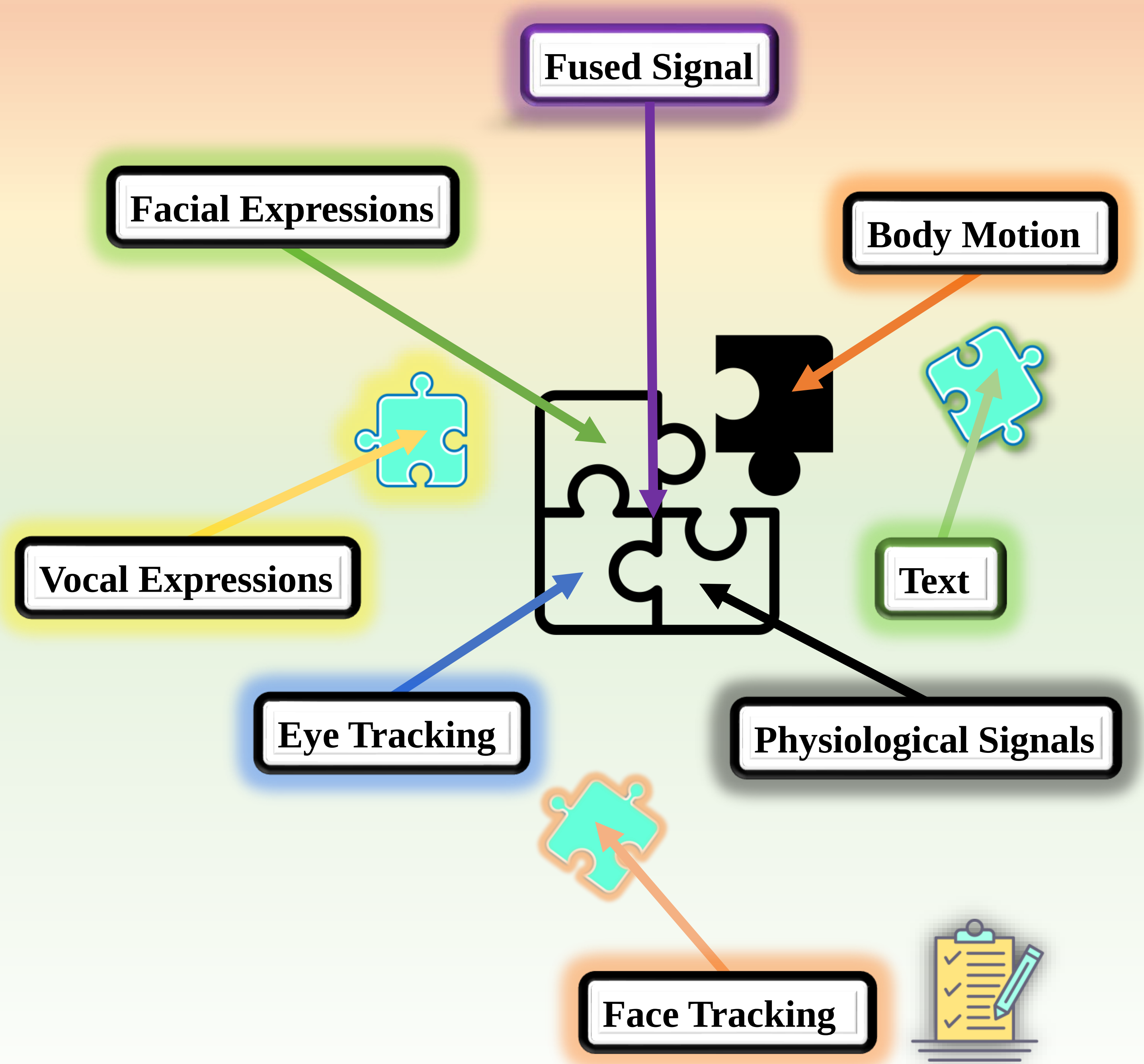
Multi-Modal Fusion

- **Outline**
  - ➢ Definition
  - ➢ Applications
  - ➢ Importance
  - ➢ Challenges
  - ➢ Methods
    - ✓ Early Fusion
    - ✓ Late Fusion
    - ✓ Intermediate Fusion (Hybrid Fusion)



By: Seyed Muhammad Hossein Mousavi
mosavi.a.i.buali@gmail.com
Lugano, Switzerland
October 2024

Multi-Modal Fusion

- **Definition**
  - Multimodal fusion refers to the process of combining data from multiple modalities (e.g., text, body tracking, images, audio, video, or sensor data) to **improve decision-making or prediction in machine learning models**.
  - By integrating information from different sources, multimodal fusion aims to **capture complementary information**, **providing a more comprehensive understanding and improving model accuracy**.
  - It has different names depending on the application:
    - **Data Fusion - Often used interchangeably with multimodal fusion**, particularly in applications involving sensors and signals from multiple sources.
    - **Sensor Fusion** - Commonly used in **robotics and automotive technologies**, where data from various sensors (like cameras, radar, and GPS) are combined.
    - **Information Fusion** - This term is generally used in contexts where **different types of information (not necessarily sensory)** are integrated.
    - **Feature Fusion** - This specifically refers to the **combination of features extracted from different modalities** before further processing, such as in machine learning models.
    - **Early Fusion, Late Fusion, and Intermediate Fusion (Hybrid Fusion)** - These terms describe the **stage at which fusion occurs** in the processing pipeline
      - **Early fusion** combines data at the input level.
      - **Late fusion** combines outputs near the decision stage.
      - **Intermediate Fusion (Hybrid Fusion)** combines features after some individual processing but before the final decision-making stage.

Multi-Modal Fusion

- **Definition**
  - **Early fusion/feature level/input level**
    - ✓ The terms "**early**" or "**input level**" in the context of multimodal fusion refer to the stage at which different data modalities are combined **in the processing pipeline**.
    - ✓ This **involves integrating data from various modalities at the very beginning of the data processing workflow**. The integration happens **before any significant analysis or modeling occurs**.
    - ✓ In early fusion, **raw data or initial features extracted from each modality are combined** to create a **single, unified dataset**. This dataset is then used as the input for subsequent processing steps, such as feature engineering, dimensionality reduction, or directly feeding into a machine-learning model.
    - ✓ Handling a **single combined dataset can sometimes simplify the modeling process** because we are working with one dataset rather than multiple separate sets.
    - ✓ **In summary, early or input-level fusion is about combining data at the earliest possible stage in the data processing sequence, aiming to maximize the use of all available information from different modalities right from the beginning.**

Multi-Modal Fusion

- **Definition**
  - **Late fusion / decision level**
    - ✓ Late fusion refers to the practice of **combining the outputs from independent models, each trained on different modalities**, at a late stage in the processing pipeline. This combination occurs after significant individual processing of each modality.
    - ✓ The integration at this stage typically **involves merging predictions, decisions, or processed features from each separate model.**
    - ✓ **Decision level fusion is essentially another term for late fusion**, emphasizing the point at which the integration occurs. This method **combines the final outputs, such as classification decisions or prediction scores, from models that have processed their respective modalities separately**.
    - ✓ The fusion involves methods like **majority voting, weighted averaging, or more sophisticated algorithms.**
    - ✓ **Both terms describe a fusion approach that waits until the end of the processing sequence to integrate data, focusing on combining final model outputs rather than raw data or intermediate representations**.
    - ✓ **This method is particularly useful when each data type requires distinct processing or when the independence of modalities is crucial to maintain until the final stages of analysis.**

Multi-Modal Fusion

- **Definition**
  - ➢ **Intermediate Fusion / Hybrid Fusion**
    - ✓ Intermediate fusion refers to integrating data from multiple modalities **at a point between the initial data input stage and the final decision-making stage**.
    - ✓ **This method involves merging features or decisions that are not as raw as in early fusion nor as final as in late fusion**.
    - ✓ In this approach, **features or partial decisions from different modalities are combined after some independent processing but before a comprehensive model or decision process is applied**.
    - ✓ This allows for the preservation of some modality-specific information while still taking advantage of the potential cooperation between different data types.
    - ✓ **Hybrid fusion is often used interchangeably with intermediate fusion**, emphasizing the **blend of characteristics from both early and late fusion approaches**.
    - ✓ The term "**hybrid**" highlights the **method's flexibility in merging modalities at various stages of the processing pipeline**. It can involve a **combination of feature-level integration and decision-level aggregation, making it adaptable to the specific needs** of the analysis or application.
    - ✓ In summary, intermediate or hybrid fusion is a nuanced approach that strategically **combines elements from both ends of the fusion spectrum,** offering a tailored solution that can adapt to various complexities in multimodal data analysis. This flexibility makes it especially useful in **applications where integrating modalities at only the beginning or end of the processing pipeline would not capture all necessary information** effectively.
    - ✓ It is so similar to late as it mostly uses neural networks for train but not for decision-making. After feature extraction in the early stage, a neural network will be used to train and extract more relations or features between them, and finally, another neural network does the job for the late stage of classification, like weighting.

Multi-Modal Fusion

- **Applications**
  - **Early Fusion**: Common in applications where integrating data at the feature level can help the model learn interdependencies between modalities effectively.
  - Often used in fields like image processing, where combining visual features from different sources (e.g., different camera angles or different spectral bands) can be beneficial before any analytical modeling.
  - **Late Fusion**: **Frequently used** because it allows for the development and optimization of separate models on different modalities, which is advantageous when the modalities are very different in nature or when they require distinct preprocessing steps.
  - Popular in decision-making systems, such as in robotics or multimodal biometric authentication, where different systems provide independent assessments that are then combined.
  - **Intermediate Fusion**: Less common than early and late fusion, but it is gaining traction as computational methods become **more sophisticated.**
  - Useful in complex scenarios where neither early nor late fusion alone provides optimal results, such as in some advanced machine learning and AI applications involving complex data types.
  - Overall, **late fusion tends to be more common across a broader range of applications due to its flexibility and the ease of integrating existing systems or models**.
  - In case of having **body motion and physiological signals**:
    - ✓ For an emotion recognition task using **body motion and physiological data** with different timestamps, handling the data synchronization and temporal alignment becomes crucial. Given this scenario and the nature of the data, **late fusion or intermediate fusion** is suitable.

Multi-Modal Fusion

- **Importance**
  - **Enhanced Accuracy and Robustness**: **Combining data from multiple modalities generally leads to more accurate and robust decision-making or prediction models.** Each modality may capture different aspects of the phenomenon being studied, and their integration can provide a more holistic view that reduces the uncertainty or error associated with relying on a single source of data.
  - **Comprehensive Analysis**: Multimodal fusion allows for a more comprehensive analysis by leveraging the strengths of different types of data. For instance, in medical diagnostics, **combining imaging data with lab results can lead to better patient outcomes than using either alone.**
  - **Improved Fault Tolerance**: Systems designed with multimodal data inputs are often more fault-tolerant. **If one sensor or data source fails or provides noisy data, the system can still function effectively** by relying on other modalities. This redundancy is crucial in critical applications such as autonomous driving and healthcare.
  - **Contextual Awareness**: In many real-world applications, the context provided by one modality can significantly enhance the information from another. **For instance, in emotion recognition, physiological signals might indicate stress levels, while facial expressions provide context about the type of emotion being experienced**.
  - **Operational Efficiency**: In industrial and business applications, the fusion of data from multiple sensors or sources can optimize operations, reduce costs, and improve safety by providing real-time insights that are not possible with unimodal data.

Multi-Modal Fusion

- ## Challenges
  - **Data Heterogeneity Challenge**: Different modalities often have different data types, scales, and formats. For instance, combining text, audio, and video requires handling structured and unstructured data, which **vary widely in dimensionality and feature space**. **This variability can complicate preprocessing and feature extraction, making it difficult to integrate data effectively.**
  - **Temporal and Spatial Misalignment Challenge**: Data collected from different sources **may not be aligned in time or space**. For example, sensor data might be collected at different rates, or images and audio might not be synchronized. **Misalignment can lead to significant integration challenges**, requiring sophisticated alignment techniques that can introduce delays or inaccuracies.
  - **Data Quality and Consistency Challenge**: **Variations in quality and consistency across data sources can affect the reliability of the fusion process**. Noise, missing data, or biased data in one modality can skew overall results. Ensuring consistent quality across modalities is essential but challenging, often requiring robust preprocessing and data-cleaning methods.
  - **Privacy and Security Concerns Challenge**: **Multimodal data often includes sensitive information**, which raises privacy and security concerns, especially when data integration exposes new vulnerabilities. Addressing these concerns requires secure data handling practices and may complicate data sharing and integration efforts.

Multi-Modal Fusion

- ## Methods

  ➢ Not all, but some main methods are:

  - ✓ **Early Fusion (Feature-Level Fusion) - Concatenation**: Combine feature vectors from different modalities **into a single feature vector** before feeding them into a machine learning model.
  - ✓ **Early Fusion (Feature-Level Fusion) - Principal Component Analysis (PCA):** Used **to reduce the dimensionality of concatenated features** from multiple modalities.
  - ✓ **Early Fusion (Feature-Level Fusion) - Canonical Correlation Analysis (CCA):** Identifies and fuses **correlated features across modalities.**
  - ✓ **Late Fusion (Decision-Level Fusion) - Weighted Averaging:** Combine predictions from different modality-specific models by **assigning weights to each modality's** output.
  - ✓ **Weighting features refers to assigning different levels of importance to features in a model.**
  - ✓ **Intermediate Fusion (Hybrid Fusion) - Multimodal Deep Learning:** Use deep learning networks (e.g., CNNs, LSTMs) for each modality and **then fuse the learned representations** at different levels (either feature-level or decision-level).
  - ✓ **Intermediate Fusion (Hybrid Fusion) - Attention Mechanism:** Use **attention layers to dynamically weigh contributions from different modalities**.
  - ✓ **Multimodal Autoencoders:** These **learn to compress multiple modalities into a shared latent space** and can reconstruct the input, making them effective in denoising or missing modality tasks.

Multi-Modal Fusion

- **Methods**

  - **Early Fusion (Feature-Level Fusion)**
    - ✓ **Feature Concatenation**: **Directly concatenates feature vectors from different modalities into a single, large feature vector**.
    - ✓ This simple and straightforward method is often the **baseline approach in early fusion**.
    - ✓ **Principal Component Analysis (PCA):** A statistical technique used to reduce the dimensionality of large data sets by transforming the original variables into a new set of variables (principal components), which are linear combinations of the original variables and capture the maximum variance within the dataset.
    - ✓ **Canonical Correlation Analysis (CCA):** A way to **derive features from two sets of variables by finding linear combinations of the variables in each dataset** that are maximally correlated with each other.
    - ✓ It's particularly useful for **discovering the relationships between two sets of modalities**.
    - ✓ **Deep Learning**: Uses deep learning frameworks, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to fuse and learn from **raw data inputs from multiple modalities simultaneously**.
    - ✓ Deep learning models, like CNNs or RNNs, are used to **fuse raw data from multiple modalities directly at the input level**. The network learns to extract and combine features from these different data sources into a unified representation.
    - ✓ This allows for capturing inter-modal dependencies right from the start, potentially leading to **richer and more descriptive feature sets.**
    - ✓ **Autoencoders** can be used to **directly combine raw data from multiple modalities by feeding them into a single autoencoder that learns to compress this data into a unified latent representation**. **The goal here is to reconstruct the combined input data from this compressed representation.**

Multi-Modal Fusion

- **Methods**

  ➢ **Early Fusion (Feature-Level Fusion)**

  ✓ **Generalized Multiview Analysis (GMA):** An advanced technique that **extends Canonical Correlation Analysis (CCA) to handle more than two sets of variables simultaneously**.

  ✓ GMA seeks to find a common subspace where the projections of these different datasets are maximally correlated.

  ✓ This method is particularly effective when you need to integrate and analyze data from multiple sources, such as audio, video, and textual data, **in a way that uncovers and exploits the hidden relationships among them**.

  ✓ **Convolutional Neural Networks (CNNs):** CNNs are particularly effective for processing grid-like topology data, such as images and videos. In early fusion, **CNNs can be used to process visual data from multiple sources or modalities simultaneously**, learning to extract and combine features directly from the **raw data**.

  ✓ Commonly used in scenarios involving image and video data from different cameras or sensors to recognize patterns or objects.

  ✓ **Recurrent Neural Networks (RNNs)**: RNNs are designed to handle sequential data, such as audio or time-series sensor data. For early fusion, RNNs can process inputs from multiple temporal sources, integrating features over time. Useful in applications like speech recognition, where audio data may be combined with temporal text data (like subtitles or scripts) to enhance understanding and context.

  ➢ **So neural networks can be used in all three stages: early, late, and intermediate.**

Multi-Modal Fusion

- **Methods**

  - **Late Fusion (Decision-Level Fusion)**

    - ✓ **Deep Learning:** In this approach, **separate deep learning models are trained independently on each modality**, and **their outputs (decisions or predictions) are combined at the final stage, often through simple techniques like voting, averaging, or a learned aggregation model**.
    - ✓ Each modality can be processed using the most suitable architecture for its specific data type and characteristics, with the final fusion focusing on maximizing the decision accuracy based on complementary information.
    - ✓ **Autoencoders:** At this stage, are used to **independently encode data from each modality**, **and then the encoded representations are combined for a final decision-making process**.
    - ✓ Alternatively, the outputs of separate models for each modality might be fed into an autoencoder to learn a final, unified decision-making criterion.
    - ✓ This setup maximizes the independence of modal analysis, ensuring that each modality is optimally processed before their encoded outputs are used for making the final decisions. It helps in fine-tuning the decision process based on a highly distilled set of features.
    - ✓ **Majority Voting:** It is simple. We will train multiple classifiers on multiple modalities or combinations of modalities for specific emotions. Then, for instance, we have two classes with “joy” and one with “neutral” then the joy will be selected.
    - ✓ Now, one classifier could be on body motion and return jot. Another is on heart rate and return neutral, and the third classifier is on a concatenated of body motion and heart rate and return joy. Then, the joy will be selected.

Multi-Modal Fusion

- **Methods**

  - **Late Fusion (Decision-Level Fusion)**

    - ✓ **Weighted Averaging:** This method involves **taking the outputs from different classifiers, each possibly trained on different modalities, and combining these outputs into a final decision by averaging them**.
    - ✓ Unlike simple voting, weighted averaging **allows us to assign different weights to each classifier's output**, reflecting their relative reliability or accuracy.
    - ✓ **Each classifier is assigned a weight based on its performance, importance, or reliability. For instance, if one classifier is generally more accurate in its predictions, it might be given a higher weight.**
    - ✓ The final output is calculated by taking the **weighted average of the outputs from all classifiers**. If classifiers are predicting a continuous variable or probability, this averaging directly combines these predictions into a **final score**.
    - ✓ **Weighted averaging is particularly useful in scenarios where different modalities contribute unequally to the decision-making process, allowing a more nuanced and accurate integration of their outputs.**
      - ❖ Example: If we have classifiers predicting stress levels based on body motion and heart rate with outputs as probabilities (e.g., 70% stress from body motion classifier and 60% stress from heart rate classifier), and we trust the body motion classifier more, we will weight it higher. If the weights are 0.6 for body motion and 0.4 for heart rate, the final stress probability would be calculated as $0.7\times0.6+0.6\times0.4=0.660.7\times0.6+0.6\times0.4=0.66$ or 66%.

  - There are other techniques like **Stacking, Bayesian model averaging**, and more, but they are not popular.

Multi-Modal Fusion

- **Methods**
  - **Hybrid Fusion (Intermediate Level Fusion)**
    - **Deep Learning:** Deep learning models are applied **after some initial feature extraction** has been performed separately on each modality.
    - The models then learn to integrate these features at a deeper, semantic level, often through techniques like **attention mechanisms or custom fusion layers**.
    - This stage **balances the uniqueness of each modality** with the power of combined analysis, allowing for a more sophisticated interplay between the modalities' features.
      - **Example**: Using a **CNN to process visual data and an RNN to process sequential audio data** in parallel and then **combining their features in subsequent layers before making a classification**.
    - **Auto Encoders**: At this stage, autoencoders are used **after some initial processing or feature extraction** from each modality.
    - The features are then combined and fed into an autoencoder that learns a joint representation of these features.
    - This can also involve modalities being processed by separate encoders with their latent representations being fused.
    - This allows for deeper integration of modal-specific features, capturing more complex interactions between the modalities. **The autoencoder in this case can help in further refining the feature set to emphasize those aspects that are most informative for the task**.

Multi-Modal Fusion

- **Methods**

  - ➢ **Hybrid Fusion (Intermediate Level Fusion)**
    - ✓ **Attention Mechanisms**: Attention mechanisms can dynamically **focus on different parts of data from multiple modalities, weighting the importance of features** based on their relevance to the task.
    - ✓ In natural language processing combined with image data, an attention-based model can focus on specific words and corresponding image regions to improve understanding in a task like image captioning.
    - ✓ **Graph Neural Networks (GNNs):** GNNs are used to **model relationships and interdependencies between different modalities represented as nodes in a graph, integrating information across edges during learning**.
    - ✓ In social media analysis, GNNs can integrate user textual content, image data, and interaction networks to predict user behavior or content popularity.

  - ➢ **Now, also, if we use early and late stages techniques together, then it is hybrid, too.**
  - ➢ **Like concatenation from the early stage followed by weighted averaging by the late.**