Multimodal AI

Advanced Topics in Multimodal Translation and Mapping

September 14, 2024

Overview

1 Examples of Multimodal Translation Systems

Concept of Multimodal Translation and Mapping

- **Multimodal Translation:** Integrating multiple modes of data (e.g., text, images, audio) to enhance translation accuracy and comprehensiveness.
- **Mapping:** Creating correspondences between different modalities to improve understanding and generate more accurate translations.

Goals of Multimodal Translation and Mapping

- Enhance translation accuracy by leveraging contextual information from multiple modalities.
- Improve the ability to handle ambiguous or context-dependent language.
- Enable richer, more informative translations that go beyond text-to-text mappings.
- Facilitate better human-computer interaction by understanding and generating multimodal content.

Example 1: Text and Image Translation

- **System Overview:** Uses images to provide context for translating text.
- **Applications:** Translating text within images (e.g., street signs, product labels).
- **Example:** Google Translate's feature for translating text in images using a smartphone camera.

Example 2: Video and Subtitle Translation

- **System Overview:** Translates spoken language in videos along with the visual context.
- **Applications:** Automatic subtitling and dubbing for films and online videos.
- **Example:** YouTube's automatic captioning and translation services.

Example 3: Multimodal Chatbots

- **System Overview:** Combines text and speech to improve conversational AI systems.
- **Applications:** Customer service, virtual assistants.
- **Example:** Multimodal assistants like Google Assistant and Amazon Alexa, which process both speech and contextual visual information.

Core Components of Multimodal Chatbots

- Text Processing

  (i) **Natural Language Processing (NLP)**: This is fundamental for understanding and generating text. NLP techniques are used to parse, understand, and translate text inputs. This involves tokenization, syntactic parsing, semantic analysis, and contextual understanding.

  (ii) **Translation Models**: These models, such as neural machine translation (NMT) systems, are trained to convert text from one language to another. They leverage vast amounts of bilingual text data to learn translation patterns.

How can we ensure alignment among various modalities while designing a Multimodal Chatbot?

Unified Data Representation

- **Consistent Contextual Understanding**
  - Centralized context management system.
  - Contextual embeddings for text, speech, and images.
- **Cross-Modal Interaction Models**
  - Joint learning models (e.g., transformers).
  - Alignment algorithms for different data representations.

Coherent User Experience

- **Consistent Response Generation**
  - Unified dialogue manager.
  - Synchronization of responses across modalities.
- **User Intent Recognition**
  - Robust intent recognition system.
  - Feedback loops for continuous improvement.

User Interface Design

- **Seamless Modal Transitions**
    - Clear indication of modality changes.
    - Visual and auditory cues for user guidance.
- **Multimodal Integration**
    - Interactive elements supporting multiple inputs/outputs.
    - Consistent feedback across modalities.

Testing and Validation

- **User Testing**
  - Test various multimodal scenarios.
  - Collect iterative feedback for improvements.
- **Performance Monitoring**
  - Monitor metrics such as response accuracy and satisfaction.
  - Analyze errors to identify alignment issues.

Technology Integration

- **Cross-Modal Data Fusion**
  - Data fusion techniques for combining inputs.
  - Real-time data processing.
- **Modular Architecture**
  - Scalable components for independent development.
  - API integration for cohesive operation.

Challenges in Multimodal Translation

Challenge 1: Data Integration

- **Issue:** Integrating and aligning data from different modalities (e.g., text, images, audio) can be complex.
- **Solution Approaches:** Advanced alignment techniques, multimodal embeddings, and synchronization methods.

Challenge 2: Contextual Understanding

- **Issue:** Capturing the context accurately across different modalities is difficult.
- **Solution Approaches:** Using attention mechanisms and contextual embeddings to enhance understanding.

Challenge 3: Computational Complexity

- **Issue:** Multimodal systems often require significant computational resources.
- **Solution Approaches:** Optimization techniques, efficient neural network architectures, and hardware acceleration.

Challenge 4: Ambiguity and Noise

- **Issue:** Dealing with ambiguous data and noisy inputs from various modalities.
- **Solution Approaches:** Robust preprocessing, noise reduction algorithms, and context-aware disambiguation methods.

Multiple Instance Learning in Multimodal AI

What is Multiple Instance Learning (MIL)?

- MIL[1] is a form of **weakly supervised learning**.
- In MIL, training data is organized in **bags** of instances, and a label is assigned to the entire bag rather than individual instances.
- This is useful in cases where individual labels are expensive or difficult to obtain.

[1] https://arxiv.org/abs/1612.03365

Applications of MIL

Medical Imaging

- In computer-aided diagnosis, medical images can be labeled with patient diagnosis, without needing local annotations for diseased regions.

Video/Audio

- Videos or audio files may have tags for the entire clip (e.g., "this video contains a cat and a human"), without knowing when exactly these events occur.

Text

- Document classification may involve determining whether a website (comprising several pages) is about a specific topic, even if only some pages are relevant.

Applications of MIL (cont'd)

Marketing

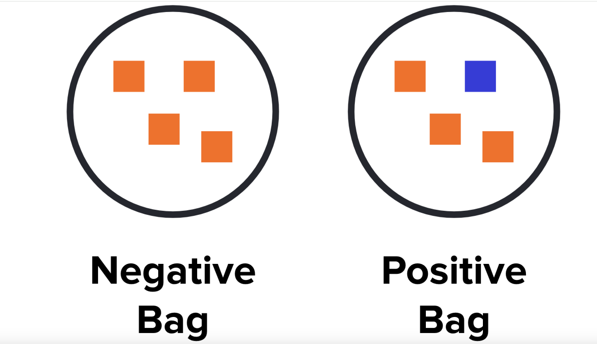
- In marketing campaigns, it may be unclear which individual within a group was influenced by the campaign, but the group's response is measurable.

Time Series

- For gas or water meters, if the total monthly consumption is known, MIL can help estimate usage at a more granular level (e.g., daily).

Representation for MIL

- In the standard MIL assumption, negative bags are said to contain only negative instances, while positive bags contain at least one positive instance. Positive instances are labeled in the literature as witnesses.



Standard MIL Assumption

- The standard assumption in Multiple Instance Learning (MIL):
  - **Negative bags** contain only negative instances.
  - **Positive bags** contain at least one positive instance.
- Positive instances are often referred to as **witnesses**.

Bag Label Definition

Let Y be the label of a bag X, which is a set of feature vectors:

$$X = \{x_1, x_2, \ldots, x_N\}$$

Each instance x_i corresponds to a label y_i.
The label of the bag Y is determined as:

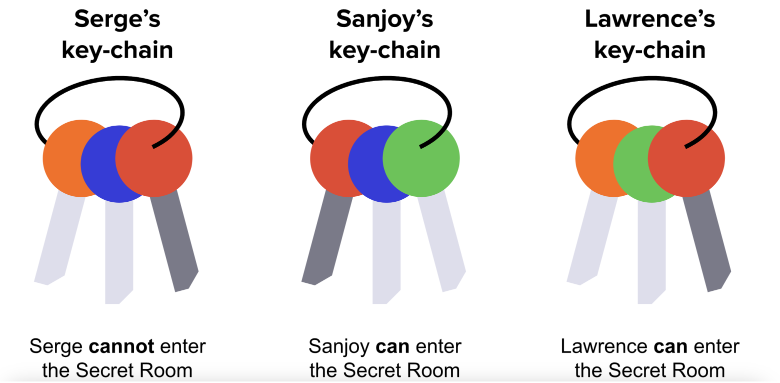
$$Y = \begin{cases} +1 & \text{if } \exists y_i : y_i = +1; \\ -1 & \text{if } \forall y_i : y_i = -1. \end{cases}$$

- If there exists at least one positive instance ($y_i = +1$), the bag is positive.
- If all instances are negative ($y_i = -1$), the bag is negative.

Intuitive Example for MIL Representation

- An intuitive example for MIL is a situation where several people have a specific key chain that contains keys.
- Some of these people are able to enter a certain room, and some aren't. The task is then to predict whether a certain key or key chain can get you into that room.

Contd...



Serge's key-chain

Serge **cannot** enter the Secret Room

Sanjoy's key-chain

Sanjoy **can** enter the Secret Room

Lawrence's key-chain

Lawrence **can** enter the Secret Room

Contd...

- To solve this, we need to find the exact key that is common for all the "positive" keychains – the green key.
- We can then correctly classify an entire keychain – positive if it contains the required key, or negative if it doesn't.
- This standard assumption can be slightly modified to accommodate problems where a single instance cannot identify positive bags, but by its accumulation.
- For example, in the classification of desert, sea, and beach images, images of beaches contain both sand and water segments. Several positive instances are required to distinguish a "beach" from a "desert"/"sea".

Task/Prediction: Instance Level vs Bag Level

- In some applications (e.g., object localization), the goal is to classify **instances** within a bag rather than the bag itself.
- The bag label simply reflects the presence of a target entity within the bag.
- **Key Point**: Bag classification performance is not always indicative of instance classification performance.
    - A *False Positive* in a negative bag misclassifies the entire bag.
    - In positive bags, it does not affect the bag's label or loss at the bag level.

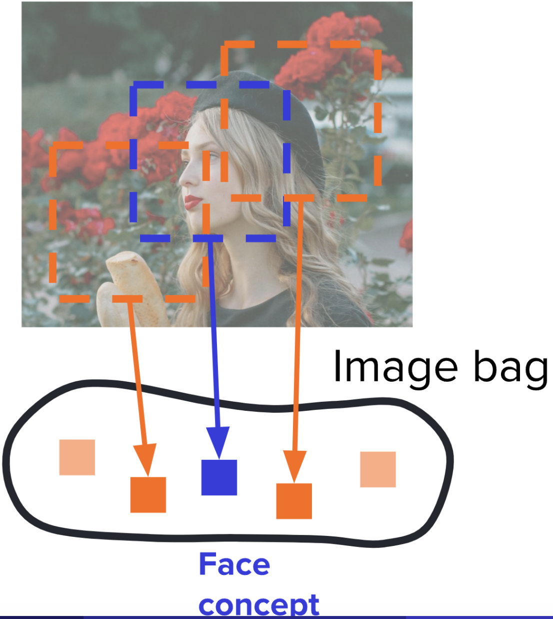
Bag Composition

- Many MIL methods assume that instances in positive and negative bags are sampled from distinct distributions.
- However, in real-world scenarios, relationships within the bag may violate this assumption.

Intra-Bag Similarities

- Instances within the same bag may share characteristics (e.g., similar lighting conditions in image segments).
- Overlapping patches in extraction processes also create shared features, complicating independent sampling.

Intra-Bag Similarities



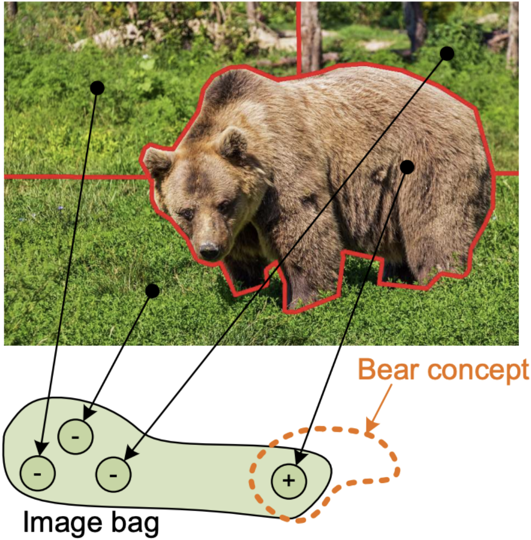
Instance Co-occurrence in Bags

- **Co-occurrence of instances** in MIL bags occurs when instances share a semantic relation.
- This correlation happens when certain objects or subjects are more likely to appear in the same environment or context.
- Examples:
  - **Environment Context**: The subject of an image is more likely to be found in one setting than another (e.g., a bear in nature rather than a nightclub).
  - **Object Correlation**: Some objects are often seen together (e.g., knife and fork).

Example: Bear and Nature

- For example, the bear is more likely to be found in nature than in a nightclub.
- Observing segments of nature in an image can help infer whether the image contains a cocktail or a bear.
- **Key Insight**: Understanding the context in which instances co-occur improves classification accuracy in MIL problems.

Example: Instance Co-occurance



Contd..

- Example of co-occurrence and similarity between instances: Three segments contain grass and forest and are therefore very similar.
- Moreover, since this is an image of a bear, the background is more likely to be nature than a nuclear central control room.

Label Noise in MIL

- Many MIL algorithms, especially those based on the **standard MIL assumption**, rely heavily on the correctness of bag labels.
- In practice, positive instances can sometimes be found in **negative bags** due to:
  - Labeling errors
  - Inherent noise in the data
- Example in computer vision:
  - A negative image (e.g., a house) might still contain positive patches (e.g., flowers).
  - The image may not be annotated as a flower image, despite having such content.

Example: Label Noise in Audio Recordings

- Label noise also occurs when different bags contain different **densities** of positive instances.
- Example in audio recordings:
  - Recording 1 (R1) of 10 seconds contains only 1 second of the tagged event.
  - Recording 2 (R2) of the same duration contains 5 seconds of the tagged event.
  - **R1 is a weaker representation** of the event compared to R2.
- This noise affects the accuracy of MIL algorithms when trying to classify or predict based on such weak instances.

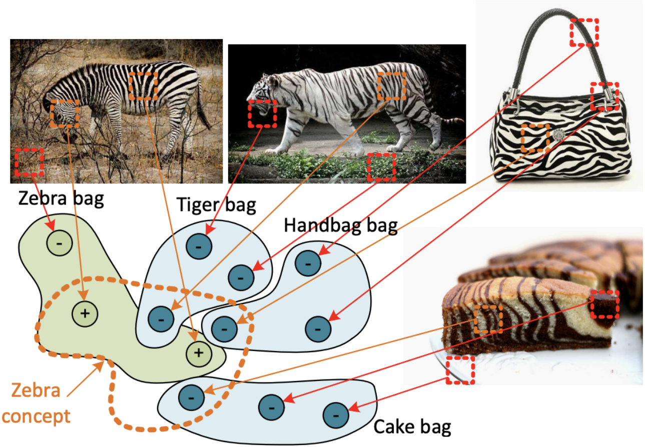
Different Label Spaces in MIL

- In some MIL problems, the label space for **instances** differs from the label space for **bags**.
- Often, the label spaces correspond to different **granularity levels**:
  - Example: A bag labeled as **car** might contain instances labeled as **wheel, windshield, headlights**, etc.
- In other cases, instance labels may not have **clear semantic meanings**.

Example: Zebra Concept Region

- The image shows an example where the positive concept is **zebra** (orange dotted region).
- This region contains various patches extracted from the zebra image.
- However, patches from **negative images** (e.g., white tiger, purse, marble cake) might also fall into the zebra concept region.
- In these cases, patches do not have clear **semantic meaning** easily understandable by humans.

Different Label Spaces



Contd...

- This is an example of instances with ambiguous labels. Zebra is the target concept and instances relating to this concept should fall in the region delimited by the dotted line.
- However, negative images can also contain instances falling inside the zebra concept region.

Bag of Words Approach

- A bag can be represented by its instances, often using techniques like image embeddings.
- The frequency of each instance in a bag is computed, creating a **histogram** that summarizes the bag's content.
- A classifier (e.g., SVM) is trained on these histograms to classify whether a bag is positive or negative.
- **Key Idea**: Each bag is represented by the occurrence of instances (similar to the Bag of Words approach in Natural Language Processing).

Earth Mover Distance SVM (EMD-SVM)

- The Earth Mover Distance (EMD) is a measure of **dissimilarity** between two distributions (e.g., via an image embedding).
- In this context, each bag is treated as a distribution of instances.
- The EMD is used to create a **kernel** in a Support Vector Machine (SVM), which compares the distributions of instances between different bags.
- **Key Idea**: The SVM leverages the EMD kernel to classify bags based on the overall distribution of instances.

Earth Mover's Distance (EMD)

Concept:

- EMD measures the minimum amount of work required to transform one distribution into another.
- Work is defined as the cost of moving weights between distributions.

Mathematical Formulation:

- Given two distributions P and Q:
  - $P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), \ldots, (p_n, w_{p_n})\}$
  - $Q = \{(q_1, w_{q_1}), (q_2, w_{q_2}), \ldots, (q_m, w_{q_m})\}$
- Flow f_{ij} represents the weight moved from p_i to q_j:

$$\sum_{j=1}^{m} f_{ij} \leq w_{p_i} \quad \forall i$$

$$\sum_{i=1}^{n} f_{ij} \leq w_{q_j} \quad \forall j$$

$$\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} = \min\left(\sum_{i=1}^{n} w_{p_i}, \sum_{j=1}^{m} w_{q_j}\right)$$

Contd...

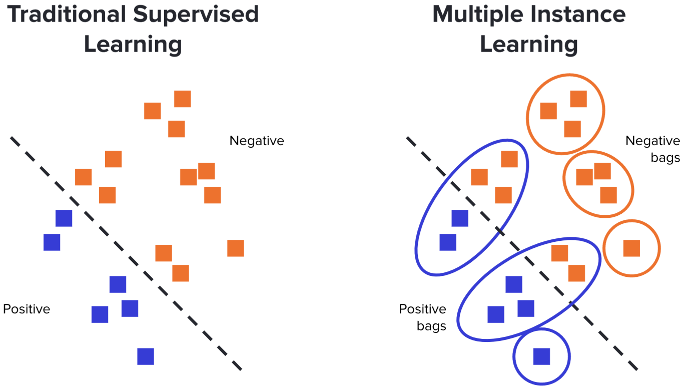
- Cost of moving weight from p_i to q_j is d_{ij}:

$$\text{Work}(P, Q) = \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} \cdot d_{ij}$$

- **EMD** is the minimum work required:

$$\text{EMD}(P, Q) = \frac{\min \left(\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} \cdot d_{ij} \right)}{\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij}}$$

Instance-Space Methods



[Dietterich et al. 1997]

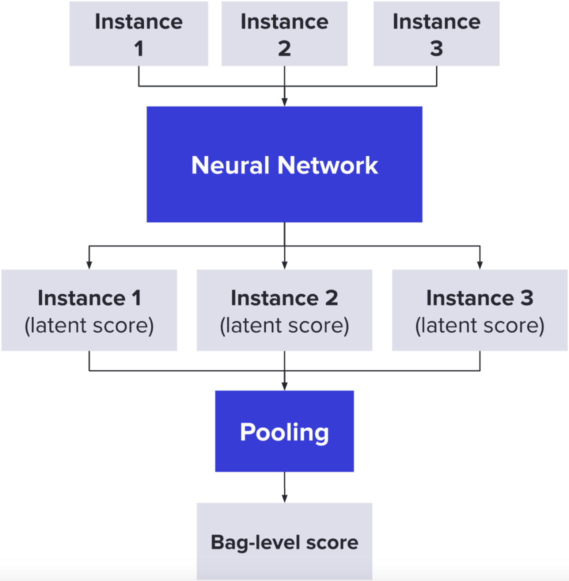
Contd...

- Alternative applications of SVMs (mi-SVM and MI-SVM) were developed for multiple instance learning applications.
- Classically, SVMs try to determine the maximum margin between instances.
- For MIL, since the goal is to have at least one instance in a positive bag as positive, the margin is changed so that condition occurs: at least one instance in a positive bag should have a large positive margin.
- After determining the decision function, the instances' class can be recovered.

Neural Nets for MIL

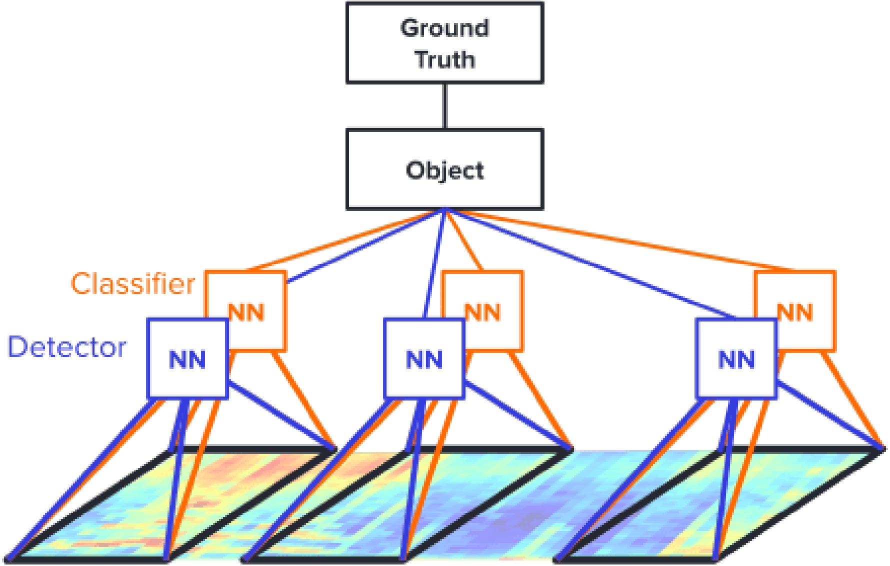
- With a bag-level label, we can have a latent space containing the probability of each segment (using a sequence-based input).
- By applying a pooling operator (max/average pooling), there's just a single score associated with a bag.
- After training, if you want to do an instance-level prediction, the last pooling layer can be removed.
- Usually, max pooling is used for classification problems, while average pooling is applied to regression problems.

Contd...


Instance 1
Instance 2
Instance 3
Neural Network
Instance 1 (latent score)
Instance 2 (latent score)
Instance 3 (latent score)
Pooling
Bag-level score

Neural Networks with Attention Mechanisms

- Attention Mechanisms can also be applied to these kinds of problems.
- Consider an architecture, for audio-level event detection, which uses both a detector and a classifier (symmetric) with just the video-level label to create two separate models.
- The output of the classifier indicates how likely a certain block has tag k.
- The output of the detector indicates how informative the block is when classifying the k-th tag. This way, the model determines how informative a block is for classifying a certain tag.


Ground Truth
Object
Classifier
Detector
NN
NN
NN
NN
NN
NN