‘Efficient AI for Media’ 특집호를 내며

배 성 호
경희대학교

최근 미디어 AI는 고해상도 영상, 장시간 비디오, 위성 영상, 3차원, 4차원 장면, 비전언어모델로 빠르게 확장되고 있습니다. 모델의 표현력은 커졌지만, 이를 학습하고 저장하며 실제 장치에서 실행하는 비용도 함께 증가하였습니다. 이제 효율성은 완성된 모델을 사후에 줄이는 부가 기술이 아니라, 어떤 정보를 센싱하고 어떤 데이터를 학습하며 어떤 표현을 저장하고 어떤 하드웨어에서 실행할지를 처음부터 결정하는 핵심 설계 원리입니다. 이번 특집호는 센싱과 추론의 통합, 학습 데이터의 압축, 복원 모델의 경량화, 3차원 장면 표현의 압축, 장편 영상 생성의 제작 문법, 대규모 멀티모달 모델의 현장 배포를 아우르는 전문가 원고 6편을 준비하였습니다.

첫 번째 기고문(“장편 비디오 생성을 위한 영상 문법 기반의 비디오 생성 기술”)에서는 짧은 클립을 넘어 여러 장면이 이어지는 장편 영상을 만들기 위한 설계 원리를 다룹니다. 긴 영상을 한 번에 출력하기보다 시나리오를 장면과 숏으로 나누고, 카메라 구도·움직임·조명·전환·편집 리듬을 영상 문법으로 명시하여 인물과 배경의 일관성을 유지합니다. 국내외 제작 사례를 통해 생성·VFX·편집을 연결하는 파이프라인을 살펴보고, 영상 문법 데이터, 상용 편집 도구 연동, 품질 검수와 저작권 관리를 향후 과제로 제시합니다.

두 번째 기고문(“Efficient AI for Media: VLM을 미디어 현장에 넣는다는 것”)에서는 비전언어모델(VLM)의 높은 성능과 실제 산업 배포 사이의 간극을 다룹니다. 영상이 만드는 방대한 시각 토큰, 도메인 변화에 따른 불안정성, GPU 메모리와 지연 제약을 짚고, 양자화·프루닝·그래프 최적화·하드웨어 친화적 매핑을 결합하는 NetsPresso의 최적화 스택을 소개합니다. 또한 전통적인 검출·추적 모델과 VLM을 결합하고, 필요한 영역만 고해상도로 다시 관찰하는 ERGO를 활용한 NVA(Nota Vision Agent)의 구조와 KBS 재난특보, 교통 관제, 산업 안전 적용 사례를 제시합니다. 향후 VLM이 World Model로 발전하더라도 메모리, 전력, 지연, 규제라는 배포 조건은 남기 때문에 효율화 기술의 역할은 더 커질 것이라는 전망도 제시합니다.

세 번째 기고문(“Toward Efficient Static and Dynamic Gaussian Splatting”)에서는 실시간 고품질 렌더링으로 주목받는 3D Gaussian Splatting(3DGS)과 동적 장면을 위한 4DGS의 효율화 연구를 체계적으로 정리합니다. 수백만 개의 Gaussian이 요구하는 큰 저장 공간과 연산량을 줄이기 위한 방법을 파라미터 압축과 구조 재설계라는 두 축으로 분류하고, 프루닝, 양자화, 엔트로피 부호화, 앵커 기반 표현, canonical deformation, level-of-detail 구조의 핵심 원리를 비교합니다. 또한 정적·동적 장면 데이터셋과 객관적 화질 평가 메트릭, 모델 크기, 렌더링 속도 등의 평가 지표를 함께 다루어, Gaussian Splatting을 제한된 장치와 장시간 동적 콘텐츠로 확장하기 위한 연구 방향을 제시합니다.

네 번째 기고문(“불확실성 인지 기반 지식 증류를 통한 고효율 위성 영상 품질 개선 연구”)에서는 저해상도 다중분광 영상과 고해상도 전정색 영상을 결합하는 PAN-sharpening을 효율화하기 위한 U-Know-DiffPAN을 소개합니다. 고성능 확산 교사 모델을 작은 학생 모델로 증류할 때 모든 픽셀을 같은 방식으로 모방시키면 경계, 작은 객체, 질감처럼 복원이 어려운 영역의 세부 정보가 약해질 수 있습니다. 이를 해결하기 위해 U-Know loss는 교사의 픽셀 단위 불확실성을 이용하여, 교사가 확신하는 영역에서는 교사의 안정적인 출력을 따르고 어려워하는 영역에서는 실제 정답 신호를 더 강하게 학습시킵니다. 이는 단순한 모델 축소를 넘어, 복원 태스크에서 반드시 보존해야 할 입력 정보를 선택적으로 전달하는 지식 증류의 방향을 보여줍니다.

다섯 번째 기고문(“추론과 센싱의 통합을 통한 효율적인 AI 기반 비전 센서 시스템 설계”)에서는 센서가 모든 데이터를 고정된 방식으로 수집한 뒤 프로세서로 보내는 단방향 파이프라인의 한계를 지적하고, 추론 결과를 다음 센싱 과정에 되먹임하는 폐루프 구조를 소개합니다. 첫 번째 사례는 직전 프레임의 손 영역에 따라 처리 영역과 해상도를 조절하는 동적 피드백 ISP와 손 자세 추정 가속기를 FPGA에 통합한 near-sensor 시스템입니다. 두 번째 사례는 객체 검출 결과를 이용해 노출 시간, 이벤트 임계값, 관심 영역 읽기를 센서 칩 내부에서 조절하는 in-sensor 이벤트 카메라입니다. 두 사례는 불필요한 데이터의 획득과 이동부터 줄일 때 실시간성, 강건성, 에너지 효율을 함께 높일 수 있음을 보여줍니다.

여섯 번째 기고문(“Dataset Condensation for Object Detection: A Survey”)에서는 대규모 학습 데이터를 소수의 합성 또는 선별 샘플로 대체하는 데이터셋 압축 기술을 객체 검출의 관점에서 조망합니다. 객체 검출은 분류와 달리 객체의 정체성뿐 아니라 bounding-box 기하, 크기 변화, 배경 영역과 전경의 조화, 다중 객체의 문맥까지 보존해야 합니다. 본 기고문은 관련 방법을 inversion 기반 합성, diffusion 기반 합성, compositional 합성으로 구분하고, 대표 방법들의 구성과 실험 조건을 비교합니다. 아울러 검출기 구조, 입력 해상도, 라벨 형식이 연구마다 달라 공정한 비교가 어렵다는 문제를 짚고, 기하 인지 목적함수, 이미지와 주석의 공동 생성, long-tail·open-vocabulary 환경, 강건성 및 개인정보 보호를 향후 과제로 제시합니다.

이번 “Efficient AI for Media” 특집호 출간을 위해 바쁘신 가운데 귀한 원고를 준비해 주신 저자 분들께 깊이 감사드립니다. 본 특집호가 미디어 AI의 효율성을 모델 크기나 연산량 하나의 문제로 한정하지 않고, 센싱에서 데이터, 표현, 모델, 하드웨어, 서비스까지 이어지는 시스템 문제로 이해하는 데 도움이 되기를 기대합니다.