

# Multi-modal Analysis

Ramaseshan Ramachandran

Multimodal learning combines different types of data or sensory channels

# INTRODUCTION

---

- ▶ We use five senses to perceive
- ▶ Mimic human learning processes
- ▶ Combine text, video/images, audio to better understand

# MULTI-MODAL PERCEPTION

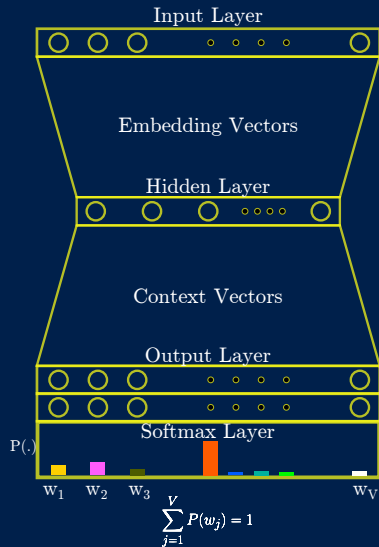
---

- ▶ A thought on good coffee
- ▶ Visualize the scene
- ▶ Sequence of related words and context
- ▶ Describe them using speech
- ▶ Exhibit emotions - nostalgic
- ▶ Some draw the south Indian coffee on a traditional vessels
- ▶ ...

## Word Embedding

- ▶ Form dense vectors for every word to capture its semantic nature
- ▶ Allows similar words to be close in a feature space
- ▶ Algorithms - HALS, COALS, Word2Vec, GloVe, FastText

# WORD2VEC MODEL



## SIMILAR WORDS FOR VIRUS

---

Vocab size	Words in the corpus
637722	222502540

Word	Similarity
virus,	0.889620
viral	0.785719
(herpesvirus)	0.764385
avirus	0.759567
fluav)	0.757418
polio-virus	0.724740
⋮	⋮
(vsv;	0.723436
(denv-2)	0.722825
(cowpox)	0.717185
⋮	⋮

- ▶ Encoder-Decoder Architectures
  - ▶ Solve the challenge of mapping long input sequences of different lengths
- ▶ Attention Mechanism: Enables models to focus on relevant information
- ▶ Transformer Architecture Advantages
  - ▶ Self attention mechanism
  - ▶ Perform parallel operations
  - ▶ Pretrained and fine-tuning capabilities for a specific content
  - ▶ Deep Network Capabilities



# CHALLENGES

---

- ▶ Deep learning models operate on numeric data
- ▶ Challenges in converting unstructured inputs to numeric formats
- ▶ How to combine multi-modal information
- ▶ Identifying relationship across the senses - both contextual (in text) and spatial (in images)

## Connecting thoughts, words and images

- ▶ Thought of a dog → description in text → understand the text → translate the meaning into visual representation

## Text-to-Image Generation

- ▶ **Input** - Chippiparai
- ▶ **Output** - The Chippiparai is a breed of sighthound from the State of Tamil Nadu in southern India. The Chippiparai has typical streamlined sighthound features with long legs and a lean and lithe frame built for speed. The breed is usually white in color, although other colors can be found.



**Input** - A lecture video (video and audio)

**Output**

- ▶ classification of the topic
- ▶ Summary of the lecture, a chapter/section of the topic
- ▶ Translation to another language,
- ▶ Create transcription in another language
- ▶ Identification of words for lip synchronization
- ▶ ...

- ▶ Transformers - CLIP, DALL-E for combining modalities
- ▶ Identify embeddings across modalities
- ▶ Fusing modalities
- ▶ Combining contextual and spatial relationships -Develop embeddings that capture deeper relationships
- ▶ Scalable Systems - Transformers reached its full potential -What next? Develop embeddings that capture deeper relationships.

# COMMON PATTERNS

---

- ▶ Classification
- ▶ Regression
- ▶ Clustering
- ▶ Dimensionality reduction
- ▶ Contextual Association

content...

# CNN: CONVOLUTIONAL NEURAL NETWORK

---

- ▶ CNN: A class of deep learning models designed for image and spatial data
- ▶ ResNet: Introduces residual connections to improve training of very deep networks

## Good at extracting the spatial features from data

- ▶ Convolutional Layers: Extract features.
- ▶ Pooling Layers: Reduce spatial dimensions.
- ▶ Fully Connected Layers: Map features to output.





advances in CV