# Multimodal Deep Learning

"Unify multimodal signals into a single vector space and thereby enable cross-modality signal processing."

## representation learning

Multimodal representation

Unimodal Embeddings

1. Visual Representation: image embeddings
2. Language Representation: Text embeddings
3. Vector arithmetic

for word and image embeddings } Word embeddings

Syntactic / semantic regularities

4. Speak Representation: i-Vector approach

## Multimodal Representation:

joint embeddings to leverage the complementarity of multimodal data to represent such concepts more accurately

① Unsupervised Training Methods:

deep Boltzmann machines

autoencoders

deep multimodal similarity model (DMSM)

generate fine-grained multimodal embeddings

deep attentional multimodal similarity model (DAMSM)

" measure the similarity between images

sub-regions and words as an additional loss function for text-to-image generation"

② Supervised Training Methods.

improve the learning of multimodal representation

| | |
|---|---|
| Discriminative factors | intra-modality generative factors |
| (supervised training) | (unsupervised training) |

③ Methods for zero-shot Learning

"certain representations may require pairwise data from different modalities simultaneously."

④ Transformer-based Methods:

Fusion of multimodal Signals

"integrate information extracted from different unimodal data sources into a single compact multimodal representation"

① Simple Operation-based Fusion

② Attention-based Fusion

③ Bilinear Pooling-based Fusion

{ Factorization for Bilinear Pooling
Bilinear Pooling and Attention Mechanisms

# Applications

## Image Captioning

## Text-to-Image Generation

1. GAN-based Methods
2. Generating High-quality images
3. Generating Semantically consistent images
4. Semantic layout control for complex scenes

## Visual Question Answering

## Visual Reasoning