Multimodal Deep Learning

Exercise Sheet 1

Date: 29th of April, 2025

Exercise 1: Identifying Multi-Modal Pipelines

For each of the following applications, state which primary multi-modal pipeline type (Translation, Alignment, or Fusion) described in the lecture would be most appropriate. Provide a brief (1-2 sentence) justification for your choice based on the pipeline's core function.

(a) **Input:** A photograph of a menu written in Japanese.
Output: An itemized list of the menu items in English.

How would the pipeline choice change if the goal was:
Input: The same photograph of the Japanese menu and the text query "Does this menu contain shrimp?".
Output: A binary "Yes" or "No" answer.

(b) **Input:** The text description "a fluffy German shepherd dog catching a blue ball mid-air in a park" and a large database containing millions of diverse images.
Output: The index or identifier of a single image from the database that most closely matches the text description.

(c) **Input:** A 10-second video clip from a patient's echocardiogram and a table containing the patient's vital signs (age, heart rate, blood pressure).
Output: A binary classification indicating the presence of significant valve insufficiency.

(d) **Input:** A paragraph from a fantasy novel describing a dragon landing atop a castle tower during a storm.
Output: A generated image visualizing this specific scene.

Exercise 2: Intuition for Manifolds

Many real-world datasets, while represented in a high-dimensional space $\mathbb{R}^D$, often have inherent structure meaning the data points effectively lie on or near a lower-dimensional geometric shape embedded within that space.

For each case described below, reason about and identify:

i. the natural **data space** $\mathbb{R}^D$,

ii. the **data manifold** $M \subset \mathbb{R}^D$,

iii. its **intrinsic dimension** m, and

iv. a sensible **coordinate space** $U \subset \mathbb{R}^m$.

Consider the following cases:

1. **Robot arm.** A planar two-segment robot arm with a fixed base, allowing shoulder rotation and elbow bend. Analyze this system according to points (i)-(iv) above.

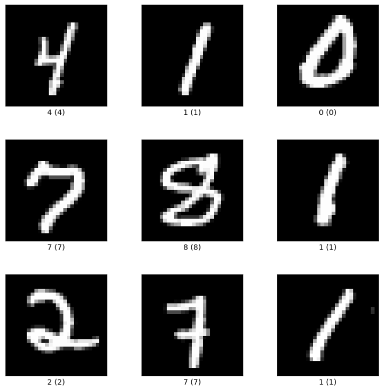

Figure 1: Examples of MNIST handwritten digits. [1]

2. **Hand-written digits (MNIST).** Each image is represented as a 784-dimensional vector (flattened 28x28 pixel image).

   First, analyze this data set according to points (i)-(iv) above. Then, address the following specific questions:

   (a) Argue qualitatively why the intrinsic dimension m is expected to be much smaller than the ambient dimension $D = 784$ (i.e., $m \ll 784$).

   (b) Suggest a type of learned coordinate space (e.g., from dimensionality reduction or generative models) that could potentially "unfold" this data manifold.

   (c) Outline a method to approximate the geodesic distance (distance along the manifold) between two image points on the manifold M.

Exercise 3: Distance in Manifolds

The distance metric used can significantly impact interpretation, especially when dealing with manifolds. Consider the *Euclidean distance* (straight-line in the Data Space $\mathbb{R}^D$) and the *Geodesic distance* (shortest path along the Data Manifold M).

For each scenario, compare the Euclidean and geodesic distance between two points A and B:

(a) **Earth.** A = London and B = Sydney.

(b) **Knotted rope.** Two points on a tangled rope lying on a table.

You can use the concept of a nearest neighbor graph to illustrate why geodesic distance can differ significantly from Euclidean distance on a manifold.

Exercise 4: Topology and Classification

The lecture discussed how deep learning involves transformations that aim to unfold complex data manifolds into simpler representations, often making non-linearly separable data linearly separable in a hidden feature space. In this exercise, you will implement and visualize this process using a simple dataset and a Multi-Layer Perceptron (MLP).

Find instructions in the accompanying file: ‘exercise1.ipynb‘.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, IEEE, 1998.