# Multimodal Deep Learning

## Answer Sheet 1

Date: April 29, 2025

### Exercise 1: Identifying Multi-Modal Pipelines

(a)

1. **Translation pipeline** because it allows a single-modal input (a photograph of a menu written in Japanese) to be processed and return as an itemized list in English. The input is an image of a menu written in Japanese, while the output is a translated menu list in English. **(Image Captioning / Image-to-Text Generation)**
2. **Fusion pipeline** because there are two modalities (an image and a text query) to generate simple classification (Yes/No). The reason that this is not an alignment pipeline is that the model needs to be able to reason based on provided inputs. **(Visual Question Answering (VQS) / Visual Reasoning)**

(b) **Alignment pipeline** because a text description input can be encoded, while images in databases can be pre-encoded, which both text input and image database can find the similarity in embedding vector. This is also called cross-modal retrieval where data is retrieved from different modalities for desired output. **(Cross-Modal Retrieval)**

(c) **Fusion pipeline** because two input modalities – a 10-second video clip and a patient's vital signs table – should be collated for determining a patient's condition. **(Video and Tabular Fusion)**

(d) **(Translation pipeline)** because a single modality input is needed to be encoded and decoded to generate an output image. **(Text-to-Image Generation)**

### Exercise 2: Intuition for Manifolds

1. Robot Arm
   1. **Data Space (D)**: $\mathbb{R}^D \in \{(x_1, y_1, x_2, y_2)\} = \mathbb{R}^4$ with $x_1$ and $y_1$ corresponding to robot shoulder and $x_2$ and $y_2$ corresponding to elbow bend.
   2. **Data Manifold ($\mathbf{M} \subset \mathbb{R}^D$)**: $\mathbb{R}^M \in \{(x_1, y_1, x_2, y_2) \mid \text{robot arm constraints}\}$.
   3. **Intrinsic Dimension (m)**: A two-link robot have at most 2 degree of freedoms (DoFs) – $\theta_1$ and $\theta_2$. Hence, $m = 2$.
   4. **Coordinate Space ($\mathbf{U} \subset \mathbb{R}^m$)**: Hence, $\mathbb{R}^U \subset \mathbb{R}^m \in \{(\theta_1, \theta_2)\} = \mathbb{R}^2$.
2. MNIST
   1. **Data Space (D)**: $\mathbb{R}^D \in= \mathbb{R}^{1\times 784}$ because MNIST is a dataset in grayscale and is also flattened.
   2. **Data Manifold ($\mathbf{M} \subset \mathbb{R}^D$)**: $\mathbb{R}^M \in \{(x, y) \mid \text{readable pixels (space of valid digit shapes)}\}$.
   3. **Intrinsic Dimension (m)**: By applying Principal Component Analysis (PCA), a flattened image dimension can be reduced. Hence, $m << 784$ – depending on data characteristic. For simplicity, $m = 1$ because there are only numbers from 0 to 9 in the MNIST dataset.
   4. **Coordinate Space ($\mathbf{U} \subset \mathbb{R}^m$)**: Hence, $\mathbb{R}^U \subset \mathbb{R}^m \in \{i \mid i \in 0 \text{ to } 10\} = \mathbb{R}^1$.

(a) Since the intrinsic dimension refers to highly featured vectors, and not all 784 pixels are considered high-feature pixels – curves, angles, or borderlines between two colors can be considered high-feature, while the black background should not be much emphasized – so the intrinsic dimension should be much smaller than 784.

**Exercise 3: Distance in Manifolds**

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

- *Euclidean Distance* is a straight-line from London to Sydney passing through Earth core.
- *Geodesic Distance* is the shortest path along data manifold $M$, which is the path that we expect to see plane flying.

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

- *Euclidean Distance* is a straight-line from two points without considering the ball of tangled rope.
- *Geodesic Distance* is the shortest path along data manifold $M$, which is the path between two points after untangled.

**Exercise 4: Distance in Manifolds**

Please see the accompanying file: 'exercise1.ipnyb'.