Data Science

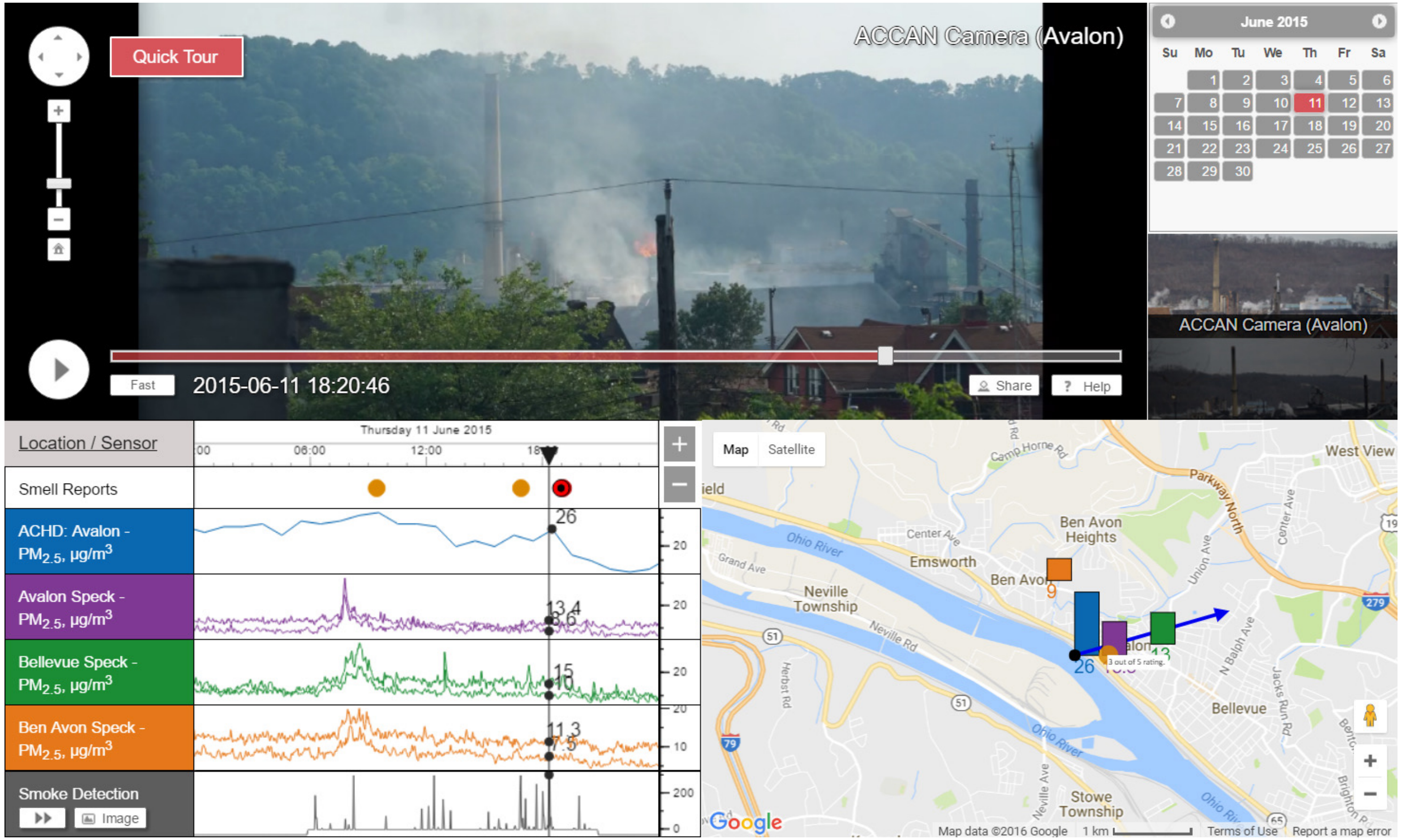
Lecture 11: Multimodal Data Processing

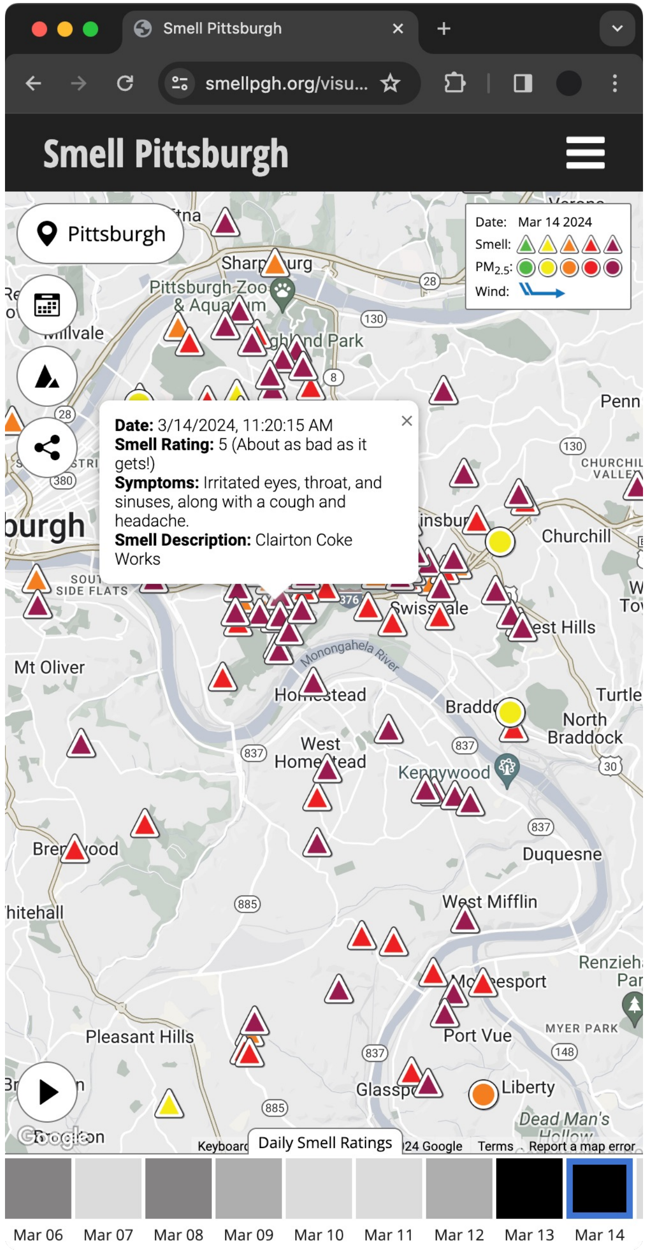
UNIVERSITY OF AMSTERDAM

Lecturer: Yen-Chia Hsu

Date: Mar 2024

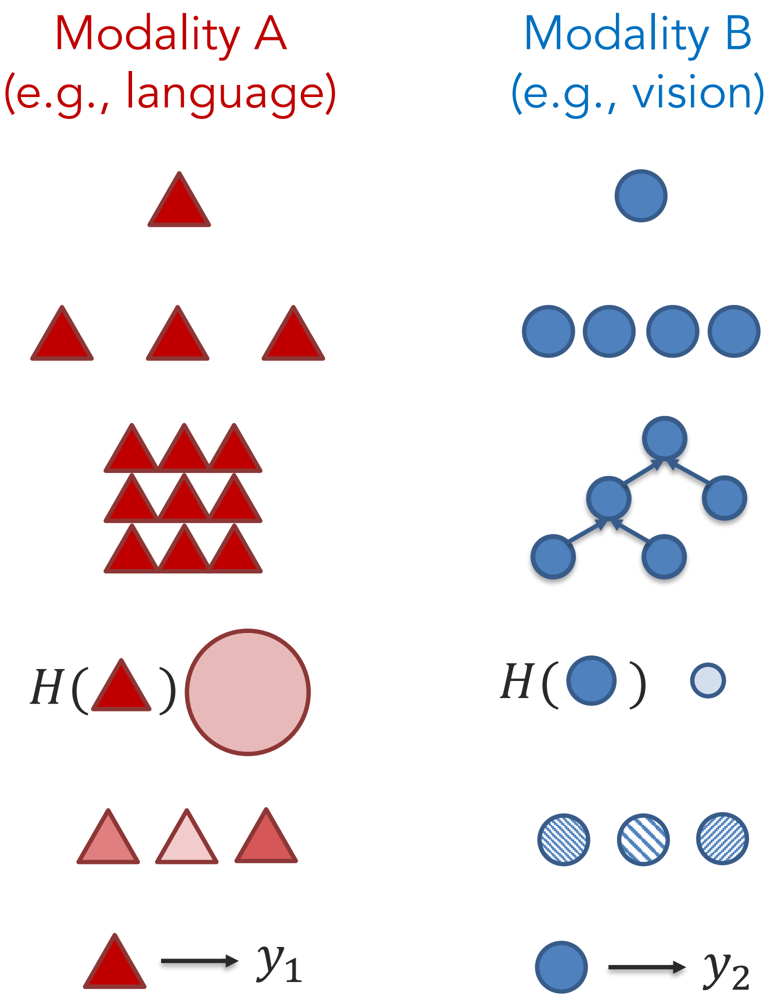
A modality means how a natural phenomenon is perceived or expressed. Multimodal means having multiple modalities.





Different modalities can have different characteristics.

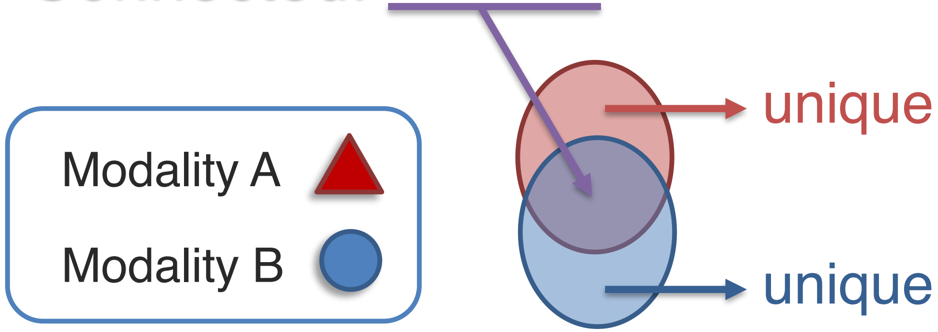
1. **Element representations:**
   Discrete, continuous, granularity
2. **Element distributions:**
   Density, frequency
3. **Structure:**
   Temporal, spatial, latent, explicit
4. **Information:**
   Abstraction, entropy
5. **Noise:**
   Uncertainty, noise, missing data
6. **Relevance:**
   Task, context dependence

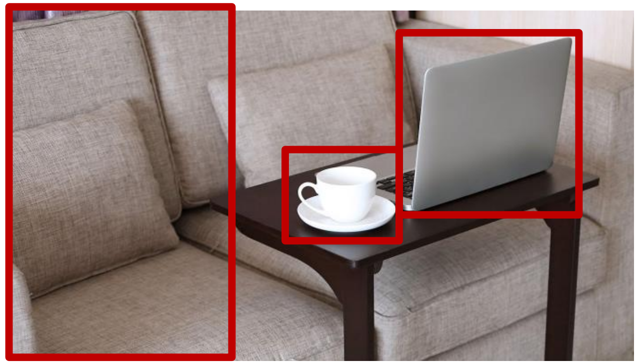


Source: CMU 11-777 MMML Course Lecture 1.1 -- https://cmu-multicomp-lab.github.io/mmml-course/fall2023/schedule/

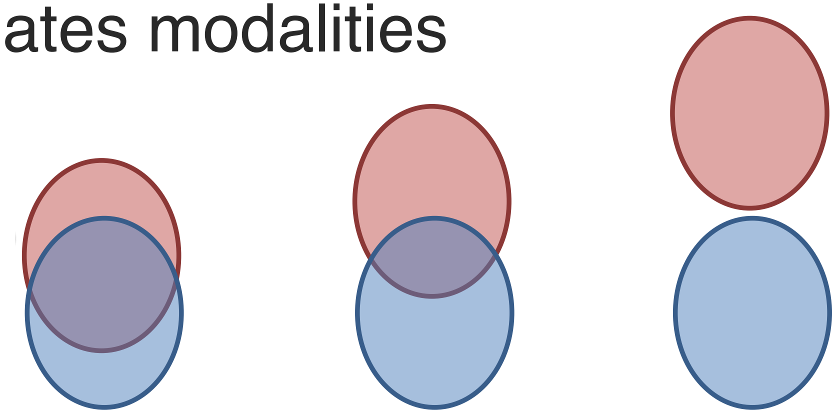
Different modalities can share information with different levels of connections.

Connected: Shared information that relates modalities





A ***teacup*** *on the* ***right*** *of a* ***laptop***
in a ***clean room****.*



strong connection

weak connection

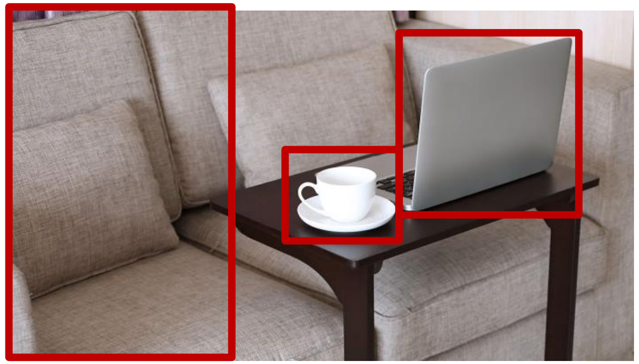
no connection

Source: CMU 11-777 MMML Course Lecture 1.1 -- https://cmu-multicomp-lab.github.io/mmml-course/fall2023/schedule/

The shared information can be connected in different ways.

Association

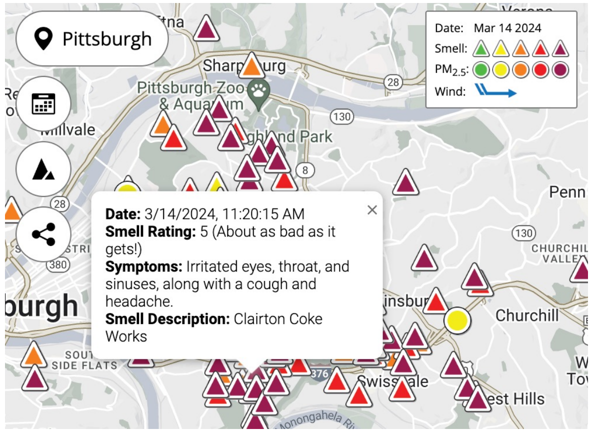
e.g., correlation, co-occurrence



A ***teacup*** *on the* ***right*** *of a* ***laptop*** *in a* ***clean room****.*

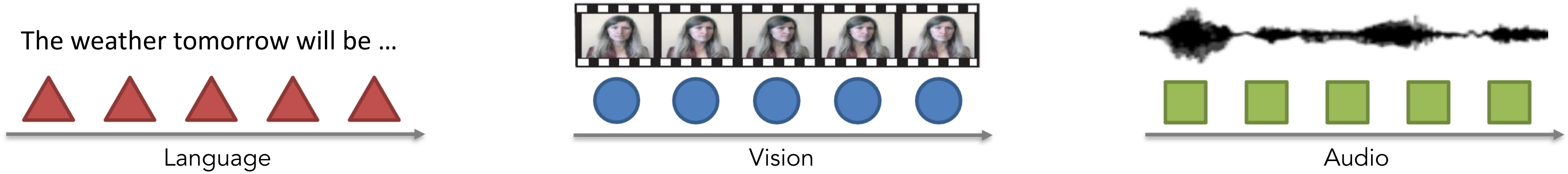
Dependency

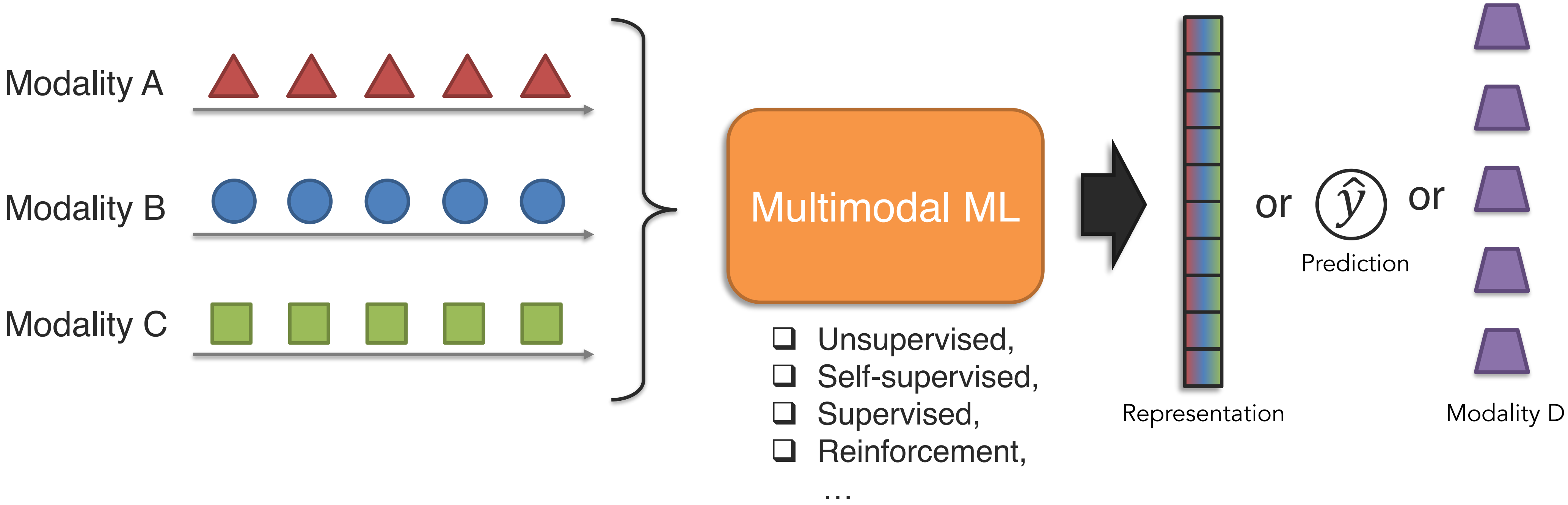
e.g., causal, temporal



Source: CMU 11-777 MMML Course Lecture 1.1 -- https://cmu-multicomp-lab.github.io/mmml-course/fall2023/schedule/

Multiple modalities can exist in different parts of the machine learning pipeline.

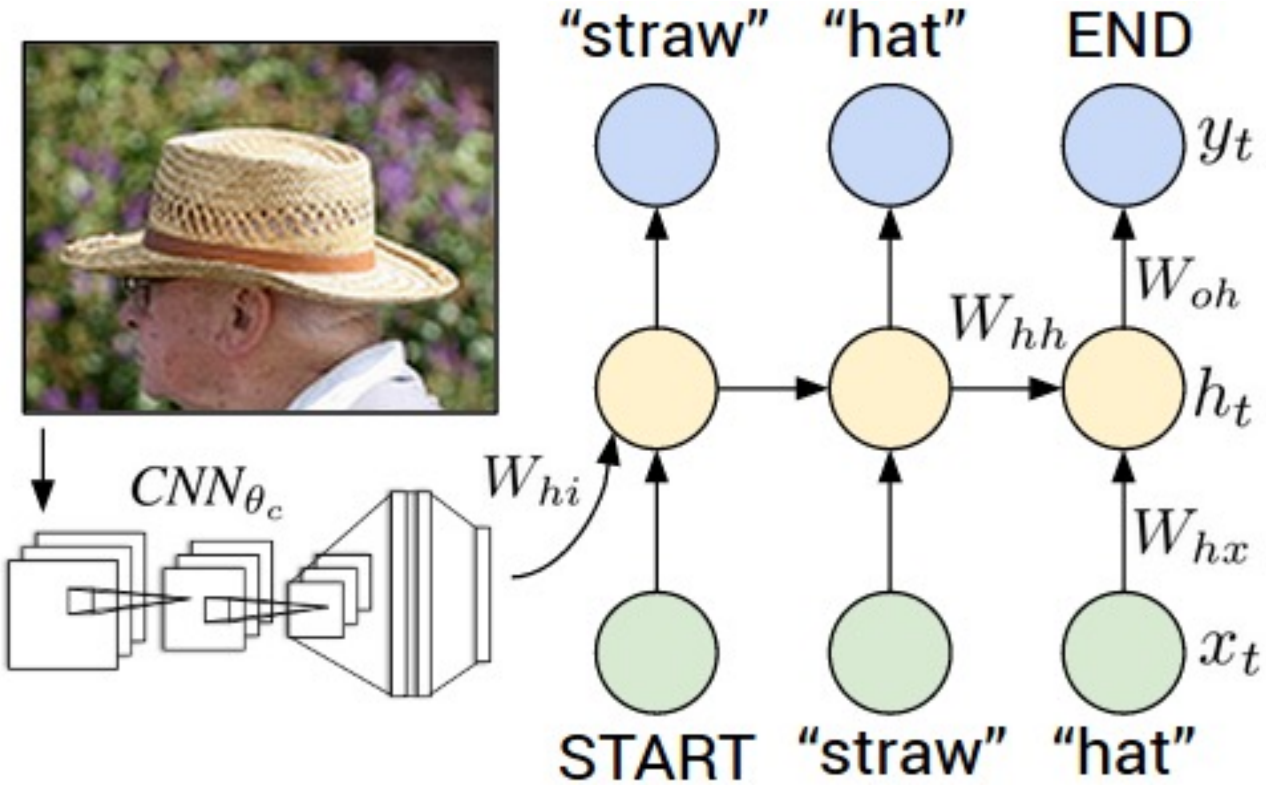




Source: CMU 11-777 MMML Course Lecture 1.1 -- https://cmu-multicomp-lab.github.io/mmml-course/fall2023/schedule/

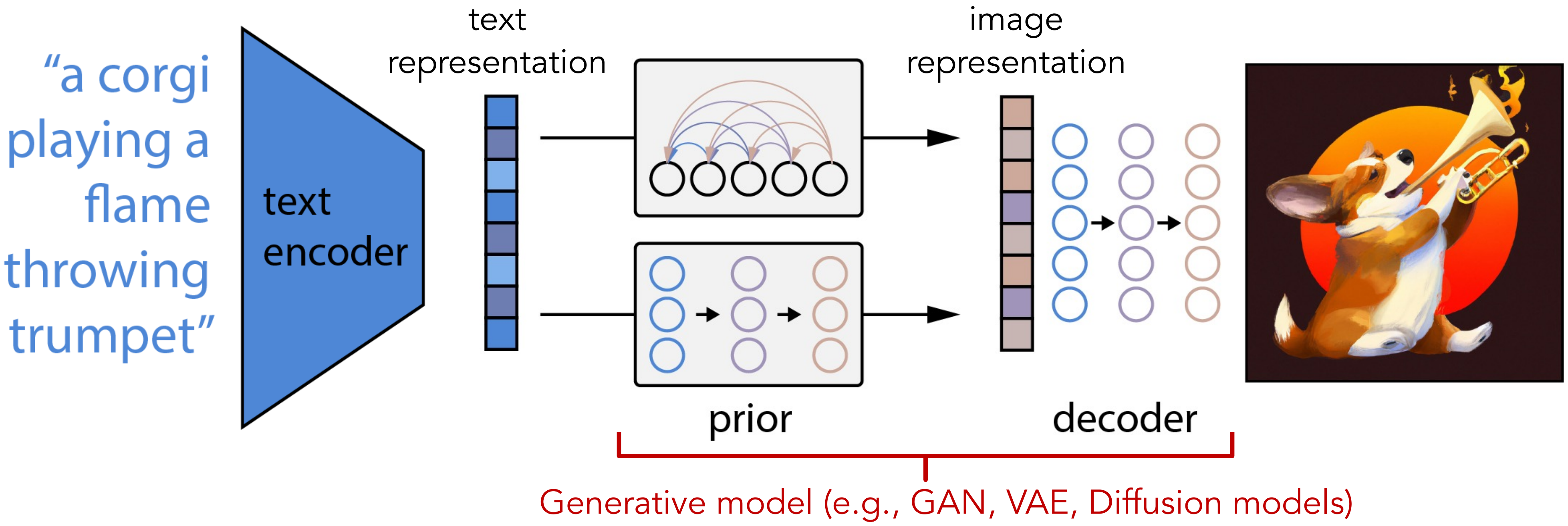

Modality A
Modality B

Image Captioning takes images as input and then outputs sentences that describe the input images (vision→language).

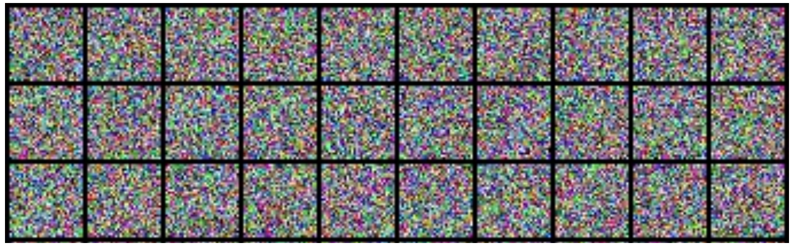


Karpathy, A., & Fei-Fei, L. (2015). [Deep Visual-Semantic Alignments for Generating Image Descriptions](). CVPR.

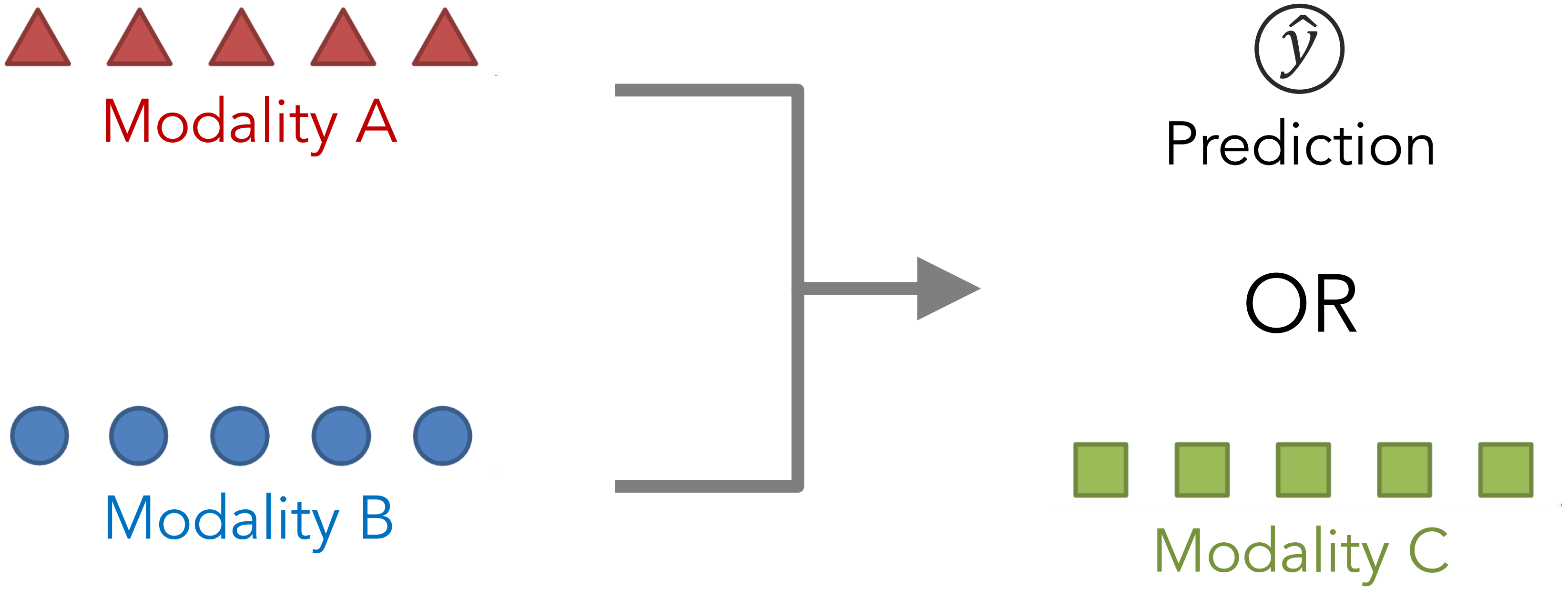
We can also take text as input and then generate images that match the input text (language→vision).



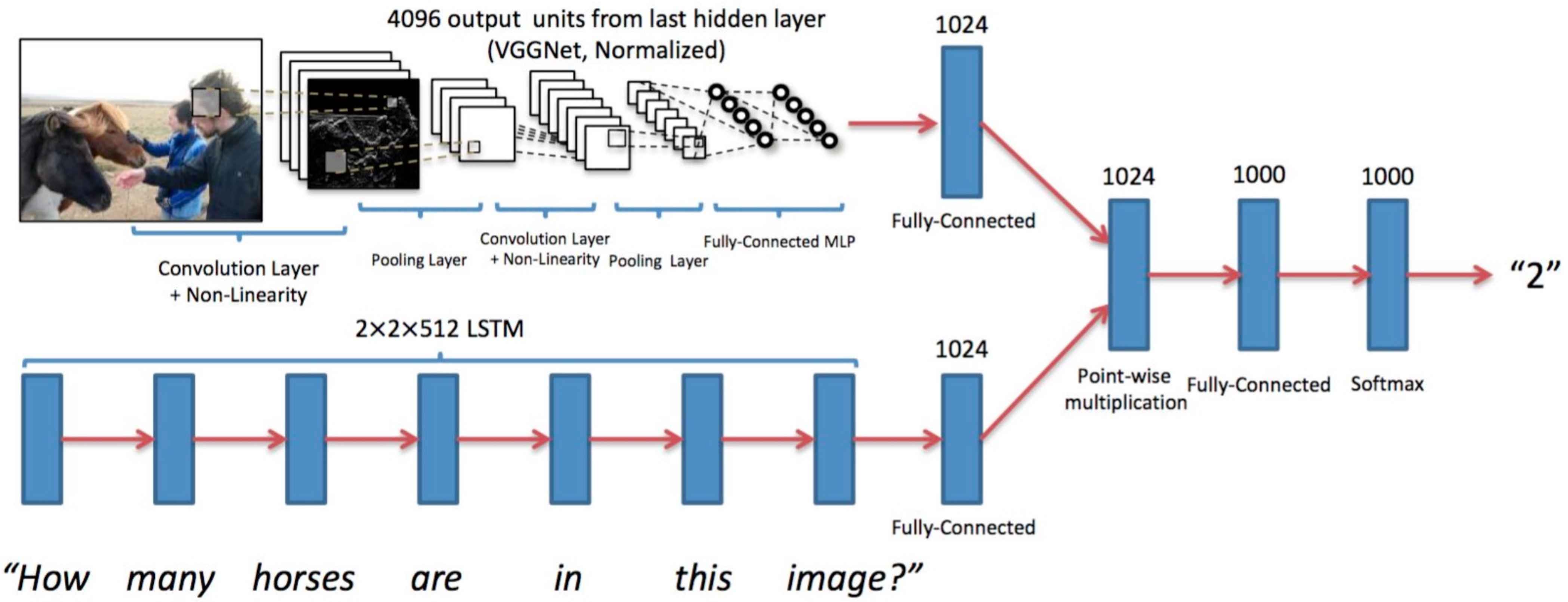
Video source of diffusion models:
https://yang-song.net/blog/2021/score/



Ramesh, A., et al. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv.


Modality A
Modality B
$\hat{y}$
Prediction
OR
Modality C

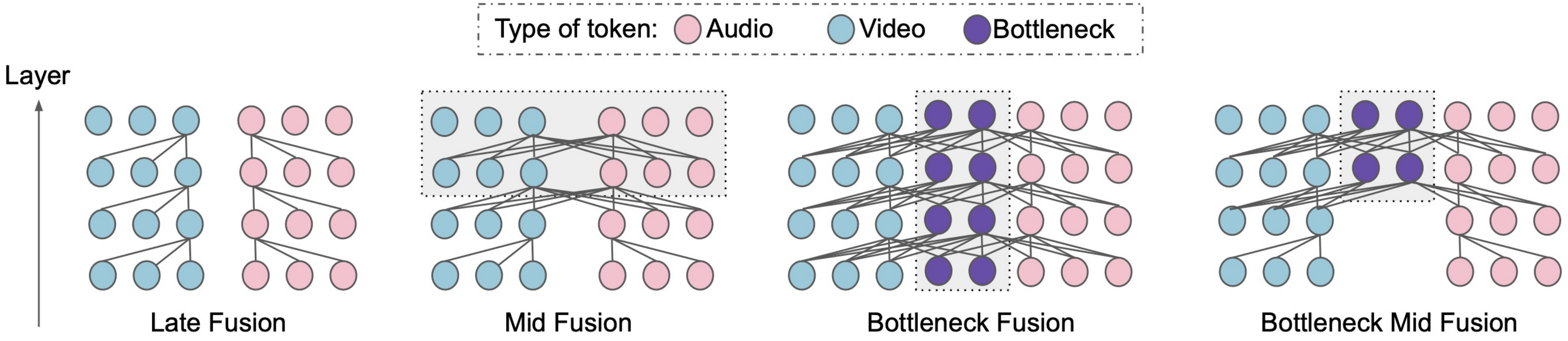
Visual Question Answering takes both images and sentences as input and then outputs a label of text-based multiple-choice answer (vision+language→label).



Zhu, Y., et al. (2016). Visual7W: Grounded Question Answering in Images. CVPR. (Figure is from Stanford cs231n course lecture 8)

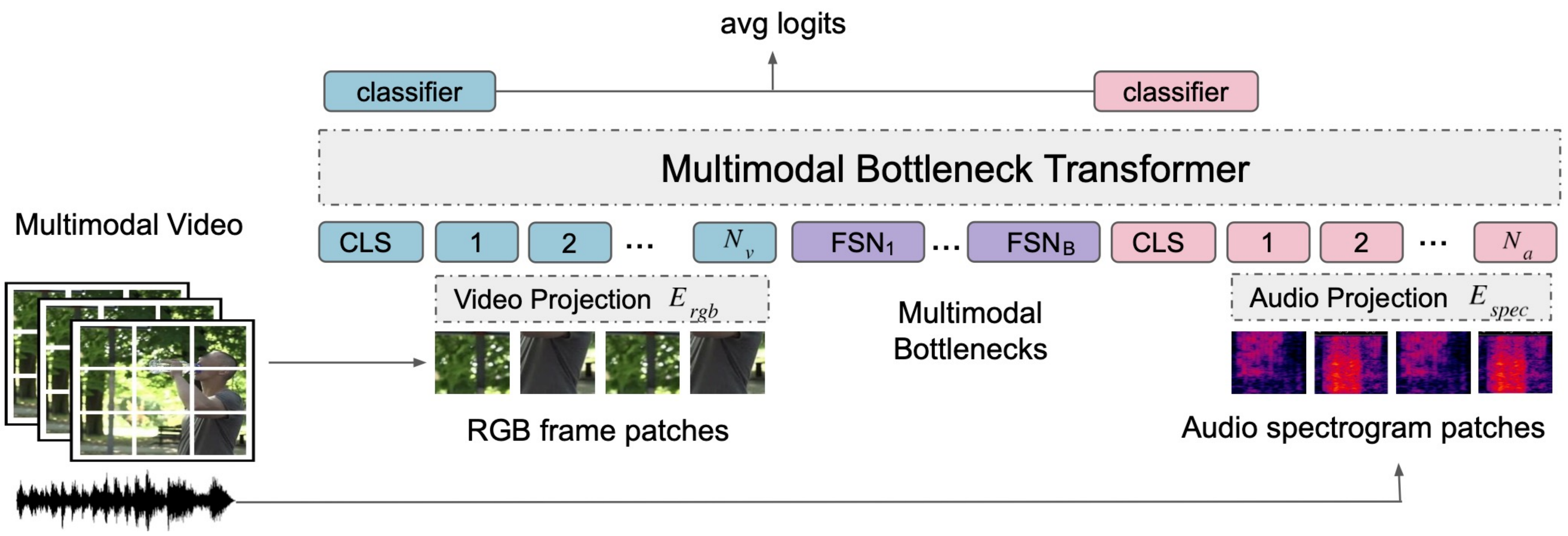
When the input has multiple modalities, we can fuse the modalities or explicitly learn their connections in the model architecture.

Source: CMU 11-777 MMML Course Lecture 1.1 -- https://cmu-multicomp-lab.github.io/mmml-course/fall2023/schedule/



Nagrani, A., et al. (2021). [Attention Bottlenecks for Multimodal Fusion](). NeurIPS

Video Classification can use both video and audio signals to predict output categories (vision+audio→labels).



Below are image tokens (i.e., image embedding)

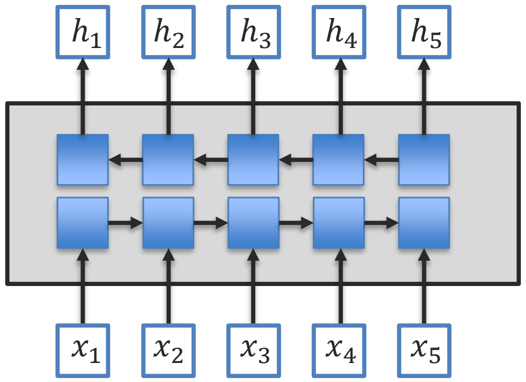
CLS 1 2 ... N_v

Below are audio tokens (i.e., audio embedding)

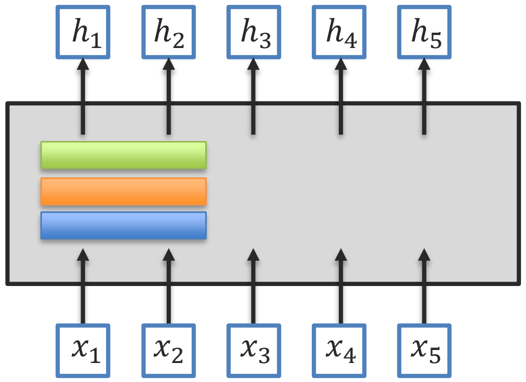
CLS 1 2 ... N_a

Nagrani, A., et al. (2021). [Attention Bottlenecks for Multimodal Fusion](). NeurIPS

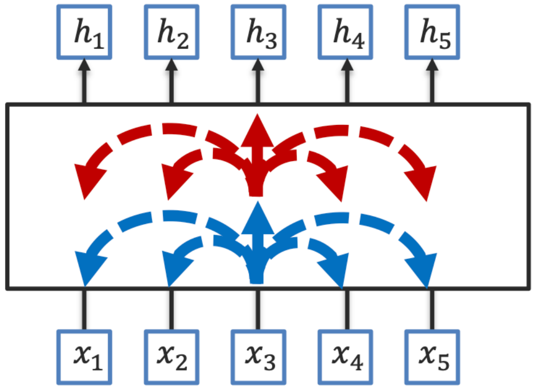
Transformers use self-attention, which is a way of encoding sequences to tells how much attention each input should pay attention to the other inputs (including itself).



Bidirectional RNN



Convolution



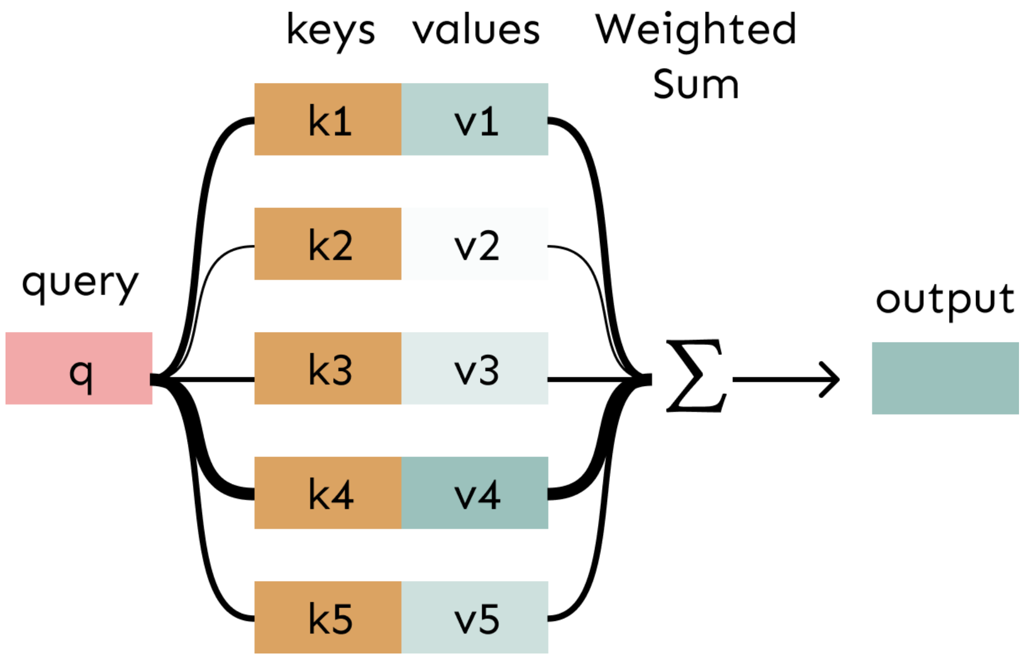
Self-attention

Source: CMU 11-777 MMML Course Lecture 4.2 -- https://cmu-multicomp-lab.github.io/mmml-course/fall2023/schedule/

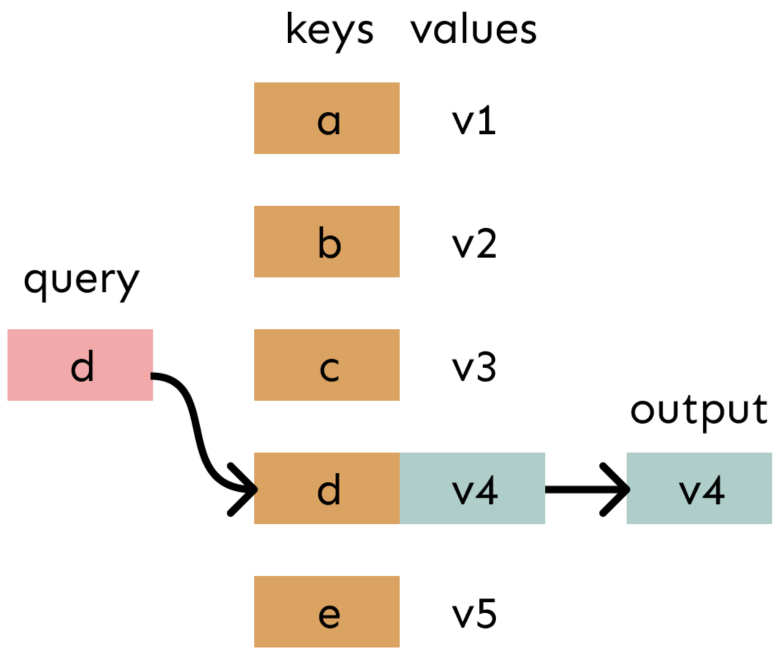
Attention is *weighted* averaging, which lets you do lookups!

Attention is just a **weighted** average – this is very powerful if the weights are learned!

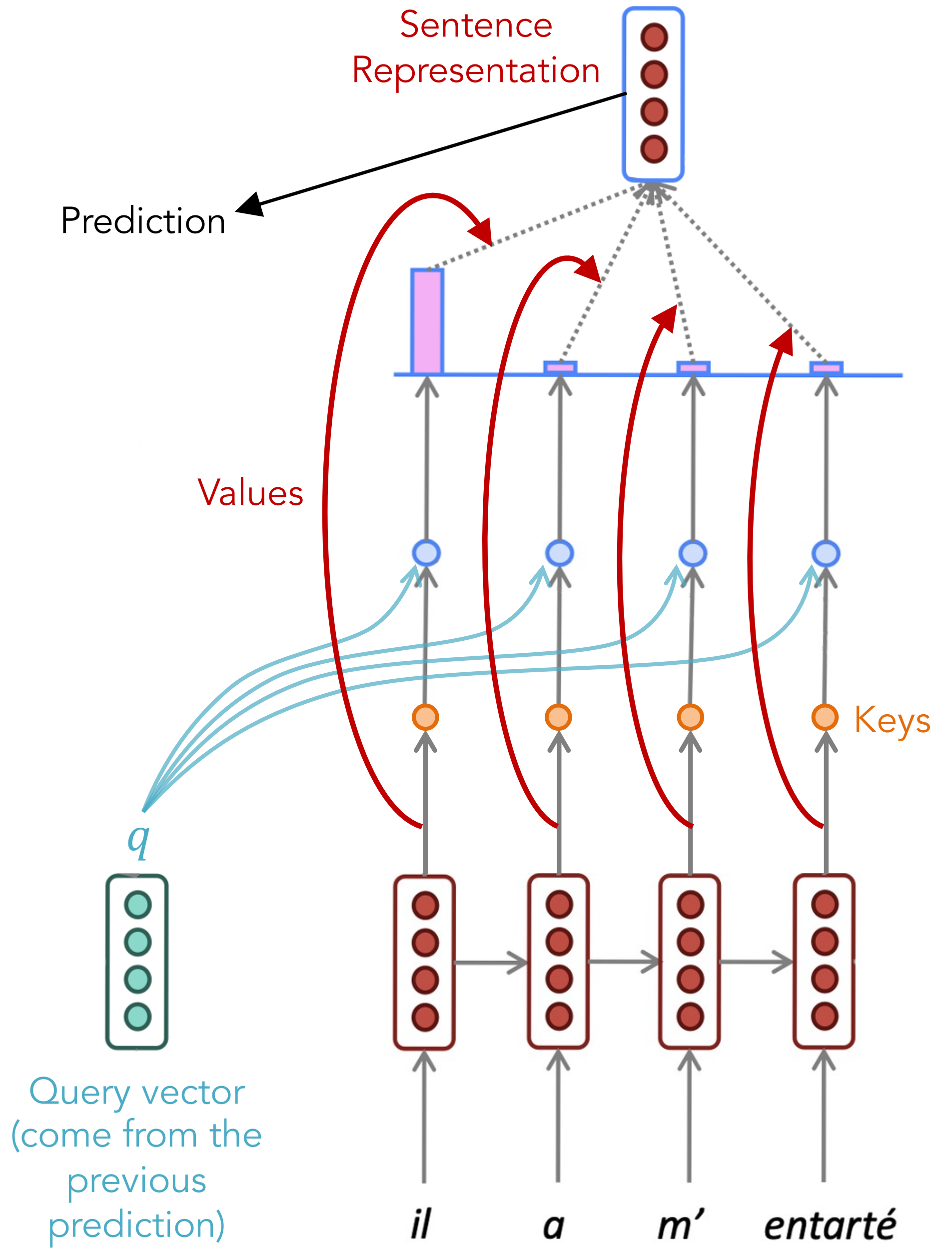
In **attention**, the query matches all keys *softly*, to a weight between 0 and 1. The keys' values are multiplied by the weights and summed.



In a **lookup table**, we have a table of keys that map to values. The query matches one of the keys, returning its value.



The figure is from the Stanford CS224N course -- https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture08-transformers.pdf



Step 5: Compute attention-weighted sum of encoder output:

- $\sum_{t=1}^{T} a_t h_t$

Step 4: Compute the attention distribution using softmax:

- $[a_1 \ a_2 \ \ldots \ a_T] = \text{softmax}([e_1 \ e_2 \ \ldots \ e_T])$

Step 3: Compute attention scores (dot product similarity):

- $e_t = q^T u_t$ — q is trainable

Step 2: Transform encoder outputs (dimension reduction):

- $u_t = \tanh(W h_t)$ — W is trainable

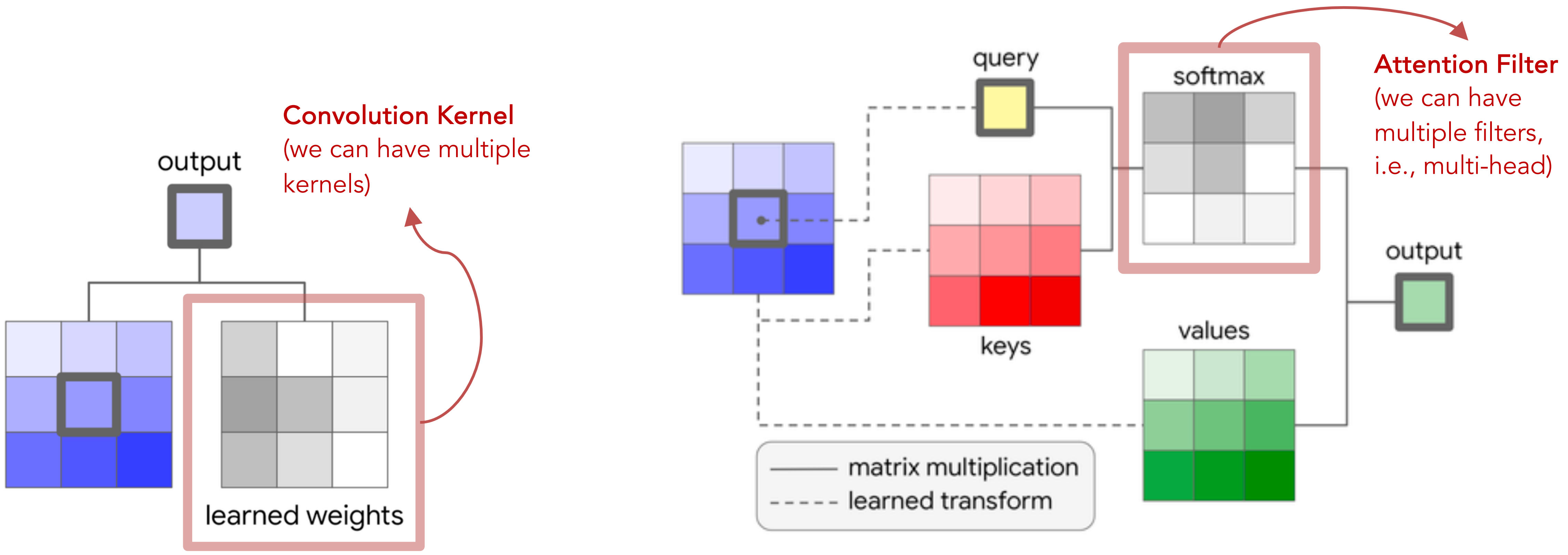
Step 1: Get the encoder output values (from the RNN):

- h_t

There are many ways of doing step 2 and 3

Source -- https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture07-final-project.pdf

Convolution layers use fixed weights (kernels) to filter information. Self-attention layers dynamically compute attention filters to show how well a pixel matches its neighbors.

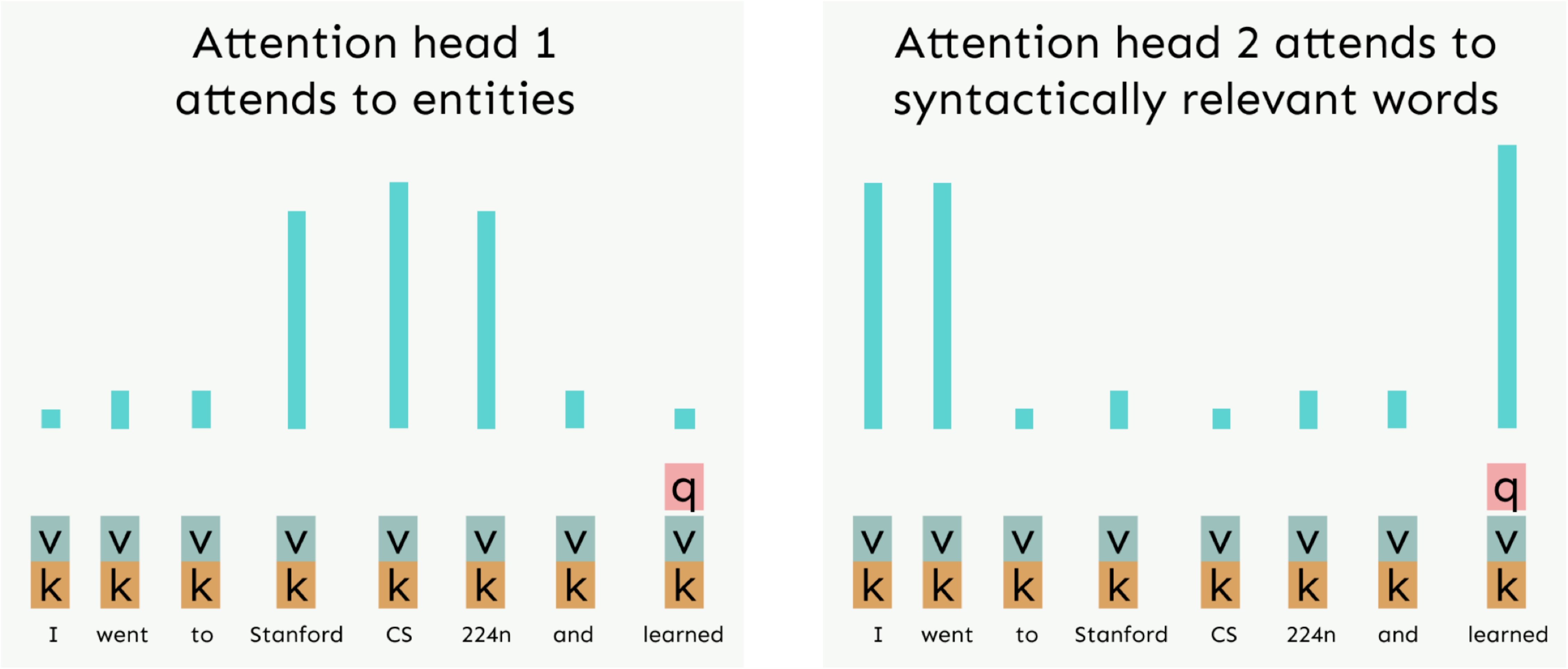


Convolution

Self-attention

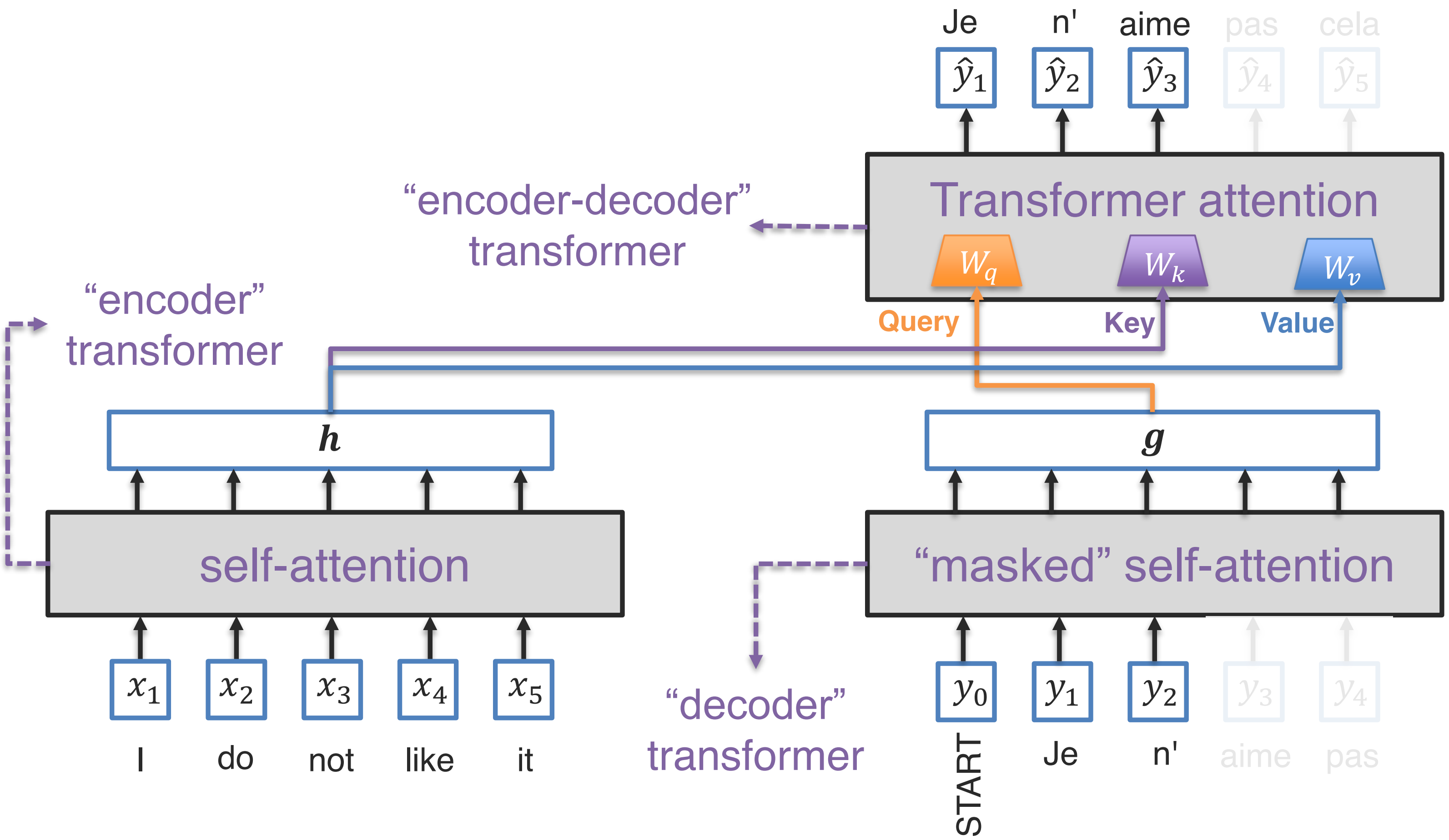
Ramachandran, P., et al. (2019). [Stand-Alone Self-Attention in Vision Models](). NeurIPS.

Transformers use multi-head attention to look at different aspects of the inputs.



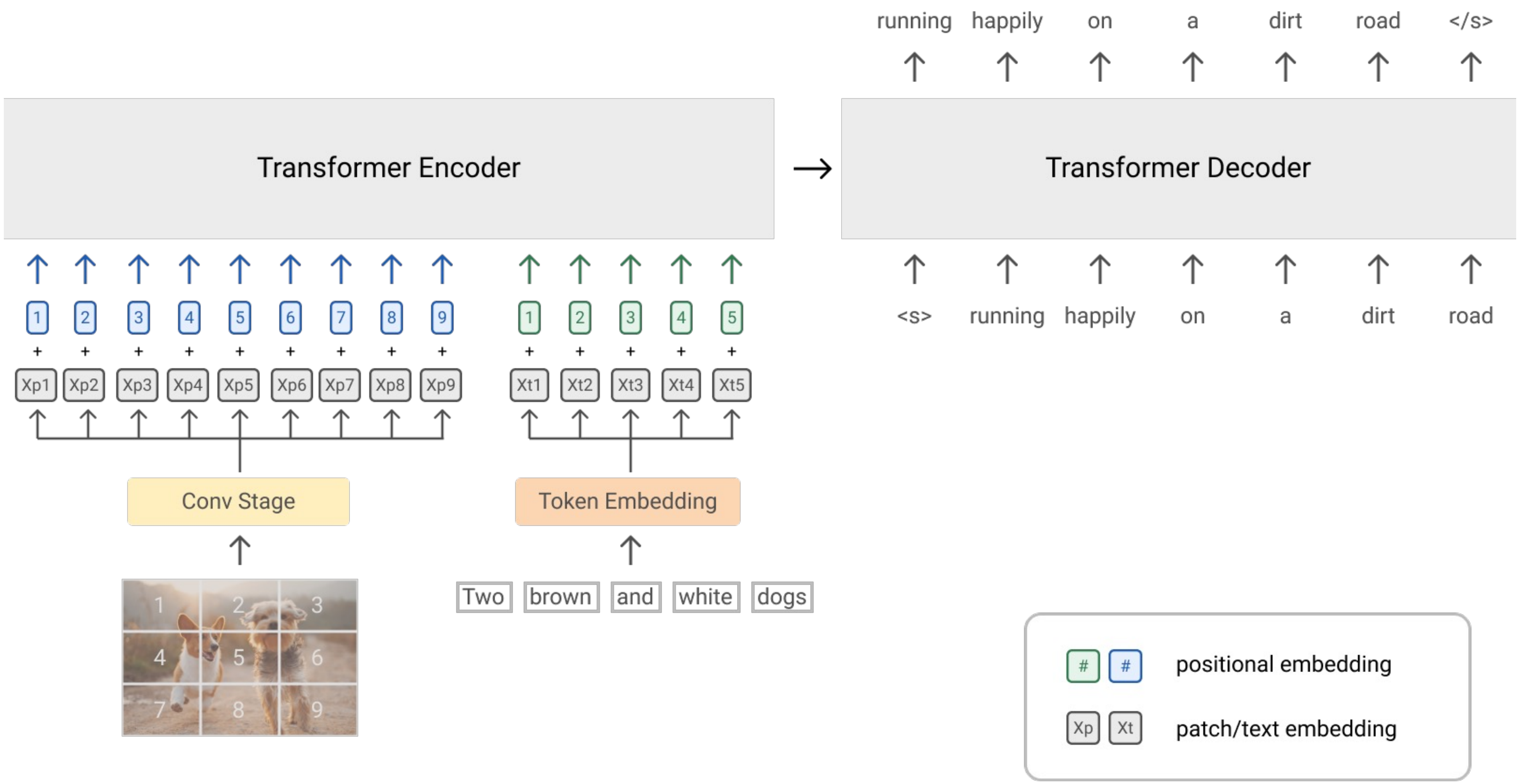
The figure is from the Stanford CS224N course -- https://web.stanford.edu/class/cs224n/slides/cs224n-2024-lecture08-transformers.pdf

Transformers are connected by two self-attention blocks (one for encoder, one for decoder) and an encoder-decoder attention block (similar to the original attention).

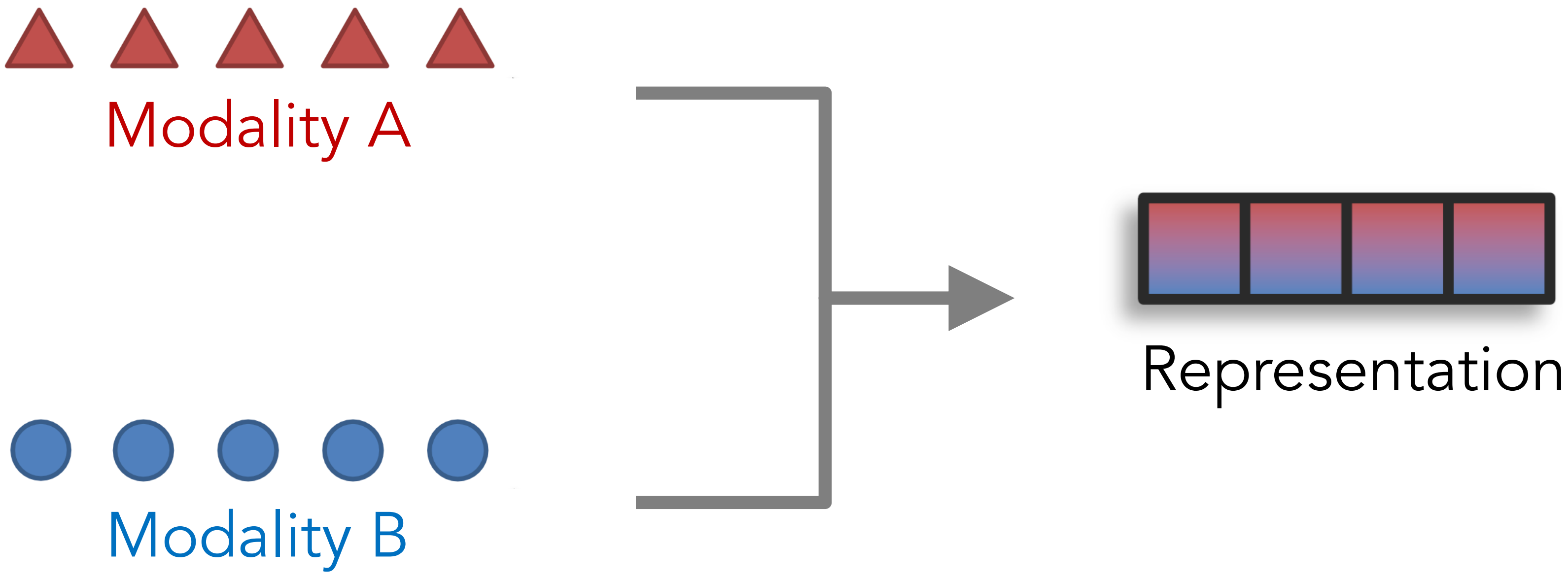


Source: CMU 11-777 MMML Course Lecture 4.2 -- https://cmu-multicomp-lab.github.io/mmml-course/fall2023/schedule/

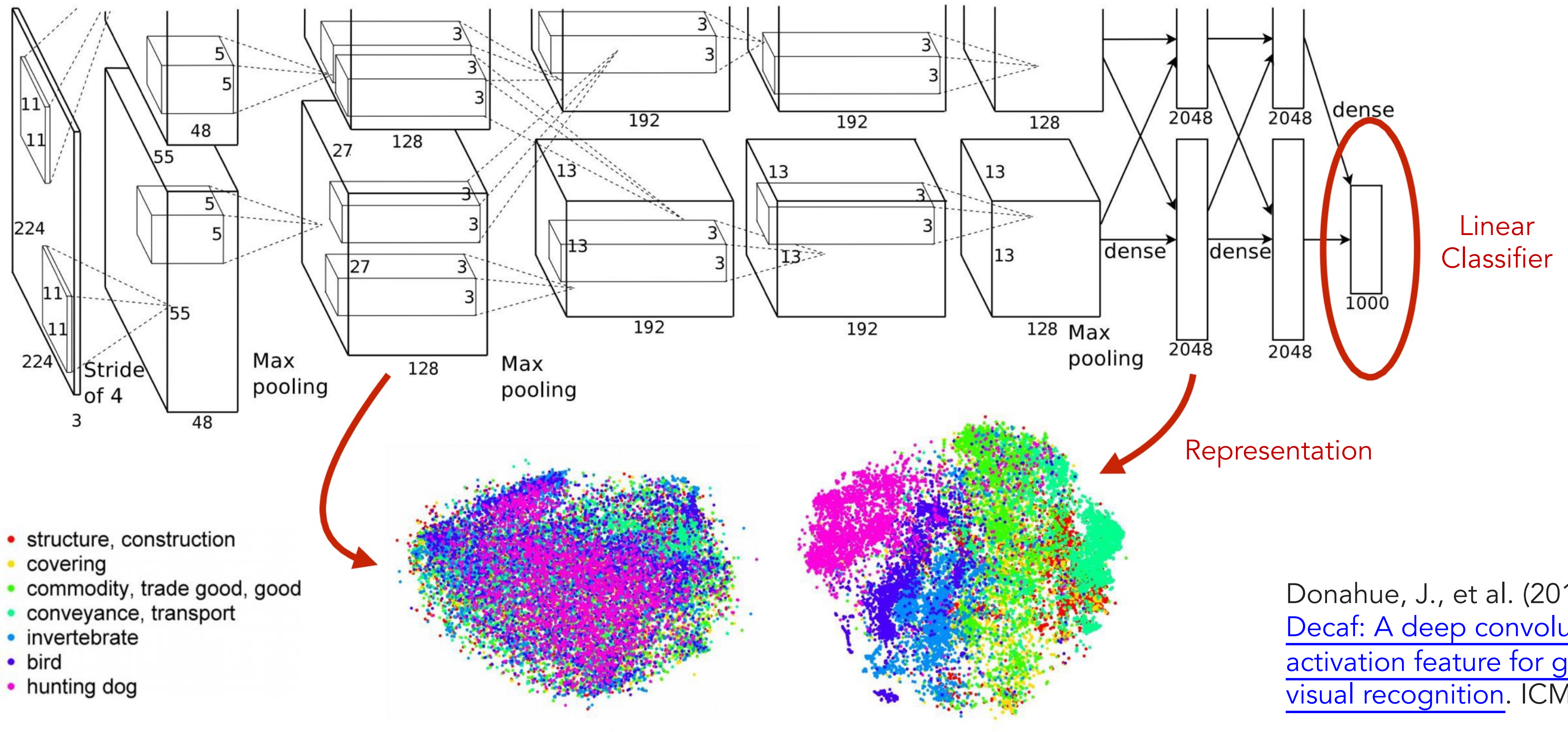
Transformers can work for multiple vision-language tasks (vision+language→language).



Wang, Z., et al. (2021). SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. ICLR.


Modality A
Modality B
Representation

Instead of completing the task directly, we can also think about how to learn a good representation (i.e., embedding) so that a linear classifier can separate the data easily.

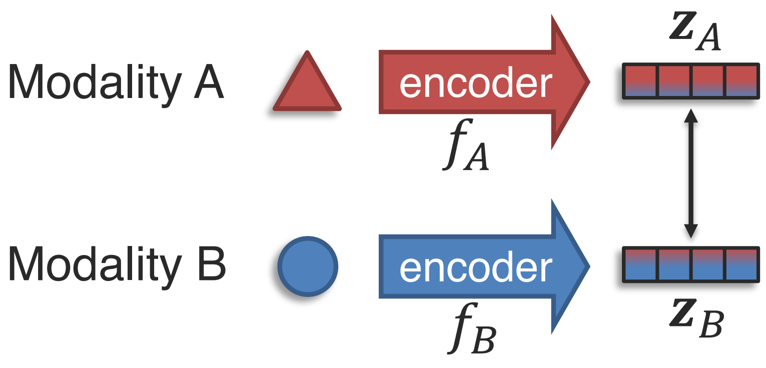


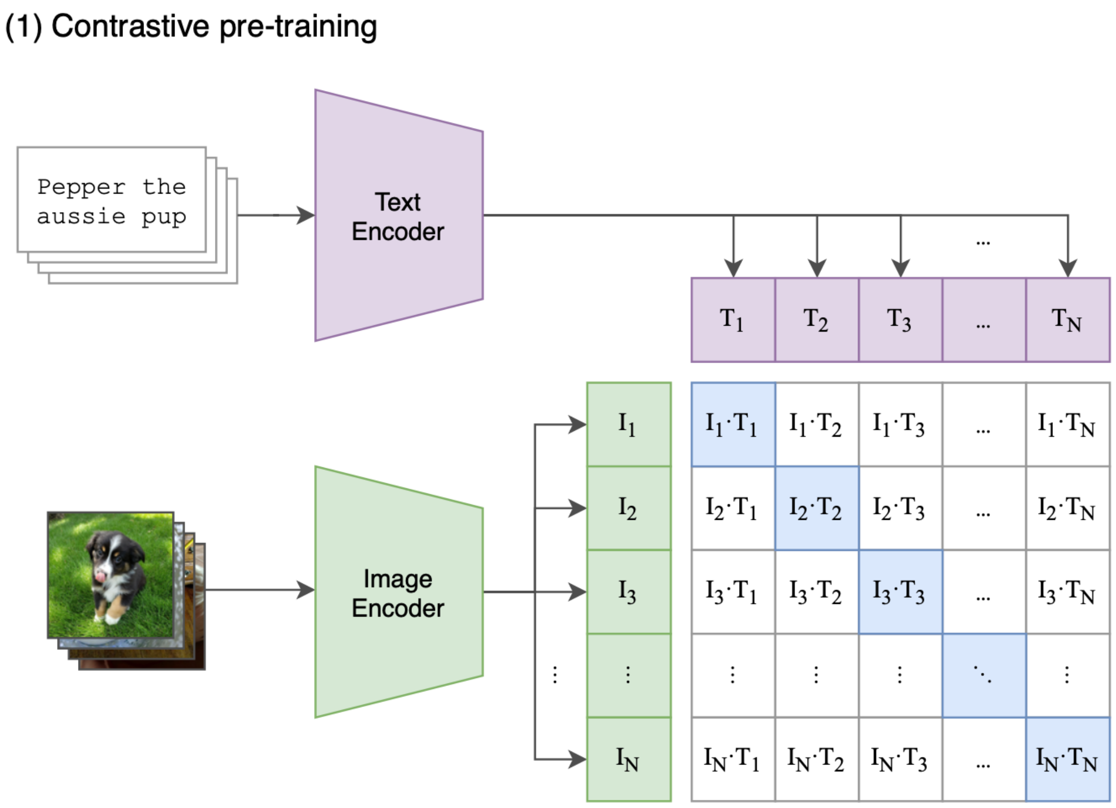
(c) DeCAF_1 (d) DeCAF_6

Donahue, J., et al. (2014). [Decaf: A deep convolutional activation feature for generic visual recognition](). ICML.

Source: Cornell CS5670 Computer Vision Course -- https://www.cs.cornell.edu/courses/cs5670/2022sp/lectures/lec21_cnns_for_web.pdf

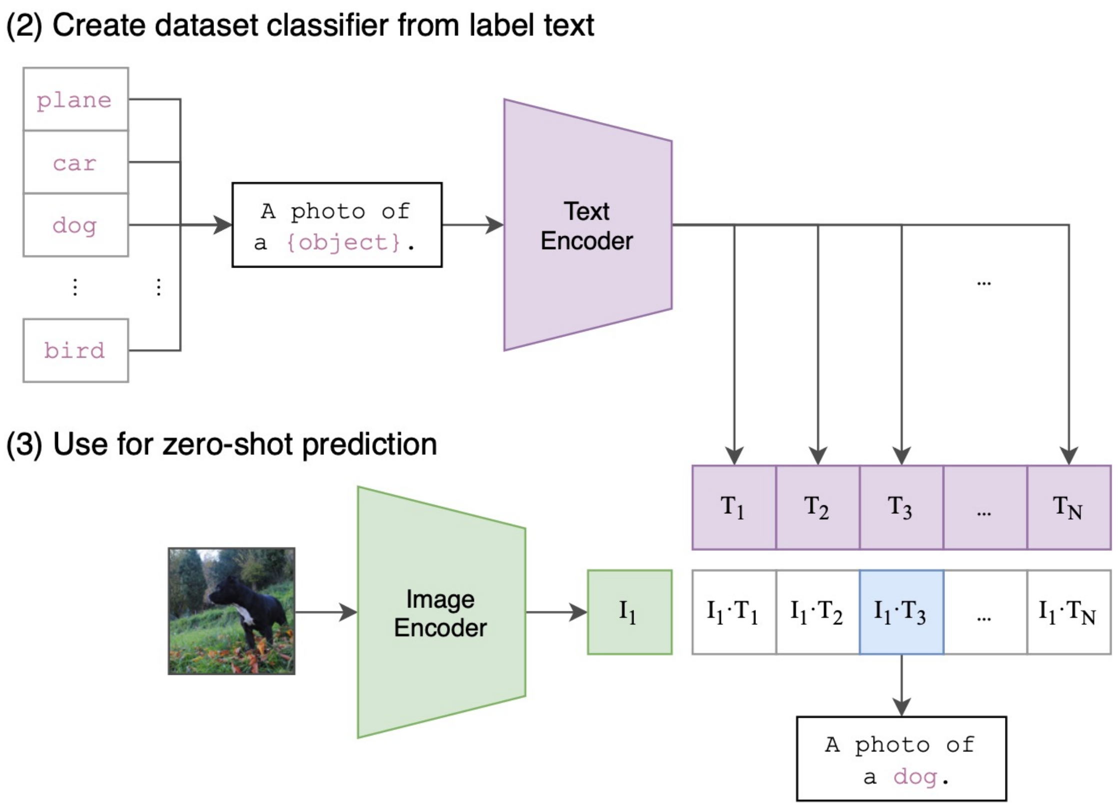
The CLIP model learns a joint text-image representation using a large number of image and text pairs (vision+language→representation).





Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML.

We can use the learned CLIP embedding to perform zero-shot prediction by taking the label with the largest similarity score between the label text and the image.

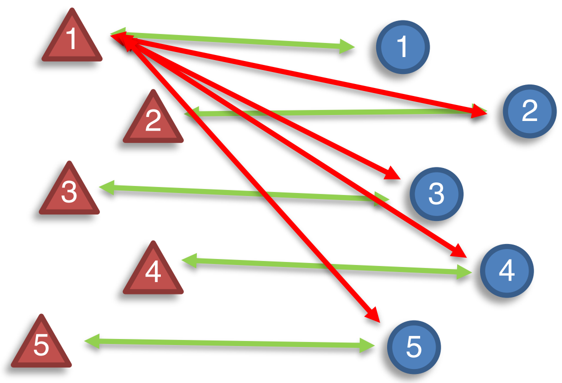


Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML.

Contrastive Learning brings positive pairs closer and pushes negative pairs far apart.

Paired data: {▲, ●}

(e.g., images and text descriptions)



Simple contrastive loss:

$$\max\{0, \alpha + sim(\boldsymbol{z}_A, \boldsymbol{z}_B^+) - sim(\boldsymbol{z}_A, \boldsymbol{z}_B^-)\}$$

positive pairs

negative pair

Similarity functions are often cosine similarity

Popular contrastive loss: InfoNCE

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}{\sum_{j=1}^{N} \mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

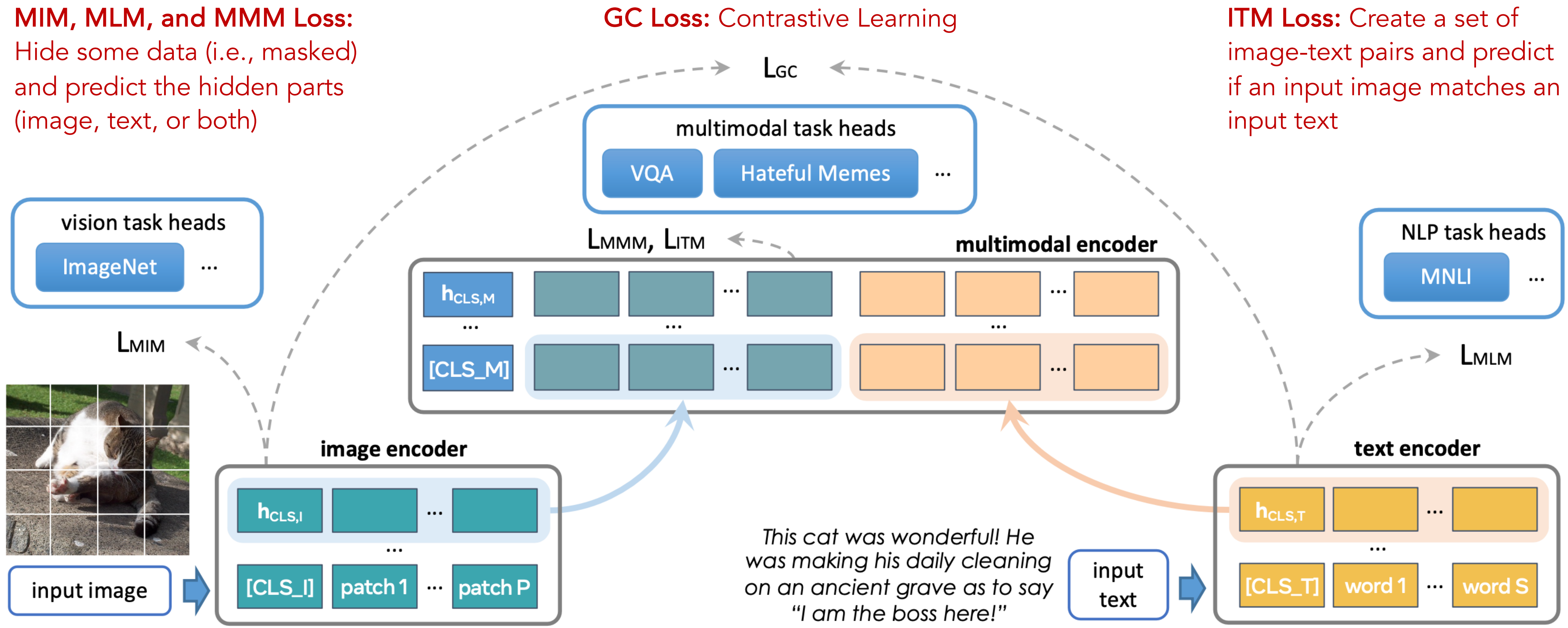
positive pairs

Similarity function can be cosine similarity

negative pairs and positive pairs

Source: CMU 11-777 MMML Course Lecture 3.2 -- https://cmu-multicomp-lab.github.io/mmml-course/fall2023/schedule/

One active research area is foundation models, which work for both unimodal (e.g., image/text classification) and multimodal tasks (e.g., visual question answering).



Singh, A., et al. (2022). FLAVA: A Foundational Language And Vision Alignment Model. CVPR.

Take-Away Messages

- Multimodal means having multiple modalities that represent multiple natural phenomena.
- Multiple modalities can exist in different parts of the machine learning pipeline.
- We can fuse the modalities or explicitly learn their connections in the model architecture.
- Self-attention is a way of encoding sequences to tells how much attention each input should pay attention to the other inputs (including itself).
- We can also think about how to learn a good representation (i.e., embedding) so that a linear classifier can separate the data easily.
- Contrastive Learning brings positive pairs closer and pushes negative pairs far apart.

Questions?