A Hierarchical Overview of Multimodal Machine Learning

And potential ideas with personal insights

Jianzhu Yao, Jun 10th

Dept. Computer Science and Technology, Tsinghua University

Outline

1. Introduction
2. Taxonomy, Methods, Applications, and Ideas
    1. Representation
    2. Translation
    3. Fusion
    4. Alignment and Co-Learning
3. The idea of multimodal integrated sparsely activated model approach

Introduction

Multimodal: see, hear, feel, smell, taste…

Modalities we explored: natural language, visual signals, vocal signals…

Importance: build models that can process and relate multimodal information

Application: recognition, multimedia retrieval, multimodal interaction and generation, media description…

Research Potential:

Representation, Translation, Fusion, Alignment and co-learning

Taxonomy, Methods, Applications and Ideas

Representation

Definition: To represent any format multimodal data for computational model.

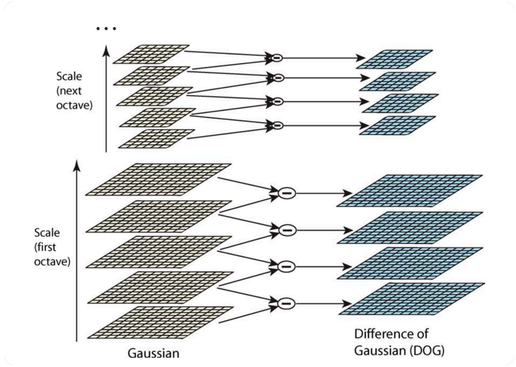
Importance: High-quality representations can make the most of data. (Backbone)

Challenges:

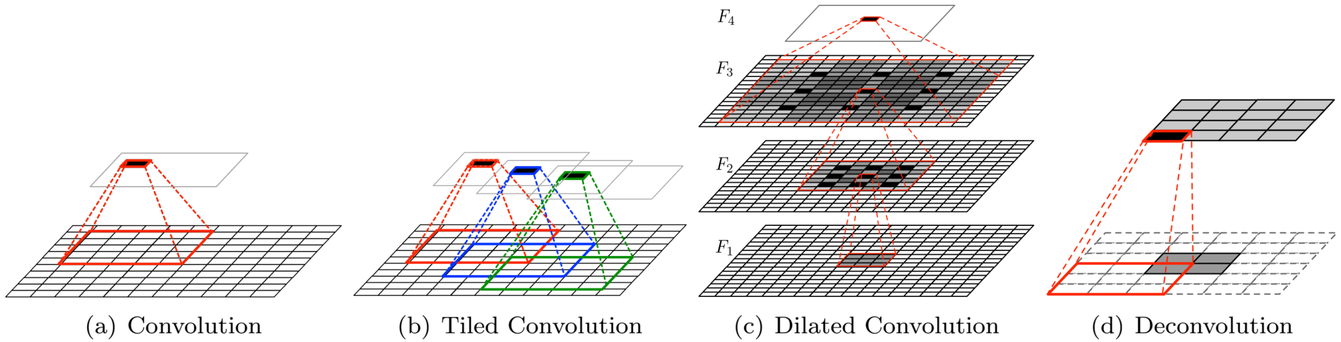
1. Data combination
2. Different level of noise
3. Missing data

Taxonomy, Methods, Applications and Ideas

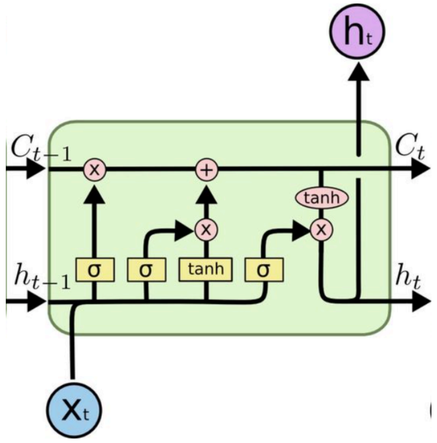
Representation: unimodal representation



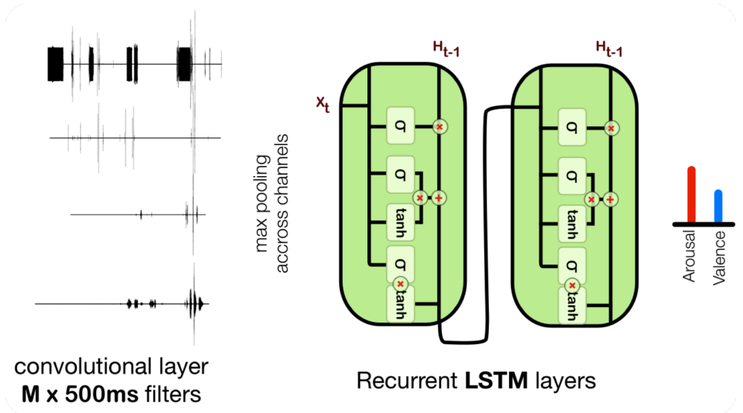
Scale invariant feature transform (SIFT)



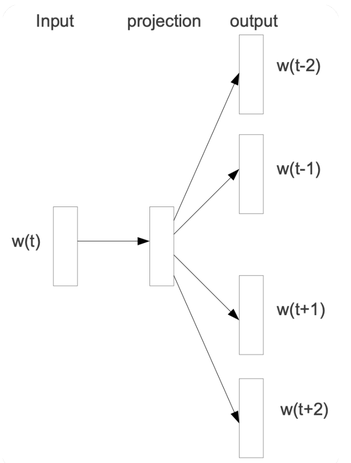
Different types of Convolution Neural Networks (CNN) for image representation



Long Short-Term Memory (LSTM) for speech recognition



LSTM-based speech emotion recognition

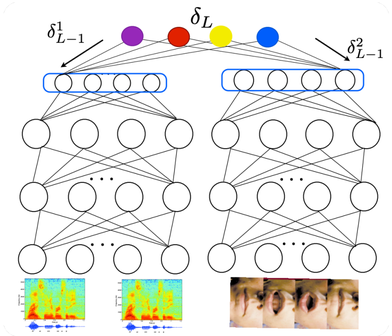


Word embeddings in NLP

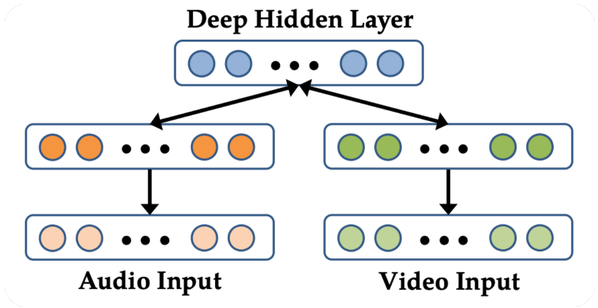
Taxonomy, Methods, Applications and Ideas

Representation: multimodal representations

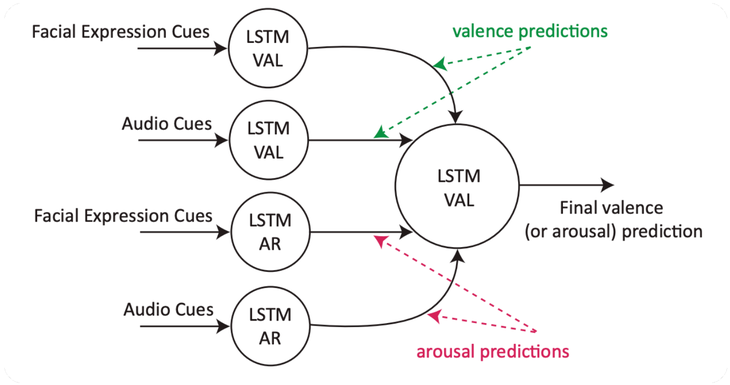
Joint Representation



Audio-Visual DNN

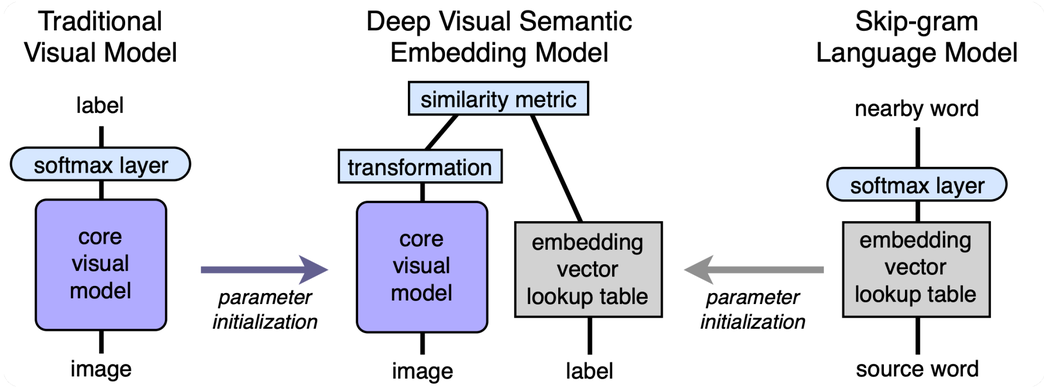


Bimodal DBN (Deep Belief Network)

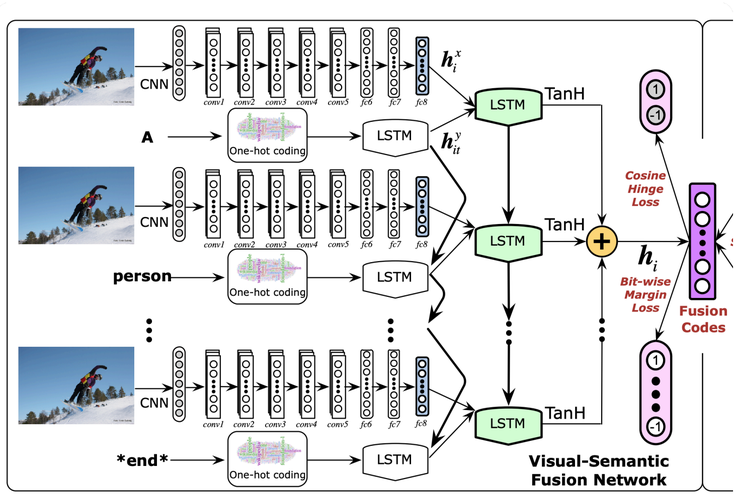


Combination of facial expression and audio cues

Coordinated Representation



Minimize the distance between modalities



A complex LSTM sentence representation model based semantic similarity constraint

Taxonomy, Methods, Applications and Ideas

Translation

Definition: Translating (mapping) from one modality to another

Example: image2text, text2image, video2text, audio2text…

Challenge: Evaluation

Human judgment, Automatic metrics (BLEU, ROUGE, Meteor, CIDEr)

Potential Idea:

Consistency, objectivity, safety, toxicity?

Machine generation detector?

GAN? How to attack the translation model?

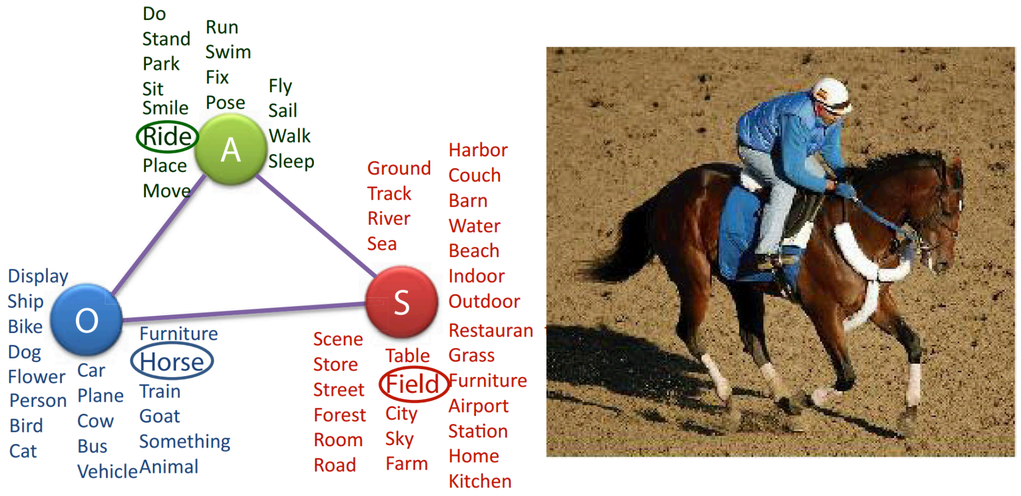
Taxonomy, Methods, Applications and Ideas

Translation: Example-Based Translation

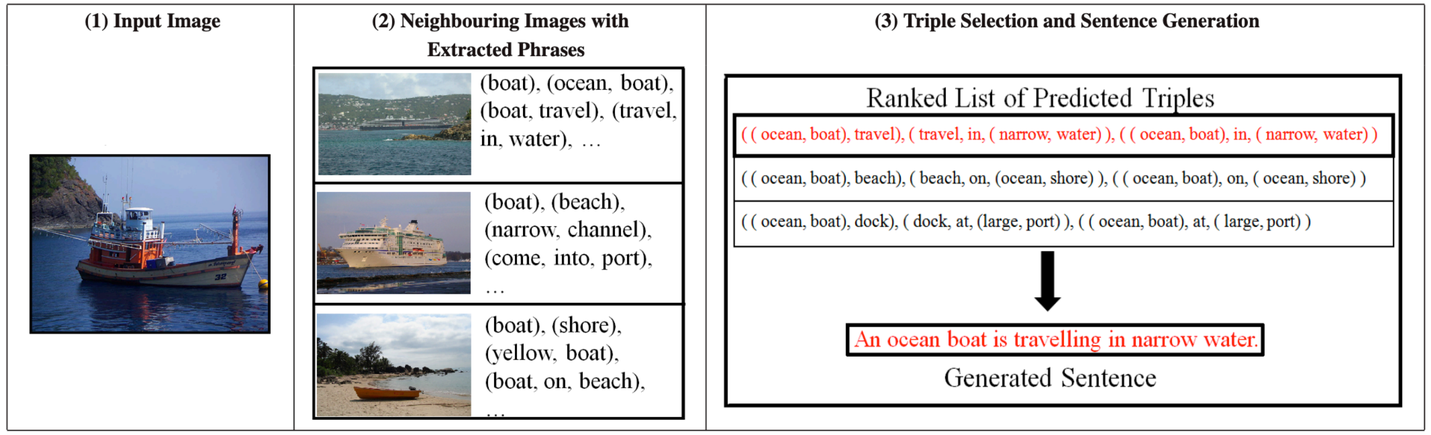
Methods: a data-dictionary approach, kind of retrieve and generation

Retrieval-based models: use just the best matched data as translate result

Combination-based models: retrieve, combine and generation



An example of Retrieval-based model



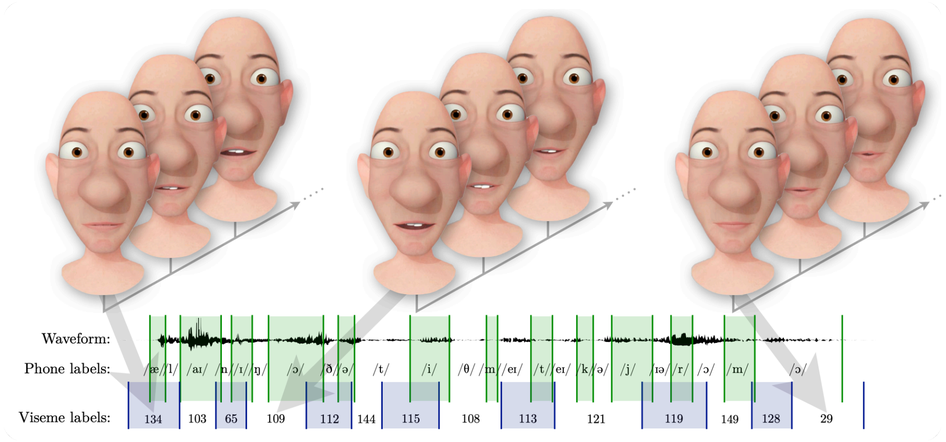
An example of Combination-based model

Taxonomy, Methods, Applications and Ideas

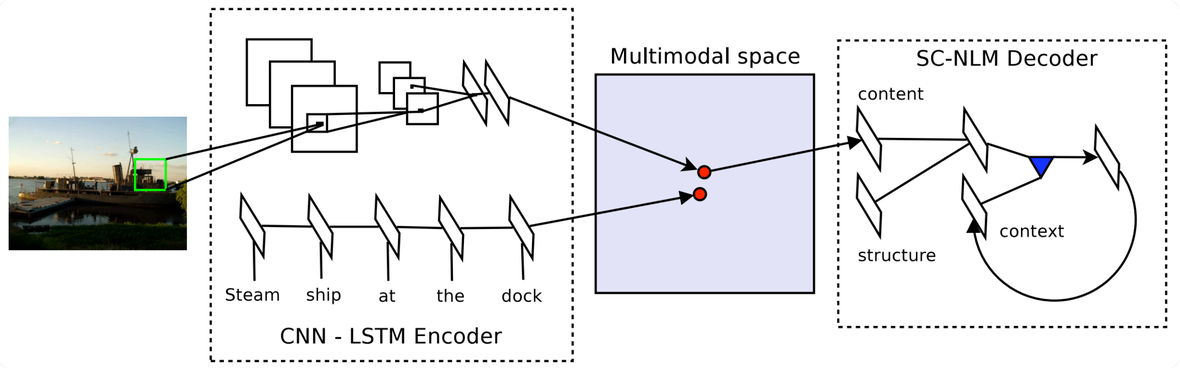
Translation: Generative Approaches

Methods: with a source modality, generate the target sequence or signal

Grammmar-based, Encoder-decoder, and Continuous generation models



An example of Continuous generation model



An example of encoder-decoder model

Taxonomy, Methods, Applications and Ideas

Translation and Alignment: Potential Application

News Video Clip Selection and Combination

Task Definition: Given a text, and plenty of video data, the task is to select the video clip for given context to generate a smooth news video

Motivation: Video editing and

Clipping is a very complicated

Task for journalists.

Method: split -> encode -> rank

-> combine -> smoothen

WWII pilot had a request for his 101st birthday. His nursing home made it happen

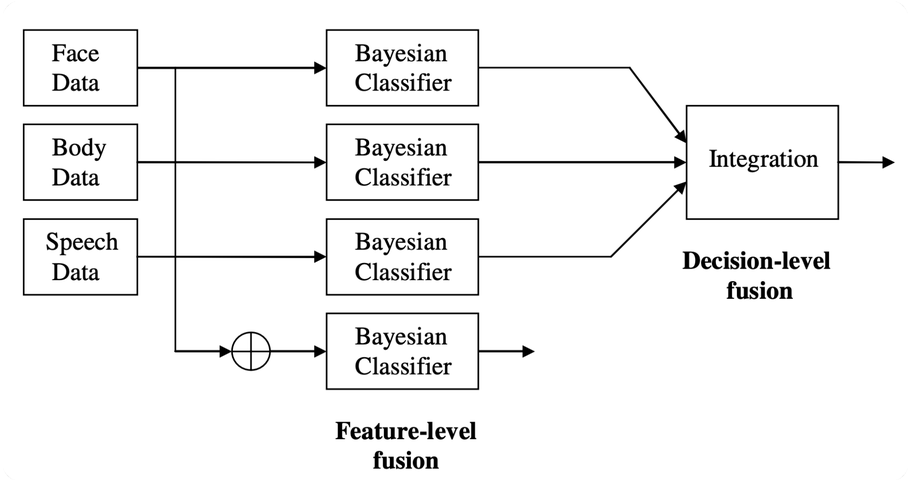
Capt. James McCubbin, a retired US fighter pilot who served during World War II, had a wish for his 101st birthday. His retirement home in Georgia helped deliver the request, and McCubbin was all smiles. Source: HLN

Taxonomy, Methods, Applications and Ideas

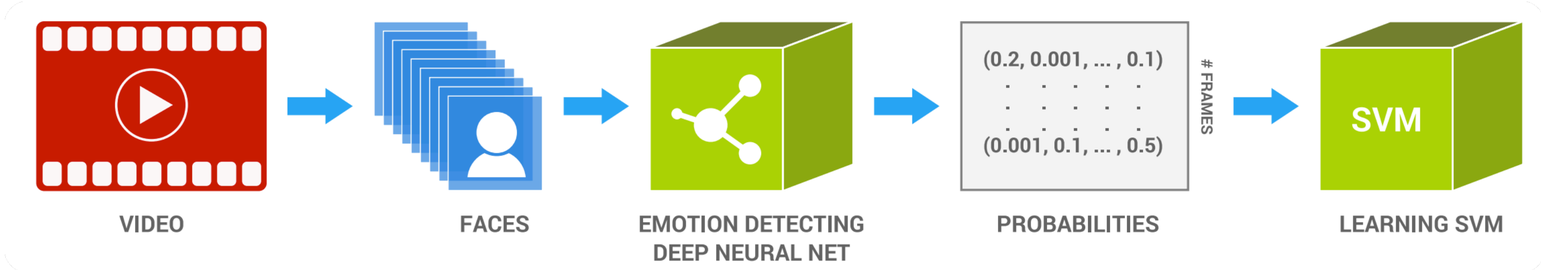
Fusion

Definition: integrating information from multiple modalities

Methods:



An example of early fusion



An example of neural network based fusion model

Taxonomy, Methods, Applications and Ideas

Fusion: Potential Approach

Multimodal Cascade (kind of modal ensemble)

Method idea:

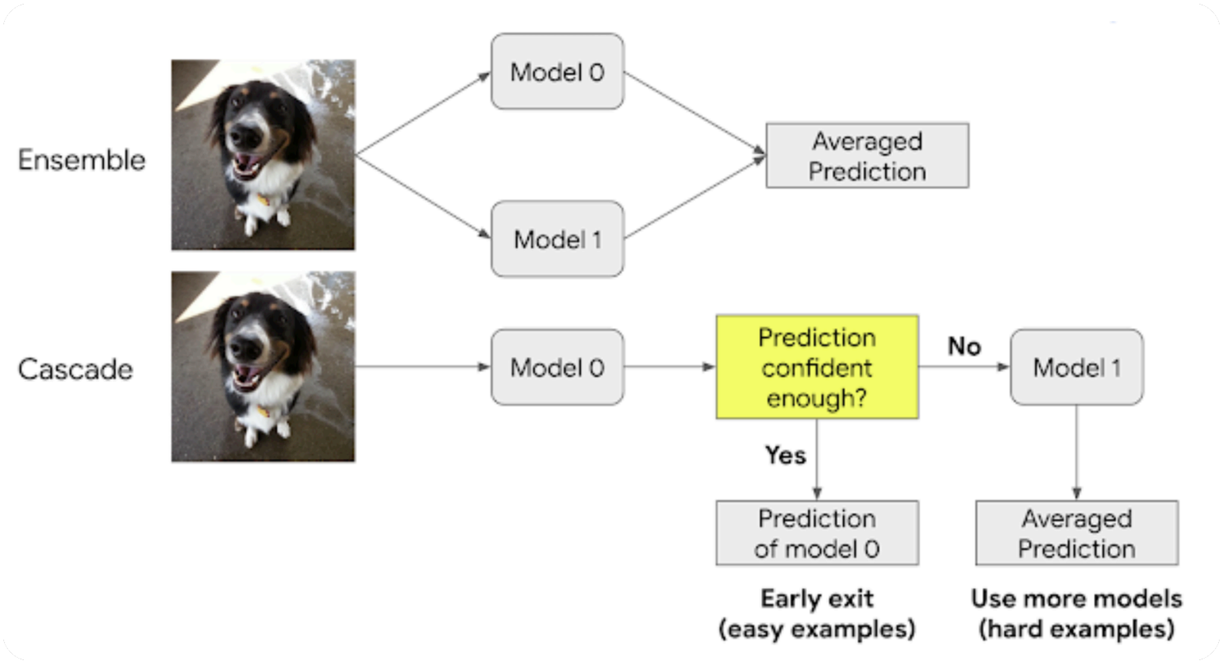
When predicting, if one modality is enough,

why use more modality data?

Motivation:

Low power consumption and latency

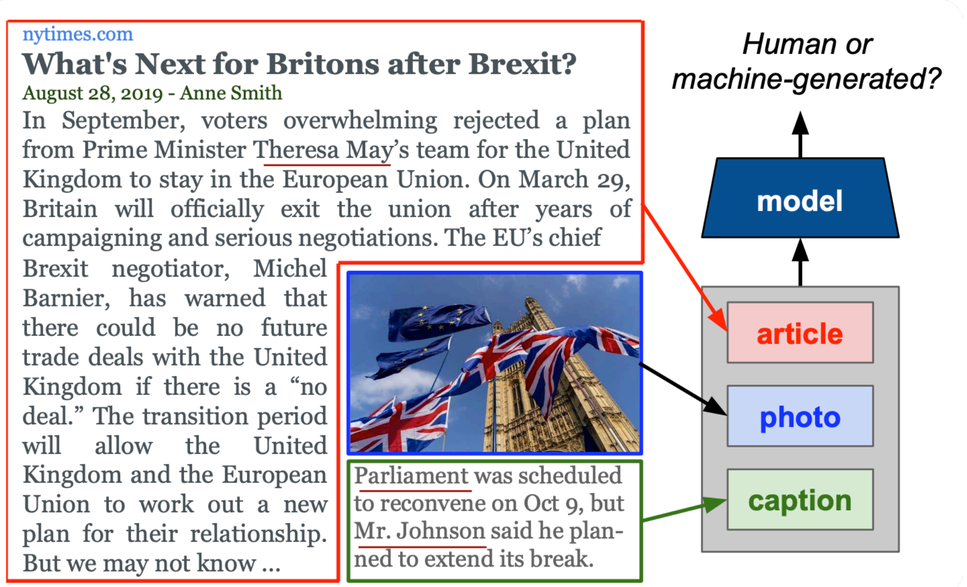
Requirements for downstream applications



An example of model cascade, but how about modal cascade?

Taxonomy, Methods, Applications and Ideas

Fusion: Potential Application



Multimedia Fake news detection

Task Definition: Given a news page with text, videos, images, social context, caption, the task is to predict whether this article is fake or machine-generated

Motivation: The emergence of fake news has seriously affected the quality and number of users of media platforms

Method: get unimodal representation -> combine and fuse -> predict

Taxonomy, Methods, Applications and Ideas

Alignment and Co-Learning

Definition:

Finding relationships and correspondences between modalities, and aiding the resource poor modality by exploiting another rich modality.

Importance:

1. Measurement of the similarity between different modalities (alignment)
2. When training data is not sufficient (co-learning)
3. Zero shot learning: through the help of other modalities

Multimodal Integrated Sparsely Activated Model

Background Introduction — Pathways, idea bases

Oct 28, 2021
5 min read

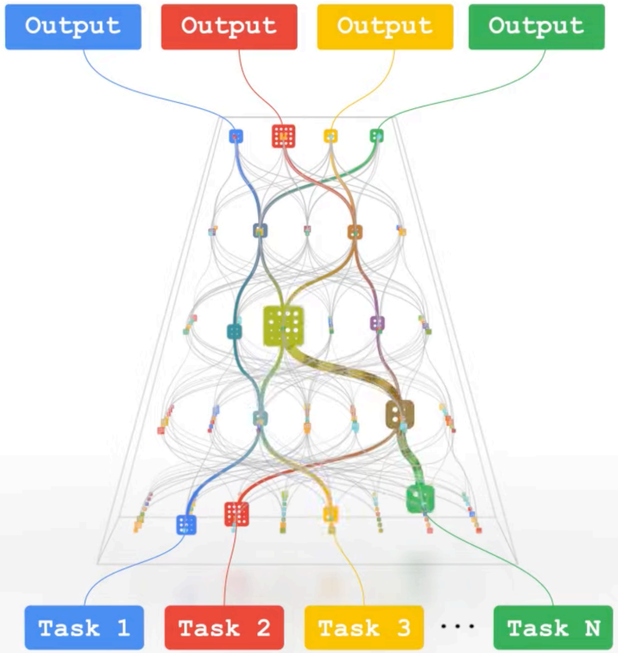
Too often, machine learning systems overspecialize at individual tasks, when they could excel at many. That's why we're building Pathways—a new AI architecture that will handle many tasks at once, learn new tasks quickly and reflect a better understanding of the world.

Jeff Dean
Google Senior Fellow and SVP, Google Research

Share

Advantage:

1. A universal model
2. Low latency and computation cost
3. Multiple senses, understand them simultaneously
4. Energy saving compared to large models



Pathways: A single model that can generalize across millions of tasks.

Multimodal Integrated Sparsely Activated Model

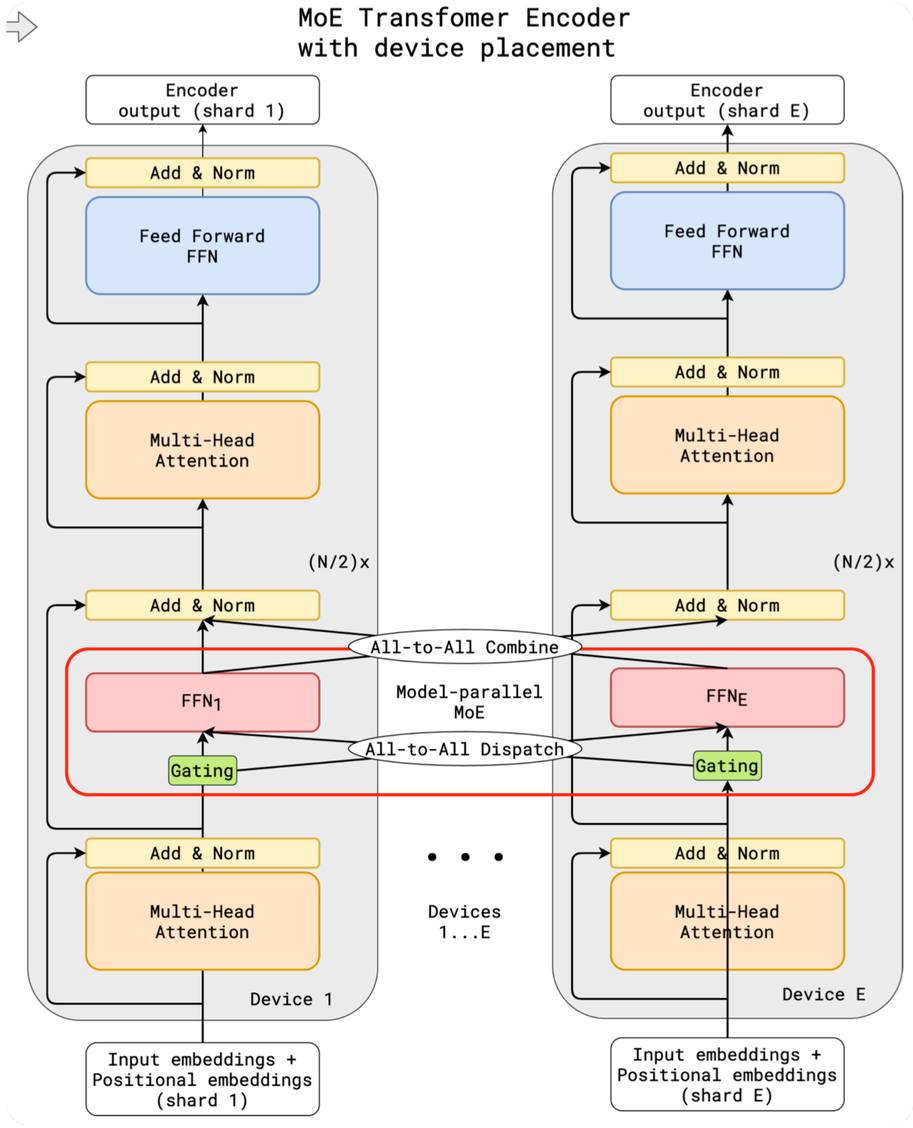
How about multimodal pathway/MoE?

Scenario:

When the model is processing the word “leopard”, the sound of someone saying, or a video of a leopard, the concept is activated simultaneously, which can make prediction less prone to mistakes and biases.

Methods:

Let the experts handle it!



What’s more?

- Negative sample generation for model-based evaluation metrics
- Video subtitles automatic generation
- Example-based model attack for data extraction
- Integrate graph structure like user network for preference prediction
- Event extraction in heterogenous data (like movies, video, image, audio)
- MultiMedia information parser
- Mixture of Multimodal Experts
- ……

Reference

1. Multimodal Machine Learning- A Survey and Taxonomy
2. Distinctive image features from scale-invariant key-points
3. Collective generation of natural image descriptions
4. Long short-term memory
5. HMM-based word alignment in statistical translation
6. Distributed representations of words and phrases and their compositionally
7. A Review and Meta-Analysis of Multimodal Affect Detection Systems
8. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network
9. Deep multimodal learning for Audio-Visual Speech Recognition
10. Learning Grounded Meaning Representations with Auto-encoders
11. Multimodal Learning with Deep Boltzmann Machines
12. Multimodal Deep Learning
13. Extending Long Short-Term Memory for Multi-View Structured Learning
14. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence – Arousal Space
15. eViSE: A deep visual-semantic embedding model
16. Deep Visual-Semantic Hashing for Cross-Modal Retrieval
17. Every picture tells a story: Generating sentences from images
18. Choosing Linguistics over Vision to Describe Images
19. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models
20. Video In Sentences Out
21. Dynamic units of visual speech
22. Expressive visual text-to-speech using active appearance models
23. Modeling Latent Discriminative Dynamic of Multi-Dimensional Affective Signals
24. Emotion recognition through multiple modalities: Face, body gesture, speech
25. EmoNets: Multimodal deep learning approaches for emotion recognition in video,