Multimodal Deep Representation Learning for Disaster Information Management
Yudong Tao, Yuexuan Tu, Mei-Ling Shyu
Department of Electrical and Computer Engineering

Introduction
Disaster Information Management
- Natural disaster causes lots of damages.
- Rapid responses are essential to mitigate human and property losses.
- Large amounts of multimedia data are generated on social media in real-time and can be potentially helpful.
- An automated disaster information management system that can collect, analyze, and synthesize the data on the fly to produce compact and useful information is needed.

Multimedia Big Data
- Large volume: 500+ terabytes Facebook data per day;
- Multimodality: various types of data, including images, videos, audios and texts;
- Complexity: multimedia data serve to convey complex semantic information and ease the understanding of new knowledge.

Challenges and Limitations
- Mixed semantic meanings confuse the computer model to detect and classify concepts;
- One modality may have limited resources;

Multimodal Deep Learning extracts distinct information contained in different modalities and adequately fuses them to learn more comprehensive knowledge for better semantic understanding.

Contribution
- Incorporate sequential information from both audio and visual models.
- Extract the most discriminative and high-level feature representations.
- Propose a novel fusion technique to synthesize both audio and video features based on Multiple Correspondence Analysis (MCA) algorithm [5].

Proposed Framework
The proposed framework is shown in Figure 1, which can be divided into three main components, namely the visual feature extraction model, the audio feature extraction model, and the fusion model. Both feature extraction models compute the likelihoods of a concept appearing in the corresponding input video or audio clip. The features (or likelihoods) are then taken as the inputs by the fusion model to eventually learn/mine the semantic meaning of the videos.

Visual Feature Extraction
- Inception-v3 factorizes convolutions into smaller convolutions (e.g., 7x7 convolutions -> 3x3 convolutions) and also adds batch normalization to the deeper layers.
- An LSTM unit is built on top of the spatial component to analyze the temporal correlations among the video sequences to generate visual features capturing both spatial and temporal information.

Audio Feature Extraction
- The workflow diagram of audio feature extraction is shown in Figure 2.
- SoundNet is a pre-trained model, utilizing a student-teacher diagram to learn the acoustic representations from unlabeled videos.
- Bidirectional GRU layers are implemented to capture the temporal features in the audio data.

MCA-Based Fusion
- The scales of features from both models are different, while the ranking of the likelihood can be considered as the same scale and more suitable to be integrated.
- MCA captures the correspondence among the categorical variables (e.g., ranking) and maps them to a continuous space.
- As shown in Figure 3, our proposed fusion model applies MCA to the score rankings of both models and then use Support Vector Machine (SVM) to conduct the final classification.

Dataset and Experimental Results
Dataset Overview
- Videos crawled from YouTube in 2017 and divided into clips;
- 1000 clips from Hurricane Harvey for training + 450 clips (from Hurricane Irma) for testing;
- Manually assigned each clip a label among nine concepts: (a) demonstration, (b) emergency response, (c) flood and storm, (d) human relief, (e) damage, (f) victim, and (g) briefing.

Experimental Result Highlights (Table 1)
- Measures: accuracy, weighted F1, and macro F1;
- Upper part: single modality model; Bottom part: fused model;
- Comparing Line 2 & Line 3: Temporal Correlation is important for visual data and improve the results a lot;
- Comparing Line 3 & Line 4: Direct SVM fusion can improve accuracy, while F1 scores decrease -> the audio model performs worse for minority concepts and the fusion confuses the visual model;
- Comparing Line 4 & Line 5: MCA improve all the measures and eliminate the negative effects on minority concepts from the audio model.

Conclusion
This poster presents a novel multimodal deep learning framework that considers sequential information from both audio and video models. Furthermore, an MCA-based fusion technique is proposed to synthesize the outputs from various models. In our experiments, we demonstrate how the proposed framework improves the accuracy from single-modality models.

Reference