어 콘텐츠가 자동으로 생성될 수 있는 시대가 열렸다. 기존에는 이러한 AI 모델이 대규모 데이터센터나 클라우드 서버에서 실행되었으나, 이러한 기술을 온디바이스(On-Device) 환경에서 실시간으로 구현하는 것은 상당한 도전 과제를 안고 있다. 최근에는 개인 스마트폰, 태블릿, AR/VR 기기, 엣지 디바이스에서도 실시간으로 실행될 수 있도록 온디바이스(On-Device) 환경에서의 AI 모델 경량화가 필수적인 기술로 떠오르고 있다.

기존의 딥러닝 모델은 높은 연산량과 방대한 메모리 사용량을 요구하기 때문에 주로 클라우드 서버에서 실행

되었다. 온디바이스 환경에서 AI 모델을 실행하려면, 한정된 연산 자원과 배터리 환경에서도 빠르게 동작할 수 있도록 딥러닝 네트워크의 크기를 줄이고 연산량을 최적화해야 한다. 특히, 영상 및 음성을 실시간으로 분석하는 미디어 애플리케이션은 고속 처리와 저지연이 요구되며, 사용자의 인터랙션에 즉각적으로 반응해야 한다. 하지만, 미디어 생성 작업은 대량의 데이터 연산이 요구되는 고비용 프로세스이므로, 이를 모바일이나 임베디드 디바이스에서 수행하려면 효과적인 경량화 기술이 필요하다.

딥러닝 네트워크 경량화 기술은 CNN 네트워크를 기반으로 Pruning(가지치기), Quantization(양자화), Knowledge Distillation(지식 증류) 등의 기술을 활용하여 AI 모델을 최적화하는 방식으로 발전되어 왔다. 이를 통해 클라우드 의존도를 줄이고 디바이스 자체에서 고품질 미디어 생성이 가능하도록 연구가 활발히 진행되고 있다. 또한, LLM과 LVM의 기술 발전에 따라 멀티모달을 활용한 생성형 AI의 미디어 제작사례가 증가하고 있고, 이에 맞춰 딥러닝 경량화 기술도 해당 네트워크를 기반으로 발전하고 있는 추세이다.

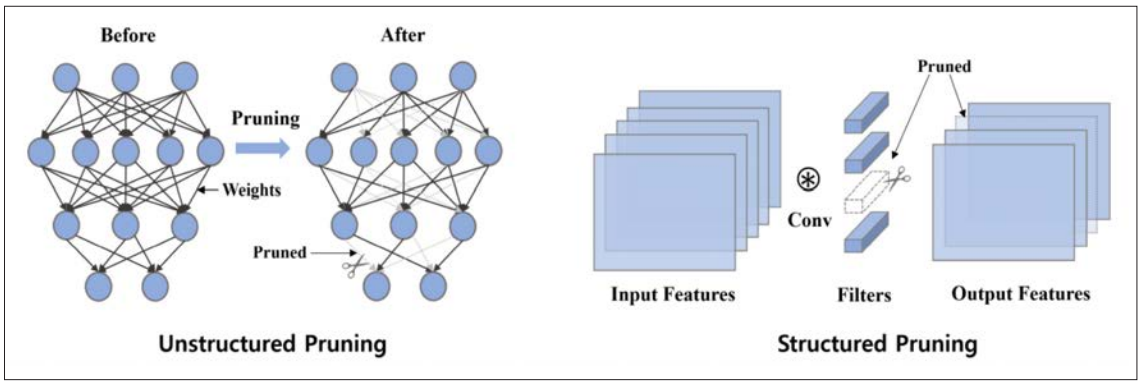
이러한 기술 발전은 창작자와 일반 사용자 모두에게 미디어 제작의 패러다임을 변화시키는 중요한 요소가 되고 있다. 예를 들어, 스마트폰에서 실시간으로 스타일 변환(Style Transfer), 음성 합성, AI 아바타 생성 등의 기능을 구현할 수 있으며, 이는 기존의 클라우드 기반 생성형 AI 서비스보다 빠르고 개인화된 경험을 제공한다. 또한, 개인정보 보호 관점에서도 데이터가 디바이스 내부에서 처리되므로 보안성이 높아진다는 장점이 있다.

Ⅱ. 딥러닝 네트워크 경량화 기술

본 장에서는 딥러닝 네트워크를 온디바이스에서 구동하기 위해 필요한 경량화 기술에 대해서 알아본다. 우선 CNN 네트워크 기반으로 발전된 경량화 기술을 알아보고 해당 기술이 생성형 AI를 위한 네트워크들에 적용된 형태와 추가적인 경량화 기술에 대해서 알아본다.

1. 온디바이스 AI를 위한 CNN 기반 딥러닝 네트워크 경량화 기술

딥러닝 모델은 높은 연산량과 메모리 요구량으로 인해 온디바이스 환경(스마트폰, IoT 기기, 임베디드 시스템 등)에서 직접 실행시키기 어렵다. 이를 해결하기 위해 Pruning, Quantization, Knowledge Distillation과 같은

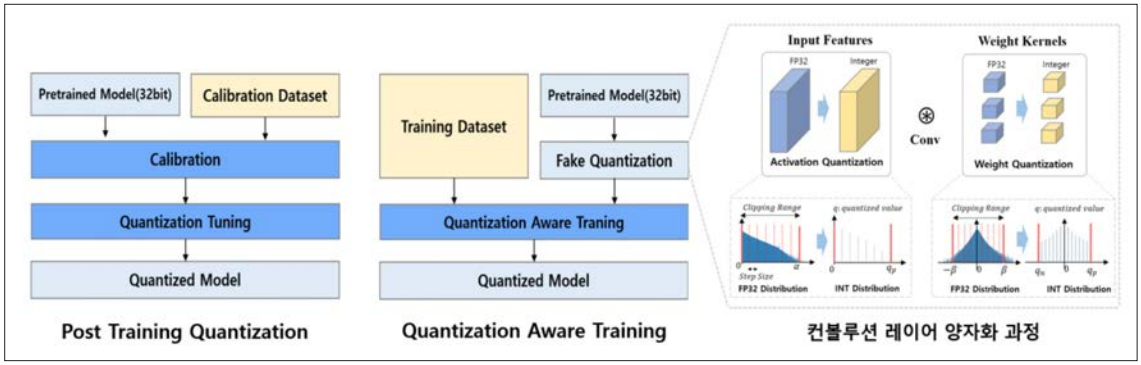


<그림 1> CNN 기반 네트워크 Pruning 기술

모델 경량화 기법이 사용된다. Pruning은 신경망 내에서 불필요한 가중치(weight)나 뉴런을 제거하여 모델의 크기를 줄이고 연산량을 감소시키는 기법이다. Unstructured Pruning은 개별 가중치의 중요도를 평가하여 낮은 값을 가지는 가중치를 제거하는 방식으로, 모델의 메모리 사용량을 크게 줄일 수 있다. 그러나 연산 속도 최적화에는 한계가 있으며, 하드웨어 가속기에서 비효율적으로 작동할 가능성이 크다. 반면, Structured Pruning은 채널, 필터, 레이어 단위로 불필요한 구조를 제거하여 연산량을 직접 감소시킬 수 있으며, 하드웨어 가속기와의 호환성이 높아 실질적인 속도 향상 효과가 크다. 그러나 성능 유지가 어려울 수 있으며, 모델의 재학습 과정이 필요할 가능성이 크다. Unstructured Pruning은 높은 압축률을 제공하지만 속도 최적화가 어렵고, Structured Pruning은 성능 최적화가 뛰어나지만 압축률이 상대적으로 낮다는 점에서 차이가 있다.

Quantization은 신경망의 가중치와 활성화 값을 더 낮은 비트(예: 32-bit → 8-bit)로 표현하여 모델의 크기를 줄이고 연산 속도를 높이는 기법이다. 일반적으로 Post-Training Quantization(PTQ)과 Quantization-Aware Training(QAT) 방식이 있다. Post-Training Quantization(PTQ)과 Quantization-Aware Training (QAT)은 모델의 연산 효율성을 높이기 위해 활용되는 대표적인 양자화 기법이지만, 적용 방식과 성능 측면에서 차이가 있다. PTQ는 사전 훈련된 모델을 양자화하여 연산을 최적화하는 방식으로, 별도의 추가 학습 없이 빠르게 적용할 수 있는 장점이 있다. 그러나 모델이 양자화로 인해 정확도가 저하될 가능성이 크며, 특히 생성형 AI 모델처럼 연산이 복잡한 경우 품질 저하가 두드러질 수 있다. 반면, QAT는 모델이 훈련되는 동안 양자화를 고려하여 학습하는 방식으로, 양자화된 상태에서도 높은 정확도를 유지할 수 있다. 다만, PTQ보다 훈련 비용이 높으며 추가적인 연산량이 필요하다는 단점이 있다. PTQ는 적용이 용이하지만 모델의 성능 저하 가능성이 있으며, QAT는 정확도를 유지할 수 있지만 계산 비용이 증가한다는 점에서 서로 다른 장단점을 가진다.

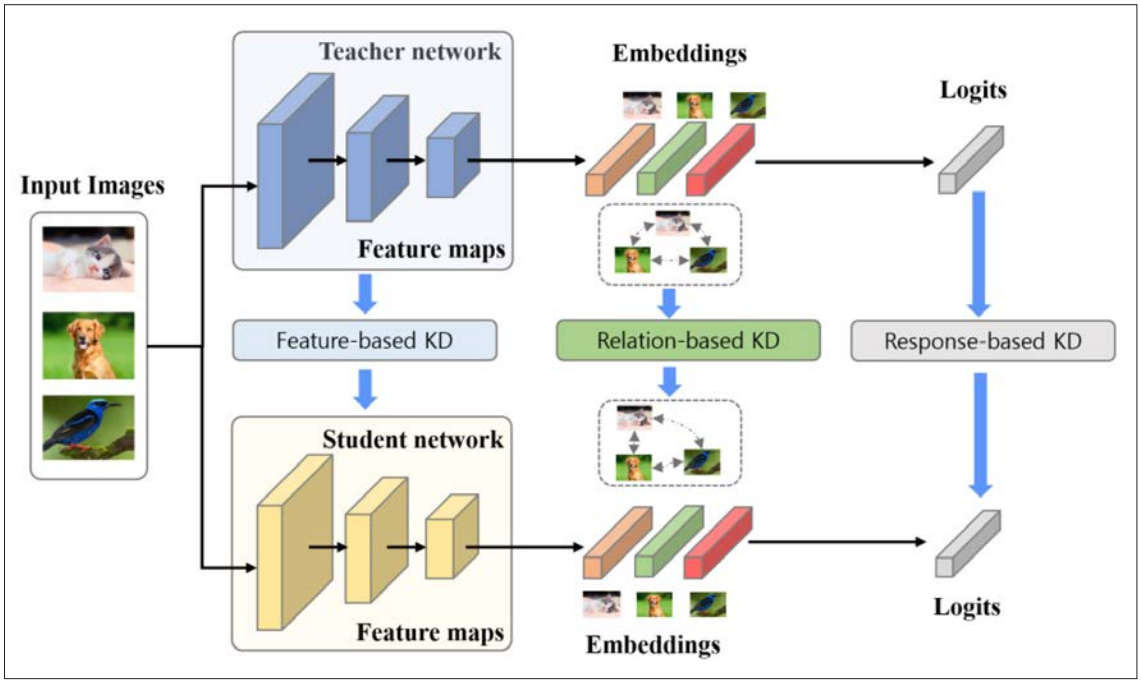
지식 증류(KD)는 큰 모델(Teacher Model)의 정보를 작은 모델(Student Model)로 압축하여 성능을 유지하면서도 연산 비용을 줄이는 기법이다. 작은 모델이 큰 모델의 예측 분포를 학습하여 성능 저하 없이 경량화된다. Knowledge Distillation은 처음 자연어 처리(NLP) 분야에서 대형 언어 모델의 성능을 경량 모델로 전이하기 위한 방법으로 제안되었다. 이후 이미지 분류 및 객체 탐지와 같은 컴퓨터 비전 분야에서도 널리 적용되었으며, 최근에는 생성형 AI 모델의 경량화에도 활용되고 있다. Knowledge Distillation은 크게 네트워크 Response 기



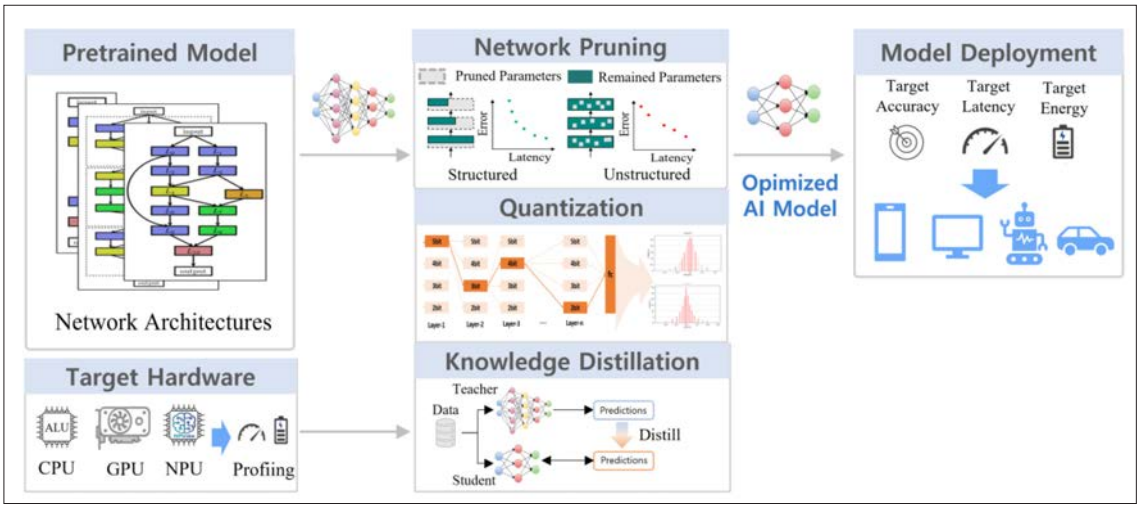
<그림 2> CNN 기반 네트워크 Quantization 기술

반 증류, 특성 맵 기반 증류, 관계 기반 증류 방식으로 구분된다. Response 기반 증류는 큰 모델(Teacher)의 출력 확률값인 로짓을 작은 모델(Student)이 학습하는 방식으로, 연산량을 줄이면서도 성능 저하를 최소화할 수 있다. 특성 맵 기반 증류는 네트워크 중간층 표현을 압축하여 경량 모델의 표현력을 향상시키는 방법이며, 관계 기반 증류는 여러 데이터 샘플 간의 관계 정보를 유지하도록 학습하는 기법이다. 이러한 기술들은 온디바이스 AI 모델을 위한 최적화 과정에서 활용되며, 모바일 기기나 임베디드 시스템에서도 고품질 미디어 생성 및 분석이 가능하도록 지원하고 있다.

앞서 언급한 세 가지 방식의 경량화 기술을 2개 이상 동시 적용하여 경량화를 최대화하기 위한 시도들도 연구되고 있다. IBM Research는 Knowledge Distillation과 Quantization을 동시에 적용하여 ResNet, 계열 CNN 모델을 INT8로 압축하면서도 원본 모델 대비 정확도 손실을 0.5% 이내로 유지했다[1]. Qualcomm AI Research는 CNN 모델에 Pruning과 QAT를 결합하여 ResNet, MobileNet 계열 모델을 기존 대비 최대 43배까지 압축했으며, 정확도는 원본의 97% 이상 유지했다[2]. Google의 MediaPipe는 TensorFlow Lite를 활용하여 Pruning과 Quantization을 동시에 적용해 손 인식, 얼굴 인식 등의 AI 모델을 온디바이스 환경에 최적화했다. 이를 통해 기존 모델 대비 크기를 최대 75%까지 줄이고, 연산 속도를 약 2~3배 향상하여 실시간 성능을 제공한다[3]. 한국전자기술연구원에서는 EfficientNet 계열 모델에 Pruning과 QAT를 결합하여 원본 모델의 크기를 86%까지 감소시키며, 정확도 손실을 0.5% 이내로 유지했다. 특히, Pruning 적용시 네트워크가 구동되는 하드웨어 특성을 프로파일링을 통해 분석하고 동작 시간 감소에 효과적인 네트워크의 구조를 제거하여 기존 Pruning 기법 대비 네트워크 추론 속도를 향상시켰다.



<그림 3> CNN 기반 네트워크 Distillation 기술



<그림 4> CNN 기반 경량화 기술 적용 파이프라인

2. 미디어 제작을 위한 온디바이스 생성형 AI 경량화 기술

생성형 AI 모델의 경우 CNN과 달리 어텐션(Attention) 메커니즘을 기반으로 네트워크 구조가 이루어진다. 대표적인 생성형 모델 아키텍처로 트랜스포머(Transformer)와 Diffusion 모델에 사용되는 U-Net이 있으며, 이러한 구조의 모델을 경량화하기 위해 기존 CNN 기반의 경량화 기술을 변형 적용하거나 아키텍처의 구조적인 분석을 통해 새로운 접근 방법을 제시하는 방향의 연구가 진행되고 있다. 먼저, 기존의 경량화 기술의 변형 적용 연구에서는 메모리 사용량 감소와 연산 속도 향상에 효과적인 Structured Pruning 기반의 경량화 기술이 연구되고 있다. 예를 들어, 트랜스포머 기반 언어 모델에서 Attention Head Pruning을 적용하여 특정 헤드를 제거하고도 성능을 유지하는 연구가 진행되었으며[4], Knowledge Distillation과 결합하여 어텐션 블록의 차원 수를 감소시키면서 온디바이스 환경에서 추론 성능을 확보한 MobileBERT가 제안되었다[5]. 이 모델은 기존 언어 모델인 BERT 대비 GLUE 태스크에서 추론 성능 손실을 0.6% 이내로 유지하며, 5.5배 빠른 추론 속도를 달성하였다. 특히, 모바일 환경(Google Pixel 4)에서 latency가 63ms로 추론이 가능하여 생성형 언어 모델을 경량화하는데 효과적인 기반 연구로서 활용되고 있다.

생성형 AI 모델은 CNN 대비 거대한 모델 크기로 인해 양자화 적용시 훈련 기반의 QAT를 적용하는데 있어, 많은 계산 자원이 필요하여, 주로 PTQ 기반의 양자화 기술의 성능을 높이기 위한 연구가 이루어지고 있다. PTQ는 데이터 정밀도가 낮아짐에 따라 QAT 대비 상대적으로 정확도 손실이 크게 발생하기 때문에, 네트워크의 핵심 연산에 대해 혼합 정밀도 연산(Mixed-Precision Computing)을 적용하는 연구가 진행되고 있다. 이러한 연구는 양자화 손실에 민감한 네트워크의 중요한 연산에 FP16이나 FP8 수준의 고정밀도를 유지하면서 나머지 연산을 4비트 수준의 낮은 정밀도로 양자화함으로써, 모델의 압축 성능을 높이는 동시에 이미지 생성 품질을 유지하였다[6]. 구체적으로, U-Net 기반의 Diffusion 모델에서 샘플링 과정은 이미지 생성 품질을 확보하는데 중요한 연산으로 양자화 적용시 높은 정밀도를 유지하고, 나머지 연산을 낮은 정밀도로 할당하는 양자화 기법이 적용되었다.

기존의 CNN 네트워크 경량화 기법과는 다른 접근 방법으로 트랜스포머 기반 생성 모델에서 압축 토큰(Compressed Token)의 사용과 임베딩 벡터 크기 축소를 통해 모델을 경량화하는 방법이 연구되고 있다. BERT, GPT 등의 언어 모델에서 학습 과정에서 중요하지 않은 토큰을 제거하는 Token Pruning을 통해 기존 모델의 연산량을 줄이고 추론 속도를 향상시키는 기법이 제안되었으며[7], Stable Diffusion 모델의 경우, 유사한 정보를 포함한 토큰을 하나의 압축 토큰으로 병합하는 Token Merging 기법이 제안되었다[8]. 또한, Diffusion 모델의 확산 과정을 기존 픽셀 공간의 임베딩 벡터에서 잠재(Latent) 공간의 임베딩 벡터로부터 수행함으로써 연산량과 메모리 사용량을 크게 줄이는 동시에 이미지 생성 품질을 확보한 LDM 모델이 제안되었다[9].

생성형 AI 모델에서 가장 연산량이 많은 단계는 디코딩 및 샘플링 과정으로, 이를 최적화하기 위해 점진적 디코딩(Progressive Decoding) 기법이 적용되고 있다. 예를 들어, 언어 모델에서는 초안을 작성한 후, 의미적으로 연관된 추가 정보를 점진적으로 확장하는 방법이 제안되었으며[10], 이미지 생성에 있어 GAN 네트워크에서 저해상도 이미지로부터 고해상도로 점진적으로 이미지를 생성하여 연산량을 줄이는 방법이 제안되었다[11]. Diffusion 모델에서도 추론 단계에서 샘플링 속도가 저하되는 문제를 해결하기 위해 확률적 샘플링 단계 축소(Stochastic Sampling Reduction)가 연구되고 있으며, 기존 샘플링 단계의 균일한 타임스텝 방식을 개선하여 최적의 스텝 크기를 탐색함으로써 적은 단계의 샘플링으로도 더 높은 품질의 이미지를 생성할 수 있는 방법이 제안되었다[12].

Ⅲ. 온디바이스 미디어 제작 서비스 현황 및 전망

1. 온디바이스 미디어 제작을 위한 생성형 AI 서비스 현황

온디바이스 환경에서 생성형 AI를 활용한 미디어 제작 서비스는 글로벌 기업을 중심으로 빠르게 확산되고 있다. 해외의 대표적 사례로는 메타(Meta)의 AR 이펙트 플랫폼인 'Spark AR'이 있다. Spark AR은 GAN(생성적 적대 신경망) 및 Diffusion 모델 등의 생성형 AI 기술을 활용하여 사용자가 스마트폰에서 직접 고품질의 AR 콘텐츠를 제작하고 공유할 수 있도록 지원한다. 클라우드 의존 없이 얼굴 추적, 가상 메이크업, 배경 교체 등 다양한 콘텐츠를 즉시 생성하여 개인화된 미디어 경험을 제공하고 있다. 애플은 온디바이스 기반 생성형 AI 기술을 자사의 모바일 앱인 'Clips'에 적용하여, 사용자의 음성을 실시간으로 텍스트화하고 자동으로 스타일화된 자막과 이모지를 삽입해 주는 기능을 제공한다. 이로 인해 사용자는 간편하게 고품질의 개인화된 비디오 콘텐츠를 제작할 수 있다. 또한, 구글의 Pixel 스마트폰 역시 Tensor 프로세서를 활용하여 실시간 배경 흐림 효과, 고해상도 이미지 확대, 실시간 HDR 영상 생성 등의 고급 미디어 제작 기능을 온디바이스에서 제공하고 있다.

한국에서도 온디바이스 생성형 AI 기술의 적용이 활발히 이루어지고 있다. 네이버의 대표적인 서비스 '제페토(Zepeto)'는 생성형 AI를 활용하여 사용자의 얼굴 데이터를 기반으로 실시간으로 3D 아바타를 생성하고 개인화된 가상 공간에서 다양한 콘텐츠를 제작하고 공유할 수 있도록 지원한다. 사용자들은 자신의 표정과 몸짓이 반영된 아바타를 통해 보다 몰입감 있는 소셜 미디어 활동이 가능해졌다. 스타트업 플라스크(Plask)는 AI 모션 생성 기술을 온디바이스로 구현하여, 스마트폰에서 촬영한 인물 영상을 즉각적으로 디지털 아바타로 전환하는 서비스를 제공하고 있다. 이를 통해 일반 사용자들도 전문 장비나 복잡한 프로세스 없이 손쉽게 고품질의 디지털 콘텐츠를 제작할 수 있다. 또한, 카카오는 카카오톡 메신저 내 카메라 앱에 온디바이스 기반의 생성형 AI 기술을 적용하여 실시간 AR 필터 및 얼굴 변형 효과를 제공하고 있으며, 이는 사용자들의 즉각적이고 개인화된 미디어 콘텐츠 제작을 촉진하고 있다.

2. 온디바이스 미디어 제작에서 생성형 AI 서비스의 전망

온디바이스 환경에서의 생성형 AI 기술은 향후 미디어 제작 분야에서 중요한 핵심 요소가 될 것으로 전망된다. 현재 디바이스 내부의 AI 연산 성능이 지속적으로 발전하고 있으며, 저전력 고성능의 모바일 및 임베디드 프로세서의 발전에 따른 과거 클라우드 환경에서만 가능했던 복잡한 미디어 콘텐츠 생성 작업이 디바이스 자체에서도 실시간으로 가능해질 것으로 예상된다. 첫째, 온디바이스 환경에서 생성형 AI 서비스는 개인화와 맞춤화 측면에서 더욱 발전할 것이다. AI 기술이 사용자의 디바이스 내에서 직접 구동되므로 개인의 데이터나 선호도를 정확히 반영한 콘텐츠 생성이 가능해진다. 예를 들어, 개인 맞춤형 아바타, 개인화된 음성 및 음악 콘텐츠 생성 등 사용자의 취향을 정교하게 분석하여 맞춤형 미디어 콘텐츠를 제공할 수 있게 될 것이다. 둘째, 개인정보 보호와 보안에 대한 요구가 높아짐에 따라 온디바이스 생성형 AI 서비스의 수요는 지속적으로 증가할 것이다. 사용자 데이터가 디바이스 내에서 처리되고 외부 클라우드로 전송되지 않기 때문에 데이터 유출이나 개인 정보 보호 문제에서 안전성을 확보할 수 있으며, 이는 더욱 많은 사용자들이 안심하고 서비스를 사용할 수 있게 하는 중요한 요인이 될 것이다. 셋째, 기술적 측면에서 볼 때, AI 모델 경량화 기술과 하드웨어 가속 기술의 발전이 두드러질 것이다. 앞으로의 디바이스는 생성형 AI 모델을 실시간으로 실행할 수 있도록 하드웨어 차원의 최적화가 더욱 강조될 것이다. 특히, 혼합 정밀도 연산과 적응형 모델 크기 조절, 동적 네트워크 구조 등 다양하고 고도화된 경량화 기법의 발전이 예상되며, 이러한 기술들은 온디바이스 AI 서비스를 더욱 효율적이고 강력하게 만들어 줄 것이다. 넷째, AR과 VR 기술과의 융합도 중요한 발전 방향 중 하나이다. 생성형 AI 기술이 실시간 3D 콘텐츠 및 가상 환경을 제작하는데 활용됨으로써 온디바이스에서 실감나는 몰입형 콘텐츠 제작이 가능해질 것이다. 이를 통해 엔터테인먼트, 교육, 쇼핑 등 다양한 분야에서 사용자의 경험을 크게 향상시키는 혁신적인 콘텐츠들이 등장할 것이다. 마지막으로, 다양한 산업과의 결합을 통한 응용 사례가 증가할 것이다. 의료, 제조, 교육, 커뮤니케이션 등 다양한 산업 영역에서 온디바이스 기반 생성형 AI 기술을 활용한 미디어 콘텐츠 제작과 자동화된 서비스가 활성화될 것이며, 이에 따라 산업 전반에 걸쳐 생산성과 효율성이 크게 향상될 것으로 기대된다.

Ⅳ. 결 론

온디바이스 AI를 위한 생성형 인공지능 경량화 연구는 기존 CNN 기반 네트워크 경량화 기법을 확장하는 것뿐만 아니라, 트랜스포머, Diffusion 모델 등 새로운 아키텍처에 특화된 최적화 방법을 적용하는 방향으로 발전하고 있다. 하드웨어 친화적인 모델 설계, 혼합 정밀도 연산, 적응형 모델 크기 조절 등의 기술이 더욱 중요해질 것으로 예상되며, 이를 활용한 모바일 및 엣지 디바이스에서의 생성형 AI 상용화가 가속화될 것이다.

약어 정리

CNN	Convolutional Neural Network
GAN	Generative Adversarial Networks
BERT	Bidirectional Encoder Representations from Transformer
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit

참 고 문 헌

[1] Esser, S.K, et al., “Learned step size quantization”, International Conference on Learning Representations (2020)
[2] Ying Wang et al., “Differentiable joint pruning and quantization for hardware efficiency”, In European Conference on Computer Vision, ECCV, 2020
[3] Google AI Edge blog, https://ai.google.dev/edge/litert/models/model_optimization?hl=ko
[4] Kyuhong Shim et al., “Layer-wise Pruning of Transformer Attention Heads for Efficient Language Modeling”, International SoC Design Conference, ISOCC, 2021
[5] Zhiqing Sun et al., “MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices”, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, 2020
[6] Tianchen Zhao et al., “MixDQ: Memory-Efficient Few-Step Text-to-Image Diffusion Models with Metric-Decoupled Mixed Precision Quantization”, In European Conference on Computer Vision, ECCV, 2024
[7] Saurabh Goyal et al., “PoWER-BERT: Accelerating BERT Inference via Progressive Word-vector Elimination”, Proceedings of the 37th International Conference on Machine Learning, PMLR, 2020
[8] Daniel Bolya, Judy Hoffman, “Token Merging for Fast Stable Diffusion”, CVPR Workshop on Efficient Deep Learning for Computer Vision, 2023
[9] Robin Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, International Conference on Computer Vision and Pattern Recognition, CVPR, 2022
[10] Bowen Tan et al., “Progressive Generation of Long Text with Pretrained Language Models”, International Conference of the North American Chapter of the Association for Computational Linguistics, NAACL, 2021
[11] Tero Karras et al., “PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY, STABILITY, AND VARIATION”, International Conference on Learning Representation, ICLR, 2018
[12] Shuchen Xue et al., “Accelerating Diffusion Sampling with Optimized Time Steps”, International Conference on Computer Vision and Pattern Recognition, CVPR, 2024

저 자 소 개

박 종 희

- 2008년 : 광운대학교 컴퓨터소프트웨어 전공 학사
- 2010년 : GIST 정보통신공학과 석사
- 2015년 : GIST 정보통신공학과 박사
- 2015년 ~ 2019년 : 현대자동차 남양연구소 책임연구원
- 2019년 ~ 현재 : 한국전자기술연구원 지능형영상처리연구센터 책임연구원
- 주관심분야 : 컴퓨터 비전, 생성형 인공지능, 온디바이스 AI

곽 종 훈

- 2012년 : 인하대학교 정보통신공학부 학사
- 2014년 : KAIST 전기 및 전자공학과 석사
- 2014년 ~ 2021년 : 현대자동차 남양연구소 연구원
- 2021년 ~ 현재 : 한국전자기술연구원 지능형영상처리연구센터 선임연구원
- 주관심분야 : 컴퓨터 비전 및 영상처리, 딥러닝 모델 경량화