# MultiModal learning

Kyungseon Lee

September 6, 2024

Seoul National University

# Outline

1. Introduction
2. AVDCNN: Audio-visual speech enhancement using multimodal deep convolutional neural networks - 2017 IEEE
3. Stable Diffusion: High-Resolution Image Synthesis with Latent Diffusion Models - 2022 CVPR
4. CLIP: Learning Transferable Visual Models From Natural Language Supervision - 2021 ICML
5. Conclusion

# Introduction

# Introduction

1. What is multimodal learning?
   - Multimodal learning is a method that takes different types of data, such as text, images, and audio, as inputs and learns from them simultaneously.
2. What is the goal of multimodal learning?
   - The goal is to maximize the joint likelihood of the data observed from different modalities.
   - To do this, we need to convert the data from each modality into a common latent space.

# Introduction

## How the data from each modality is combined?

1. Early Fusion
   - $\hat{y} = f(\mathrm{x}_1 \bigoplus \mathrm{x}_2 \bigoplus \ldots \bigoplus \mathrm{x}_n)$
2. Late Fusion
   - $\hat{y}_i = f_i(\mathrm{x}_i),\ i = 1, 2, ..., n$
   - $\hat{y}_{\text{final}} = \sum_{i=1}^{n} w_i \hat{y}_i$
3. Joint Fusion
   - $\mathrm{z}_{\text{joint}} = f_1(\mathrm{x}_1) \bigoplus ... \bigoplus f_n(\mathrm{x}_n)$; $f_i$ is updated only by the loss from $x_i$
   - $\hat{y} = g(\mathrm{z}_{\text{joint}})$

- $n$: of modality, $\bigoplus$ : concatenation symbol
- $x_i$ : embedding input vector for the i-th modality, $\hat{y}$: output
- $w_i$: the ensemble weight
- $f(.), g(.), h(.)$: Neural network models

# AVDCNN: Audio-visual speech enhancement using multimodal deep convolutional neural networks - 2017 IEEE

# AVDCNN

**Joint Fusion method**

- Objective : Noise reduction in speech data.
- Solution : Use lip movement images as extra data to enhance the speech data.

# AVDCNN

**Solution : Use lip movement images as extra data to enhance the speech data.**

1. $Z_{\text{joint}} = f_A(X_A) \bigoplus f_V(X_V)$
2. $\hat{Y} = g_A \circ g_{common}(Z_{\text{joint}})$ ; enhanced speech data
3. $\hat{X}_V = g_V \circ g_{common}(Z_{\text{joint}})$ ; reconstructed image data

$$\mathcal{L} = \min_{\theta} \left( \frac{1}{K} \sum_{i=1}^{K} \left\| \hat{Y}_i - Y_i \right\|_2^2 + \mu \left\| \hat{X}_{V,i} - X_{V,i} \right\|_2^2 \right)$$

- $\bigoplus$ : Concatenation symbol, $\mu$: weight hyperparameter.
- $X_A$, $X_V$ : An embedding vector of noisy audio and visual data.
- $Y$: The clean audio data.
- $f_A$, $f_V$: The feature extraction encoder functions for audio and visual data.
- $g_A$, $g_V$: The reconstruction decoder functions for audio and visual data.

# Stable Diffusion: High-Resolution Image Synthesis with Latent Diffusion Models - 2022 CVPR

# Stable Diffusion

**Conditional method**

- Objective: Generate the desired image each time.
- Solution : Use a different modality as a condition for image generation.

# Diffusion Model

1. Forward process
   - $q\left(\mathrm{x}_t \mid \mathrm{x}_{t-1}\right) = \mathcal{N}\left(\mathrm{x}_t; \sqrt{1-\beta_t}\mathrm{x}_{t-1}, \beta_t \mathrm{I}_D\right)$
2. Reverse process
   - $p_\theta\left(\mathrm{x}_{t-1} \mid \mathrm{x}_t\right) = \mathcal{N}\left(\mathrm{x}_{t-1}; \mu_\theta\left(\mathrm{x}_t, t\right), \sigma_t^2 \mathrm{I}_D\right)$
3. Objective function

$$\mathcal{L}_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(x_t, t\right)\right\|_2^2\right]$$

with $t$ uniformly sampled from $\{1, \ldots, T\}$.

- $D$ : Dimension of data.
- $\mathrm{x}_0$ : data $\in \mathbb{R}^D$, $\mathrm{x}_t$ : $x$ with noise added $t$ times
- $q\left(\mathrm{x}_0\right)$ : True density function of $\mathrm{x}_0$.
- $\beta_t$ : small positive hyperparameter
- $\epsilon, \epsilon_\theta$ : noise and estimated noise

# Stable Diffusion

**Solution : Use a different modality as a condition for image generation.**

1. input $z_0 = f^E(x_0)$ ; Feature extraction function using an AE.
2. In the denoising process of diffusion, we approximate true reverse process using the Transformer.

$$p_\theta\left(\mathrm{z}_{t-1} \mid \mathrm{z}_t, x_{other}\right) = \mathcal{N}\left(\mathrm{z}_{t-1}; \mu_\theta\left(\mathrm{z}_t, t, x_{other}\right), \sigma_t^2 \mathrm{I}_D\right)$$

$$\mathcal{L}_{SD} = \mathbb{E}_{x, x_{other}, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta\left(x_t, x_{other}, t\right)\|_2^2\right]$$

- $z_T$: $z$ after adding noise T times to become $N(0, I_p)$ noise.
- $\tilde{z}$: The output $z$ generated conditionally.
- $x_{other}$: the embedding vector of the other modality data as the condition.
- $\mu_\theta(.)$: A prediction function that calculates the average, using a Transformer.

# CLIP: Learning Transferable Visual Models From Natural Language Supervision - 2021 ICML

# CLIP

**Self-supervised method**

- Objective: Remove the data labeling process.
- Solution : Use the text attached to the image as input instead of a label.

## Contrastive learning

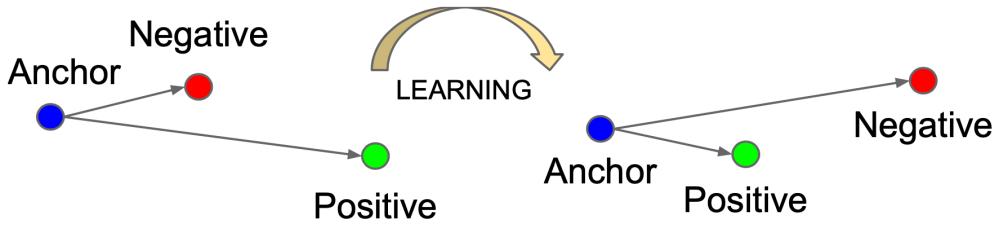


**Figure 1:** Basic contrastive learning process

Basic Contrastive loss: $L(x_p, x_q, y) = y * ||f(x_p) - f(x_q)||^2$

1. $x_p, x_q$ : input pair, $y$ : output, $f(.)$ : contrastive learning model
2. In the representation space, if an input pair is positive, it is mapped close together, and if it is negative, it is mapped farther apart.
3. This might look similar to clustering, but the goal of contrastive learning is to better learn representations through the input pairs.

# CLIP

**Solution : Use the text attached to the image as input instead of a label.**

1. $\mathsf{I}_i = f_I(\mathsf{X}_{I,i}),\ \mathsf{T}_i = f_T(\mathsf{X}_{T,i})$
2. $(I_i,\ T_i)$ ; the correct pair of the image and text data
3. $(I_i,\ T_j),\ i \neq j$ ; the uncorrect pair of the image and text data

$$\mathcal{L}_{contrast}(\mathsf{I},\mathsf{T}) = \sum_i [-log(\frac{exp(I_i \cdot T_i)}{\sum_j exp(I_j \cdot T_i)}) - log(\frac{exp(I_i \cdot T_i)}{\sum_k exp(I_i \cdot T_k)})]$$

$$\hat{y}_i = \left[\frac{exp(I_i \cdot T_j)}{\sum_k exp(I_i \cdot T_k)}\right],\ j = 1, 2, ..., n$$

- $\mathsf{X}_{I,i}, \mathsf{X}_{T,i}$: input of the image data and text data
- $I_i$: The embedding vector of the i-th image
- $T_i$: The embedding vector of the i-th text
- $f_I, f_T$: The embedding functions for the image data and text data

# Conclusion

# Conclusion

## Comments

- In multimodal learning, different papers use various methods to fuse different modalities. Among these, the Joint fusion method is effective at both extracting features from each modality and learning interactions between them.
- Multimodal learning also allows models to be trained without data labeling.
- Additionally, it can be applied in areas like harmful word filtering, human behavior recognition, data augmentation, and noise reduction.