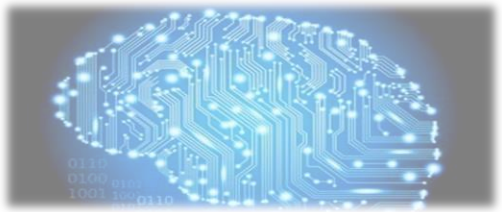
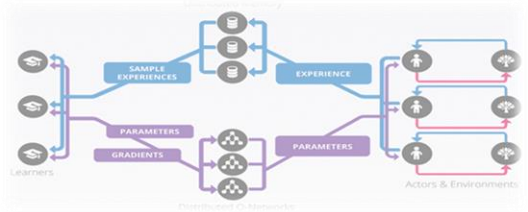

Multimodal Dataset Research

AutoML 연구실
석사과정 문아성

21.05.18

Multimodal Dataset Research

- MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records

Xu, Zhen, David R. So, and Andrew M. Dai. "MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records." arXiv preprint arXiv:2102.02340 (2021).

- Dataset: Medical Information Mart for Intensive Care (MIMIC-III)

Johnson et al. MIMIC-III, a freely accessible critical care database. Scientific data 3: 160035.

Real-world EHR data 53,423 hospital admissions *from 2001 to 2012*

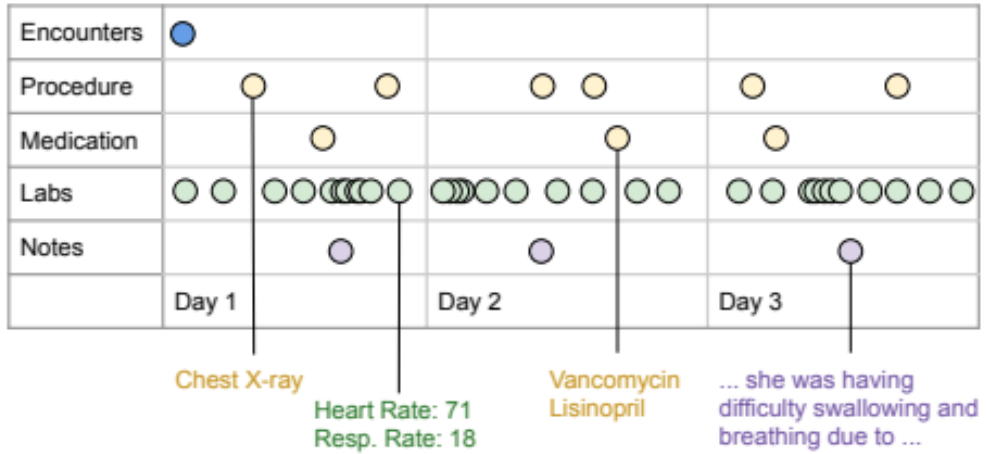

Figure 1: An illustrative example of a patient's record. It contains multiple feature modalities, including categorical features, continuous features and clinical notes across time.

Multimodal Dataset Research

- BM-NAS: Bilevel Multimodal Neural Architecture Search

Yin, Yihang, et al. "BM-NAS: Bilevel Multimodal Neural Architecture Search."

arXiv preprint arXiv:2104.09379 (2021).

- Dataset

MM-IMDB: Internet Movie Database

{posters, plots, genres, meta information} *25,959 movies*

NTU RGB+D: Action Recognition {RGB video, Skeleton pose}

56,800 action samples (40 subjects, 80 view points, 60 classes)

EgoGesture: Gesture Recognition {Video(RGB-D, Depth)}

24,161 gesture samples (50 subjects, 6 scenes, 83 classes)

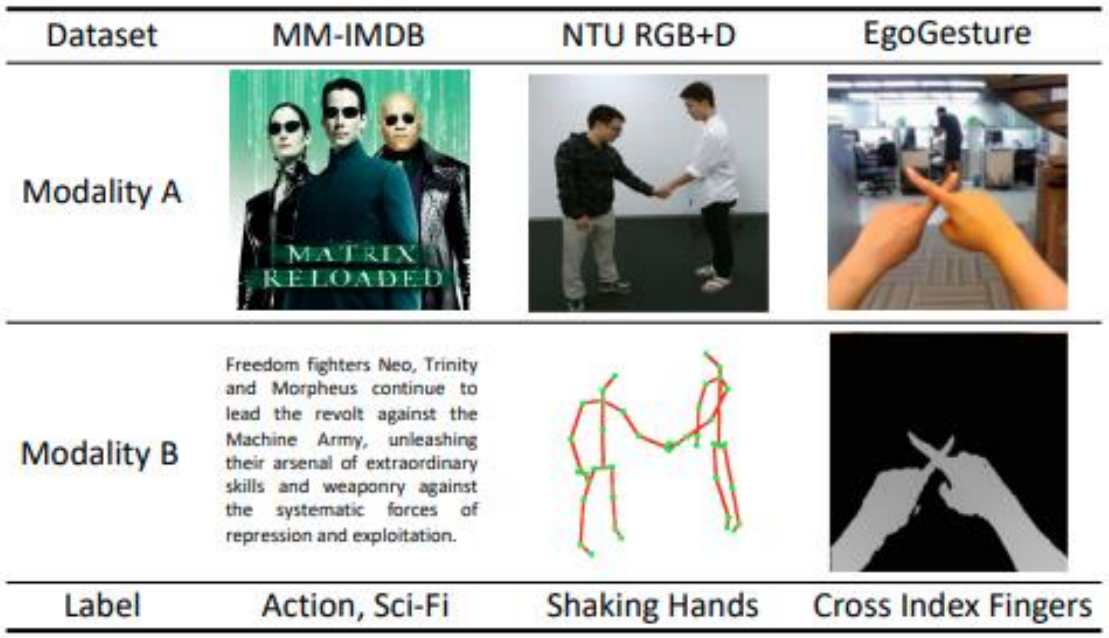

Figure 4. Examples of the evaluation datasets.

Multimodal Dataset Research

- Deep Multimodal Neural Architecture Search

Yu, Zhou, et al. "Deep Multimodal Neural Architecture Search."

Proceedings of the 28th ACM International Conference on Multimedia. 2020.

- Dataset

VQA-v2: Visual Question Answering {images, questions, answers}

200,000 images, 1,106,000 questions

Flickr 30K*: image with caption from Flickr website {images, captions}*

31,000 images

RefCOCO, RefCOCO+ and RefCOCOg: *{images, query}*

66,600 images and queries

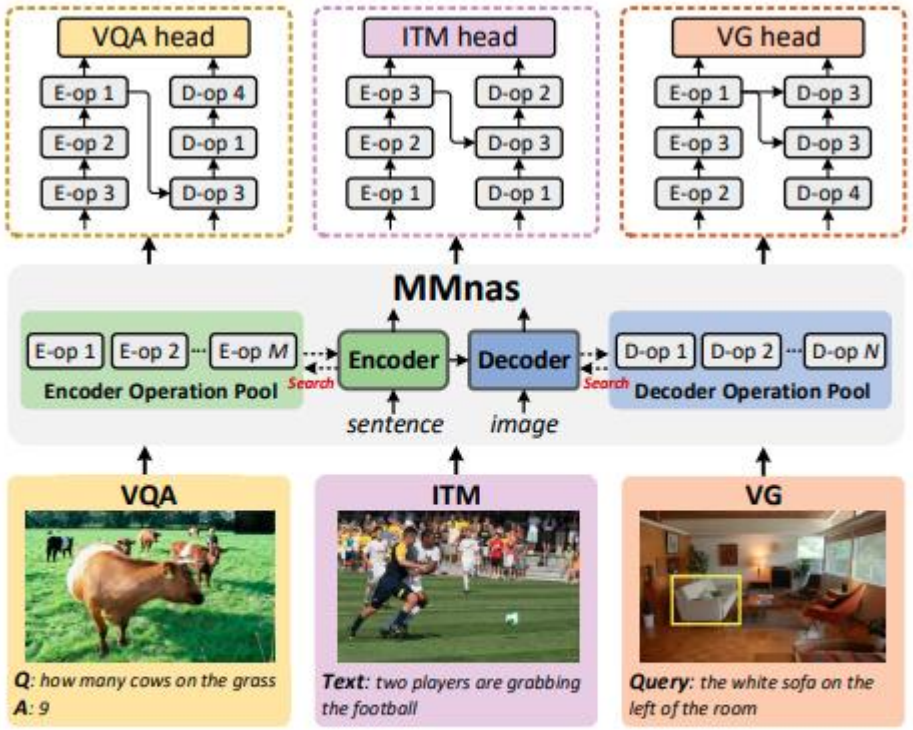

Figure 1: Schematic of the proposed generalized MMnas framework, which searches for the optimal architectures for the VQA, image-text matching, and visual grounding tasks.