Deep Learning for Computer Vision

Fall 2022

https://cool.ntu.edu.tw/courses/189345 (NTU COOL)

http://vllab.ee.ntu.edu.tw/dlcv.html (Public website)

Yu-Chiang Frank Wang 王鈺強, Professor

Dept. Electrical Engineering, National Taiwan University

2022/11/1

What to Cover Today…

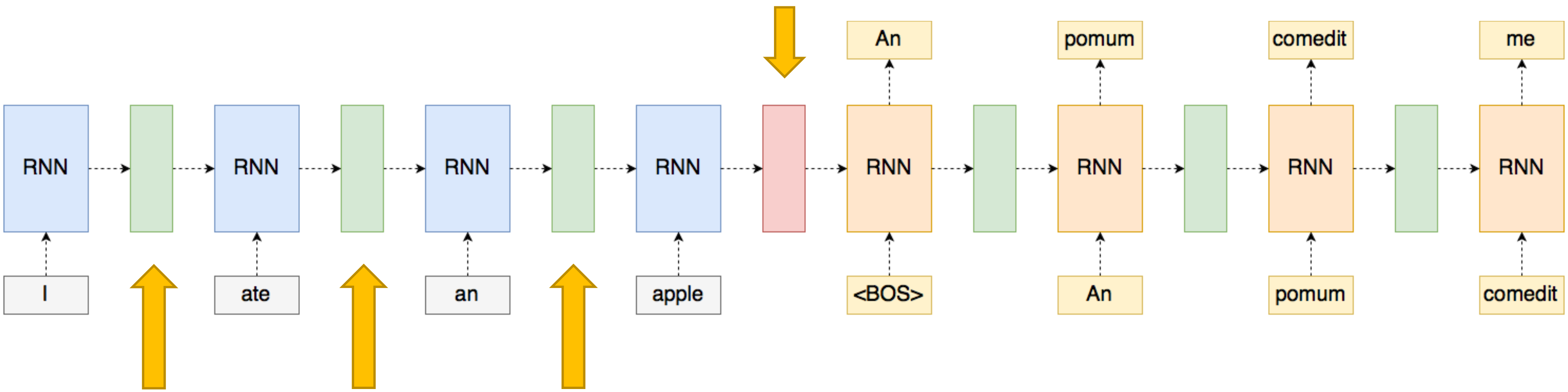
- **Recurrent Neural Network** & **Transformer**
  - Attention in RNN
  - *Attention is All You Need*: Transformer
  - Transformer for Visual Analysis
    - Visual Classification
    - Semantic Segmentation & More
- **Vision & Language**
  - Image Captioning
  - Text-to-Image Synthesis

"a corgi wearing a bow tie and a birthday hat"

Teddy bears shopping for groceries in the style of ukiyo-e

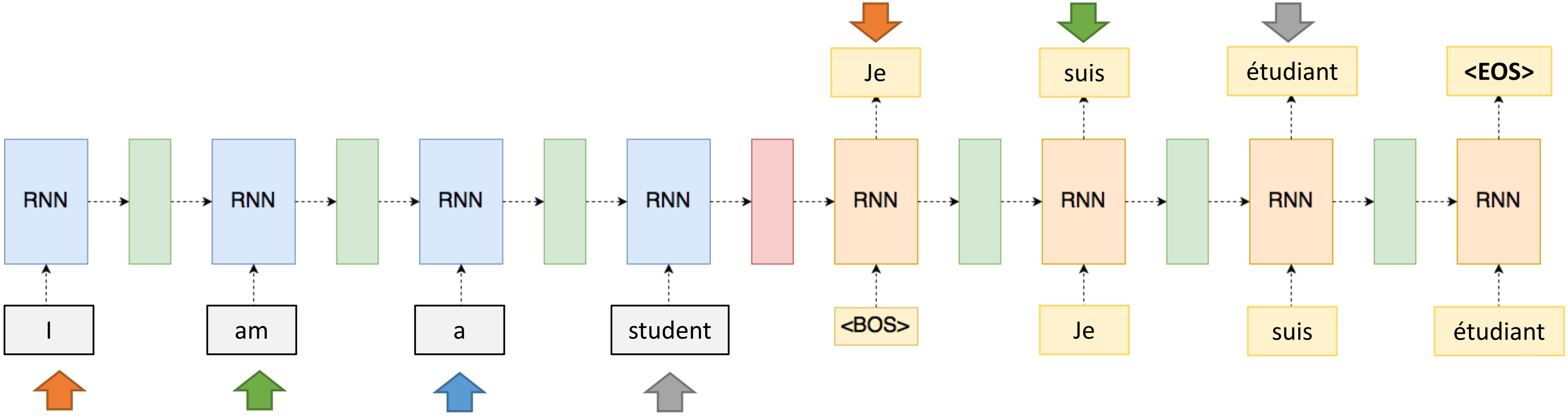
What's the Potential Problem of RNN?

- Each hidden state vector extracts/carries information across time steps (some might be diluted downstream).
- However, information of the entire input sequence is embedded into a single hidden state vector.



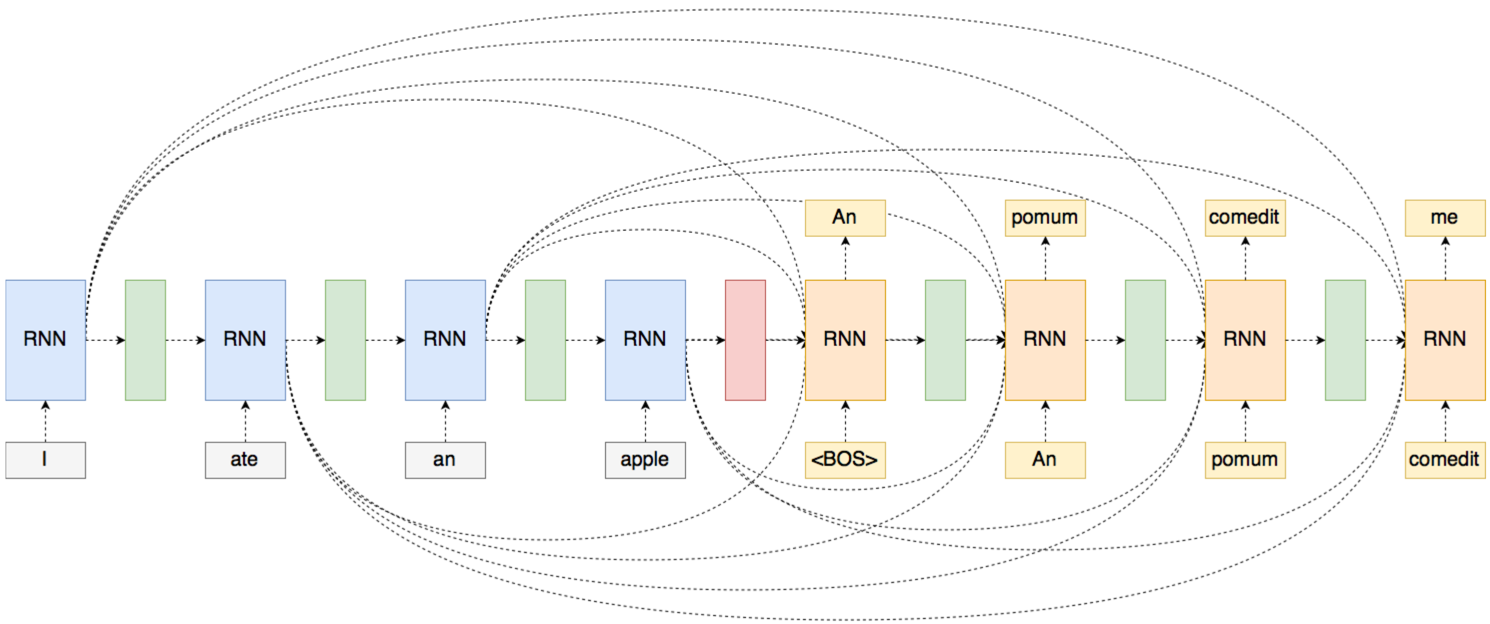
What’s the Potential Problem? (cont’d)

- Outputs at different time steps have particular meanings.
- However, synchrony between **input** and **output seqs** is not required**.**



What's the Potential Problem? (cont'd)

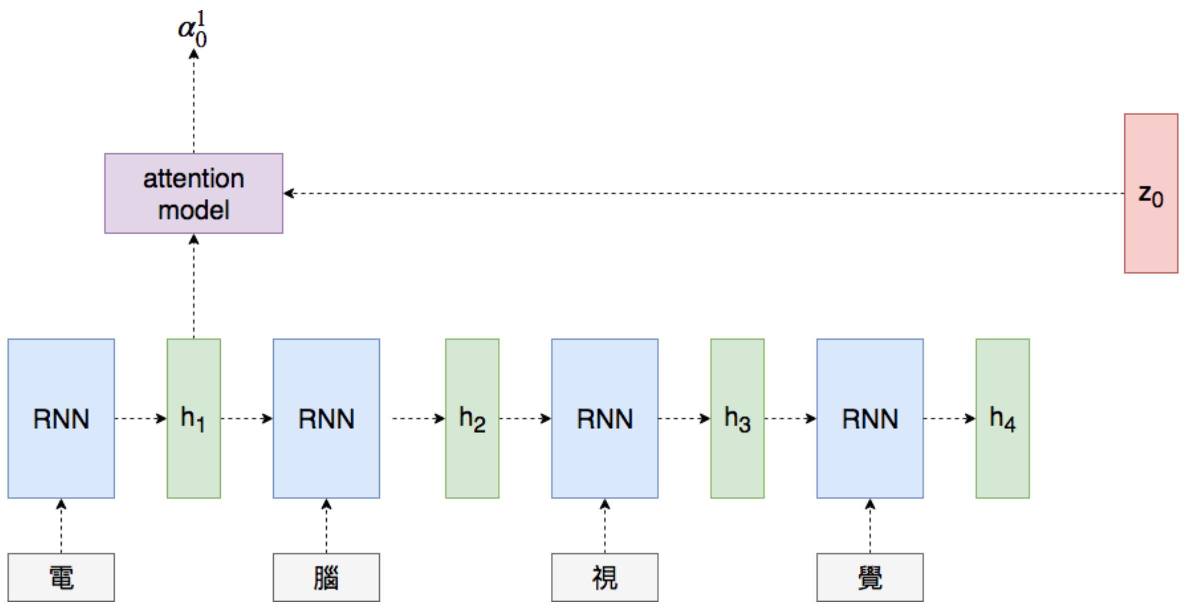
- Connecting every hidden state between encoder and decoder?



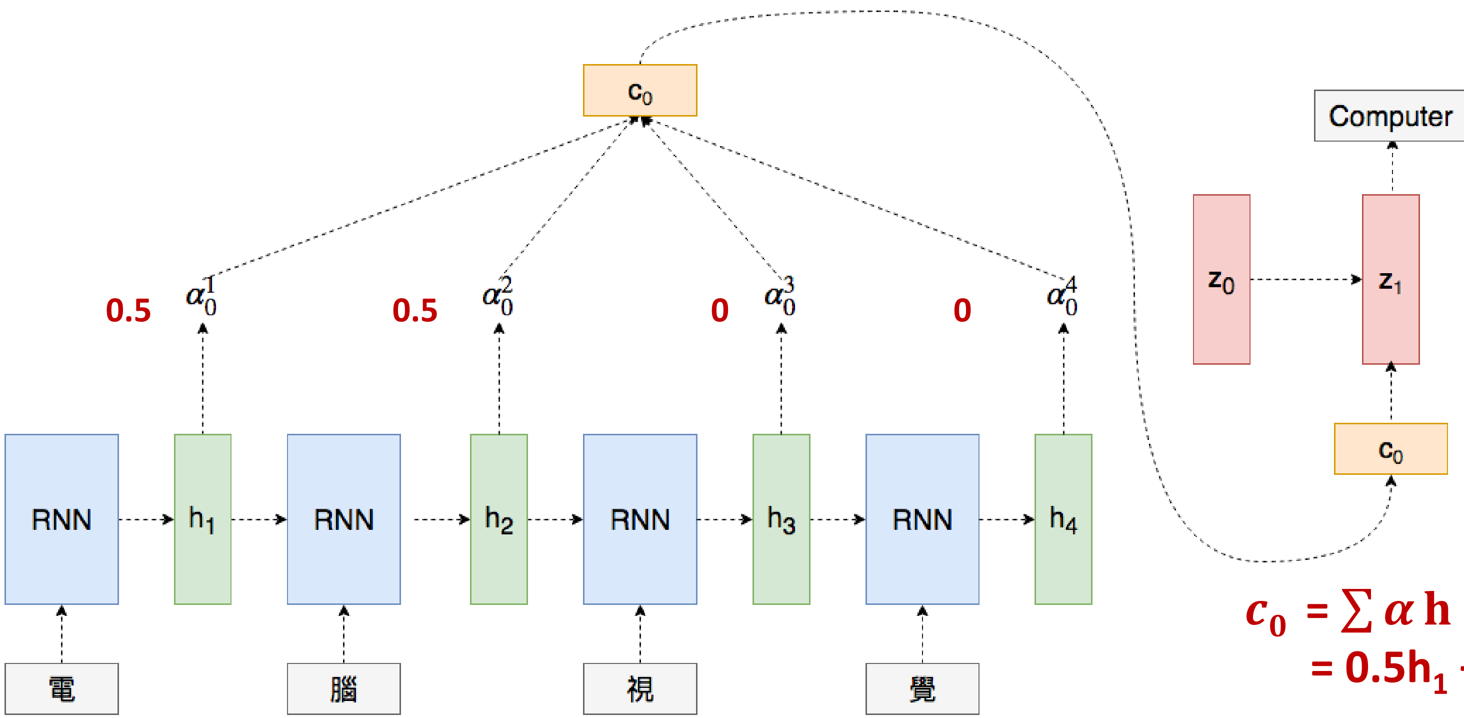
- Infeasible!
  - Both inputs and outputs are with varying sizes.
  - Overparameterized

Solution #1: Attention Model

- What should the attention model be?
  - A NN whose inputs are z and h while output is a scalar α, indicating the similarity between z and h.
- Most attention models are jointly learned with other parts of a network (e.g., classifier, regressor, etc.)
  - Will see some examples later.

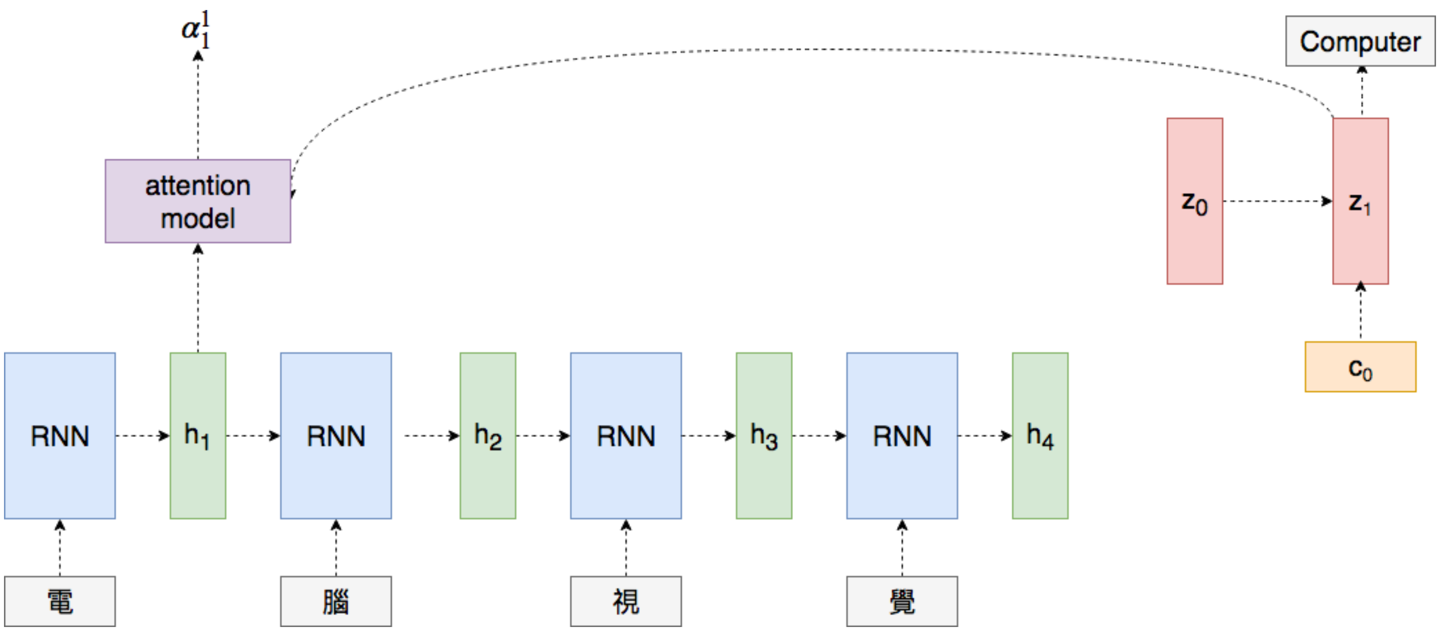


Solution #1: Attention Model

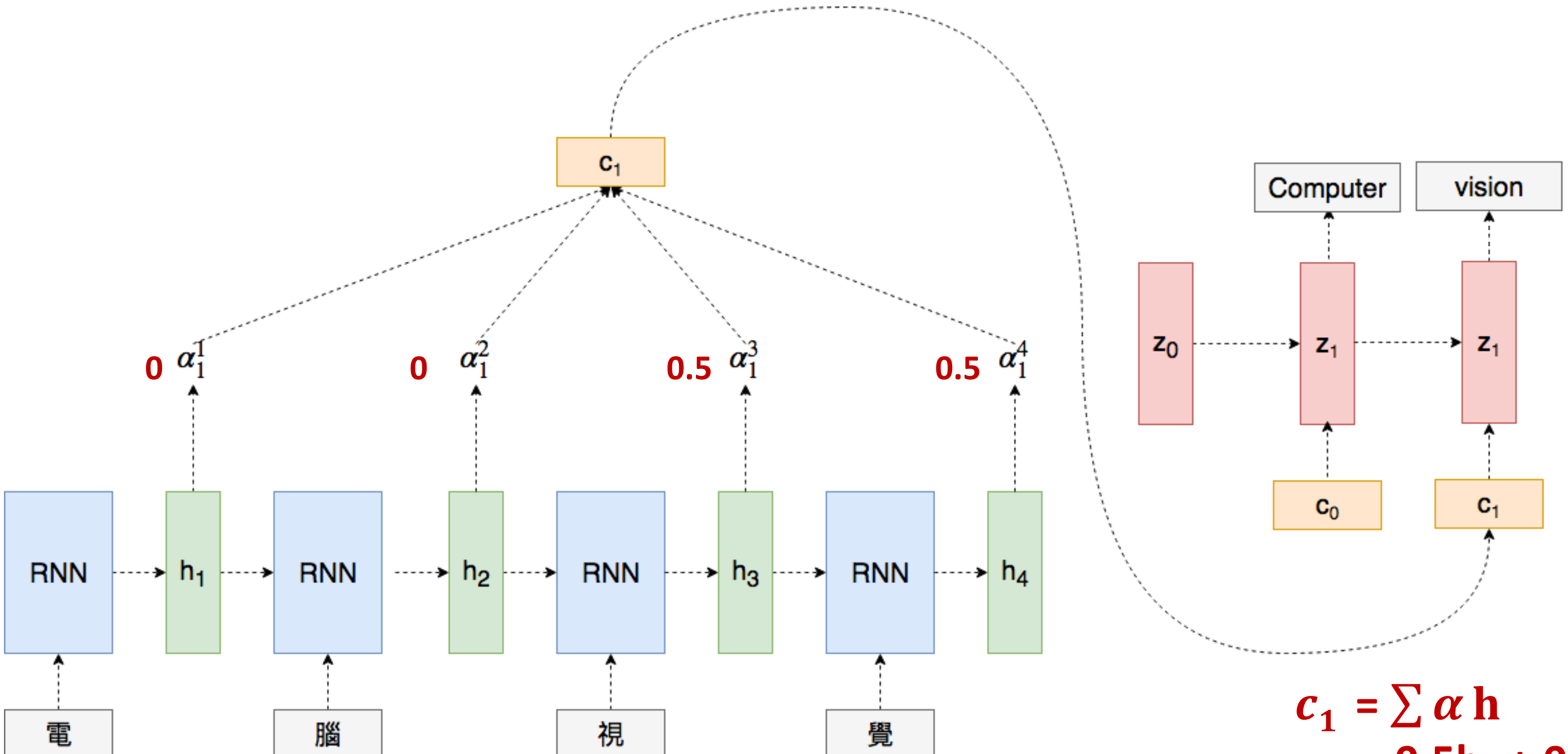


$$c_0 = \sum \alpha\, \mathbf{h}$$
$$= 0.5h_1 + 0.5h_2$$

Solution: Attention Model

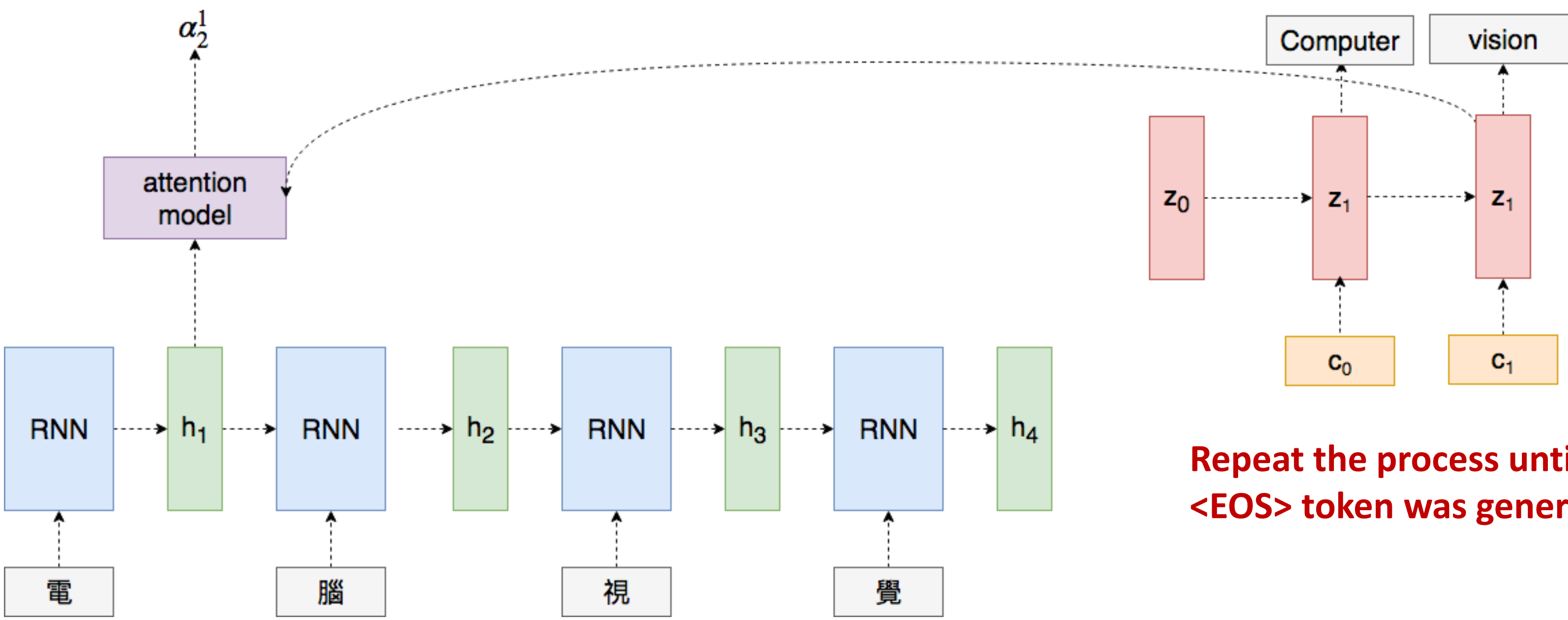


Solution: Attention Model



$$c_1 = \sum \alpha \, \mathbf{h} = 0.5 h_3 + 0.5 h_4$$

Solution: Attention Model



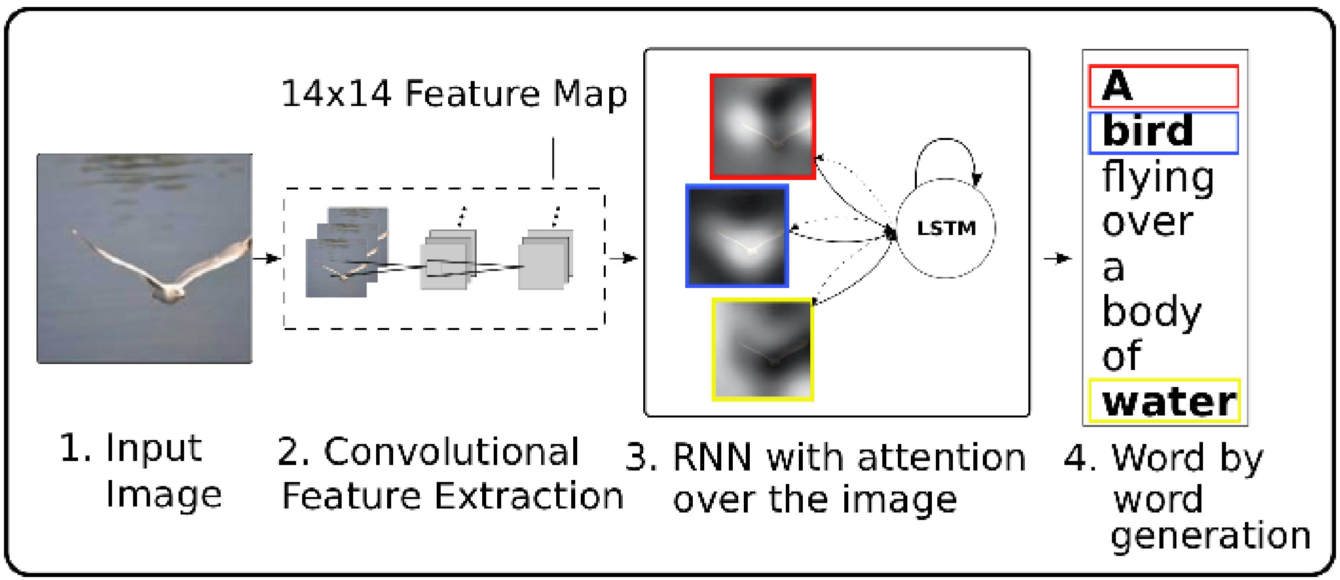
Repeat the process until <EOS> token was generated

Selected Attention Models for Image-Based Applications

- Image Captioning
  - Xu et al, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML ’15
- Visual Question Answering
  - Zhu et al, “Visual7W: Grounded Question Answering in Images”, CVPR ’16
- Image Classification
  - Mnih et al, “Recurrent Models of Visual Attention”, NIPS ’14

Image Captioning with Attention

- RNN focuses visual attention at different spatial locations when generating corresponding words during captioning.



Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

Image Captioning with Attention

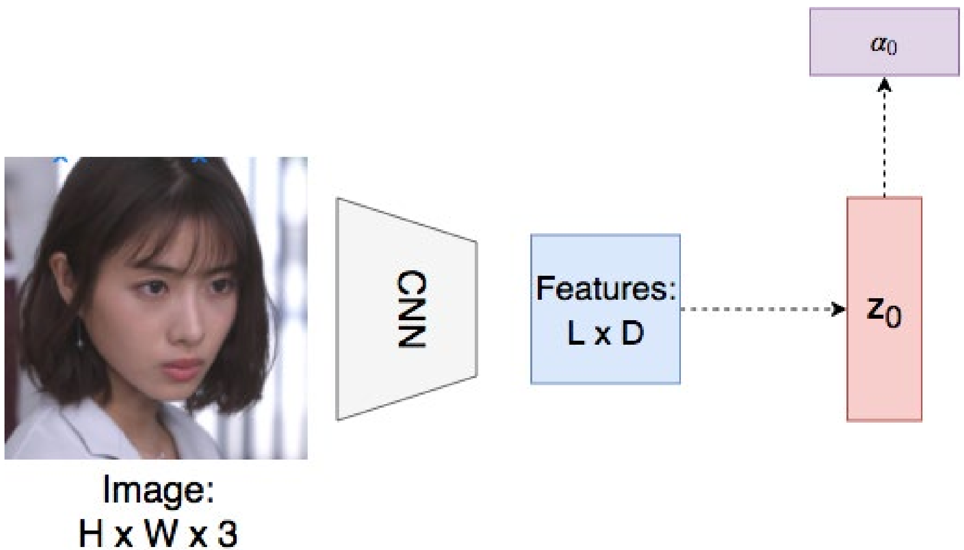
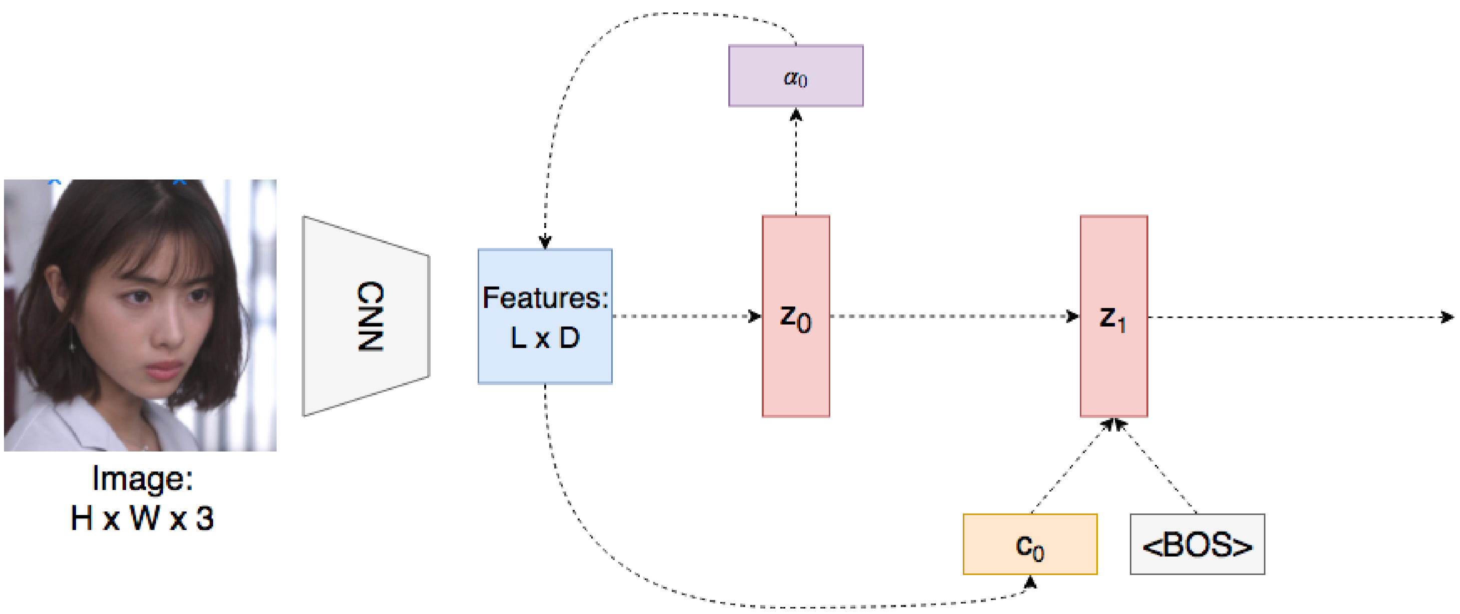


Image Captioning with Attention

Distribution of attention over L locations



Weighted combination of features

Image Captioning with Attention

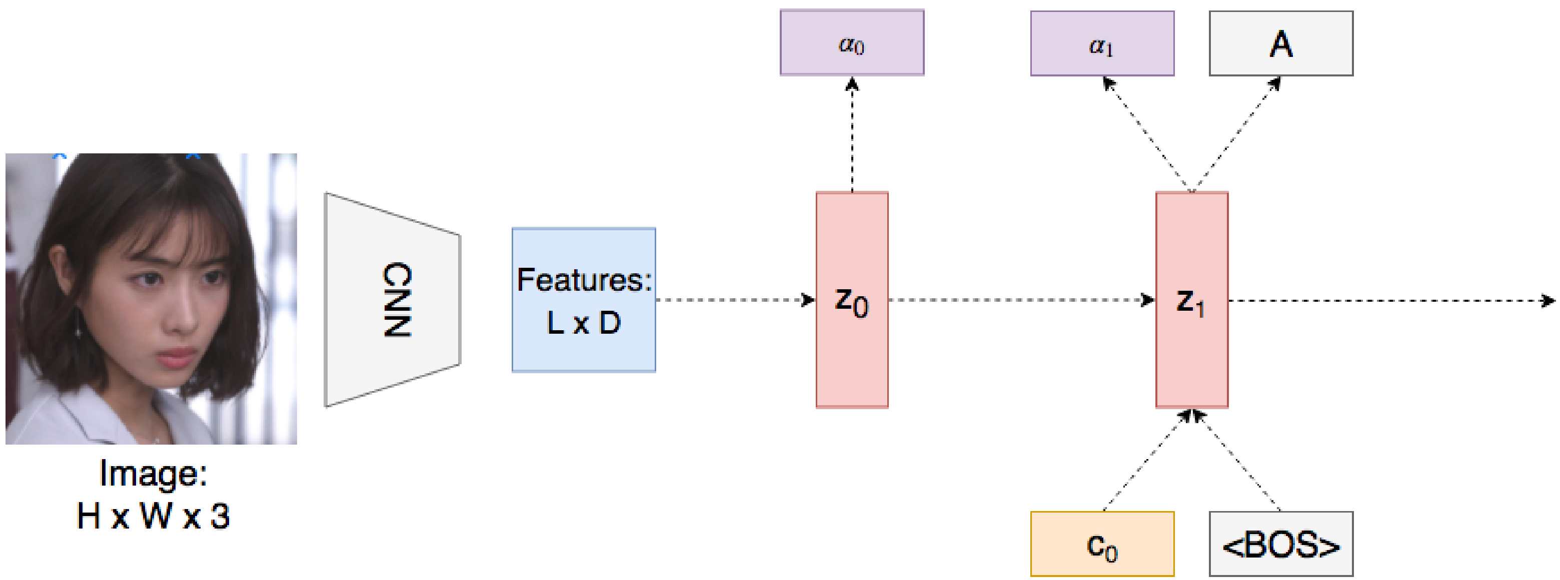


Image Captioning with Attention

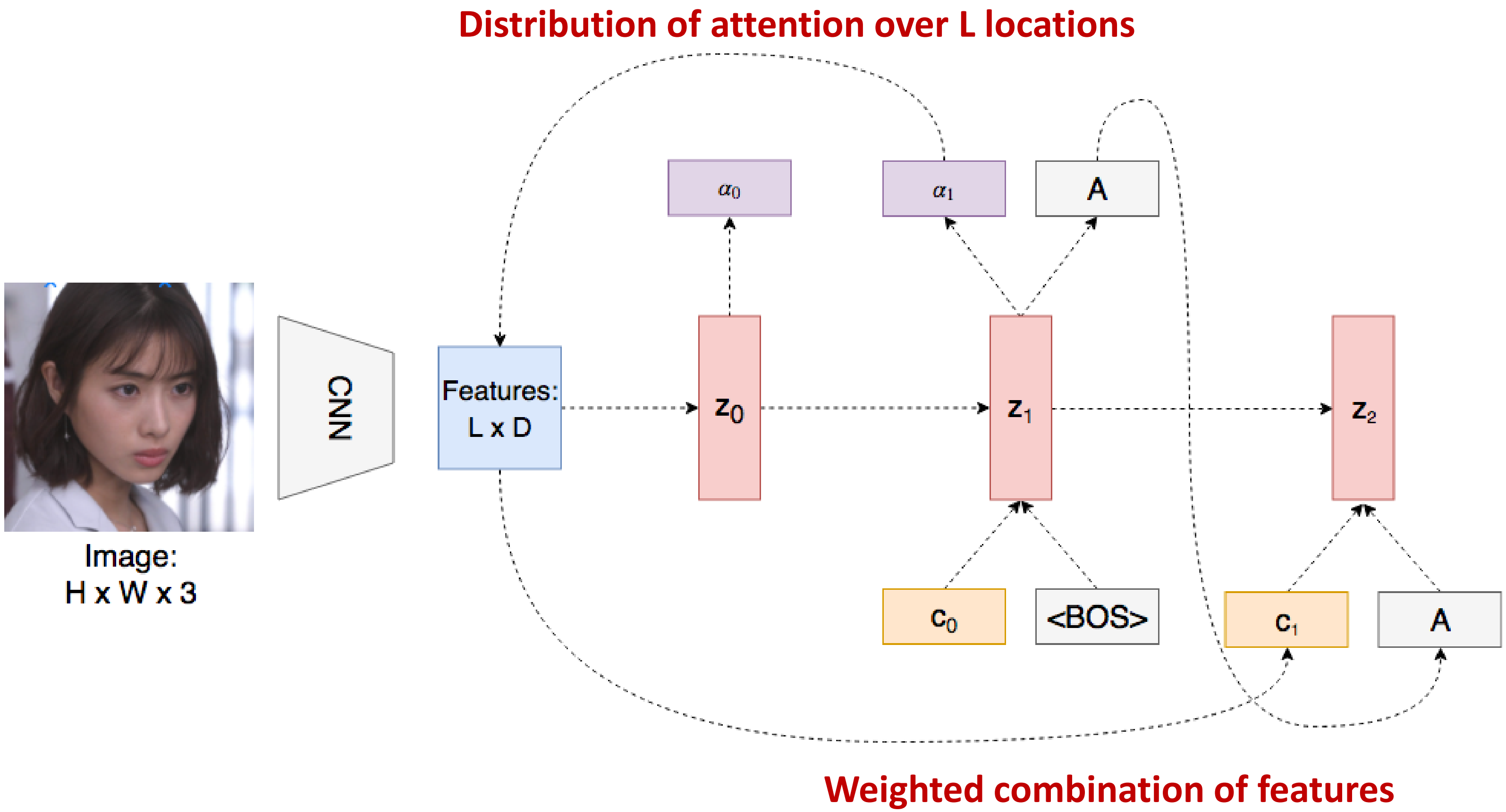
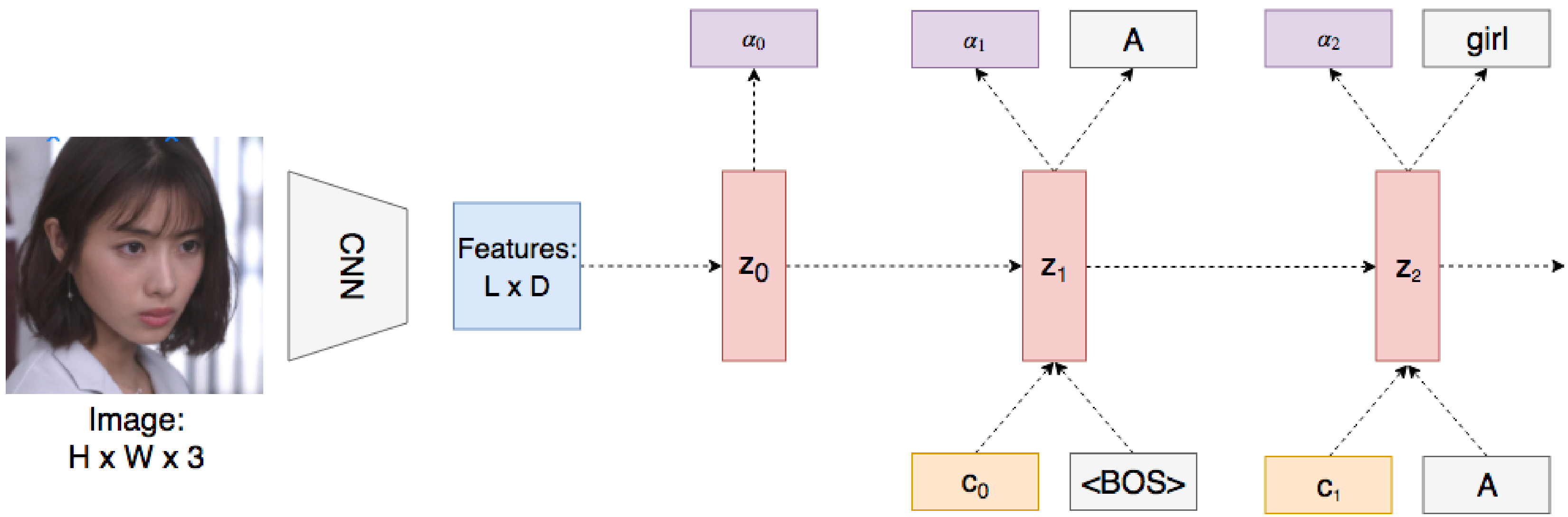


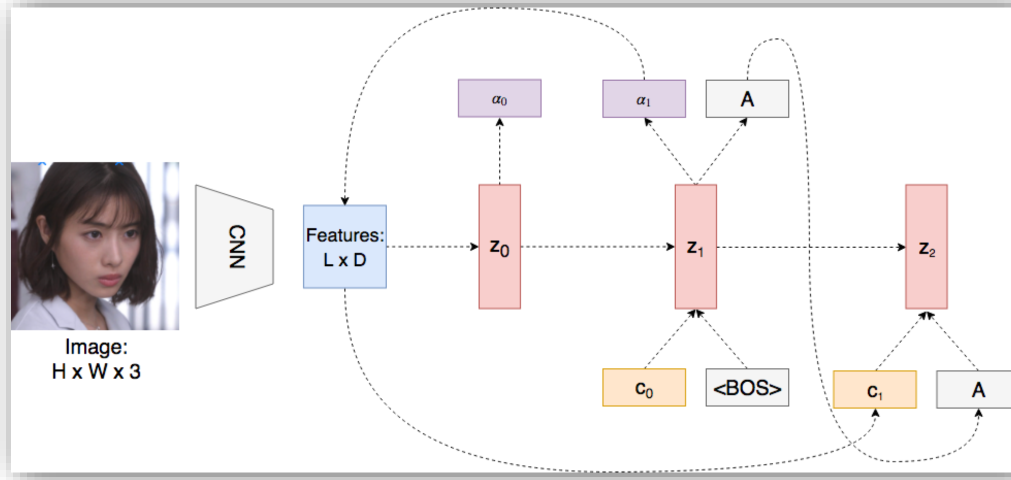
Image Captioning with Attention

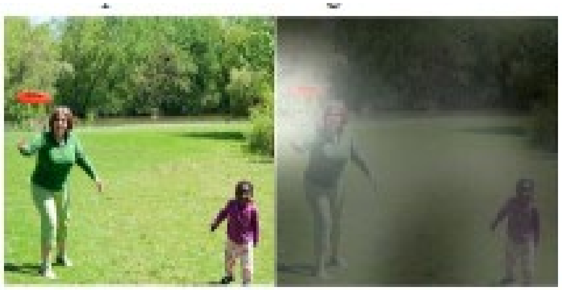


Repeat the process until <EOS> token was generated

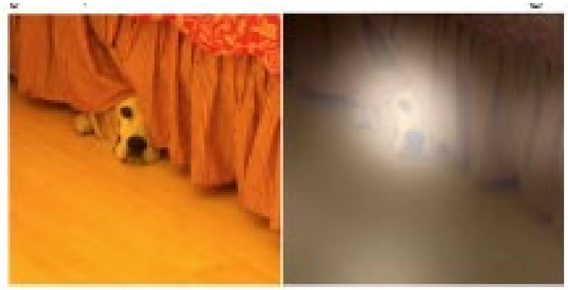
Image Captioning with Attention

- Visualization



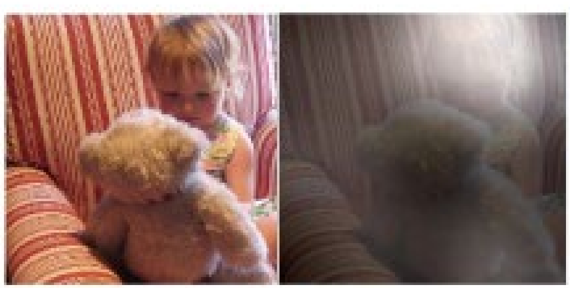


A woman is throwing a frisbee in a park.



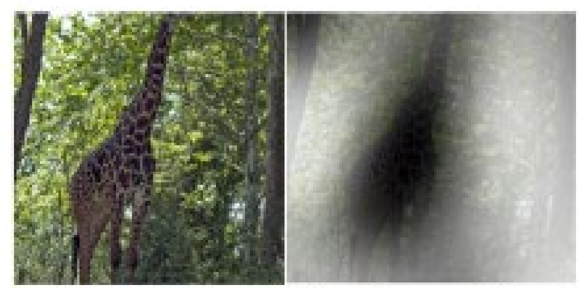
A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.



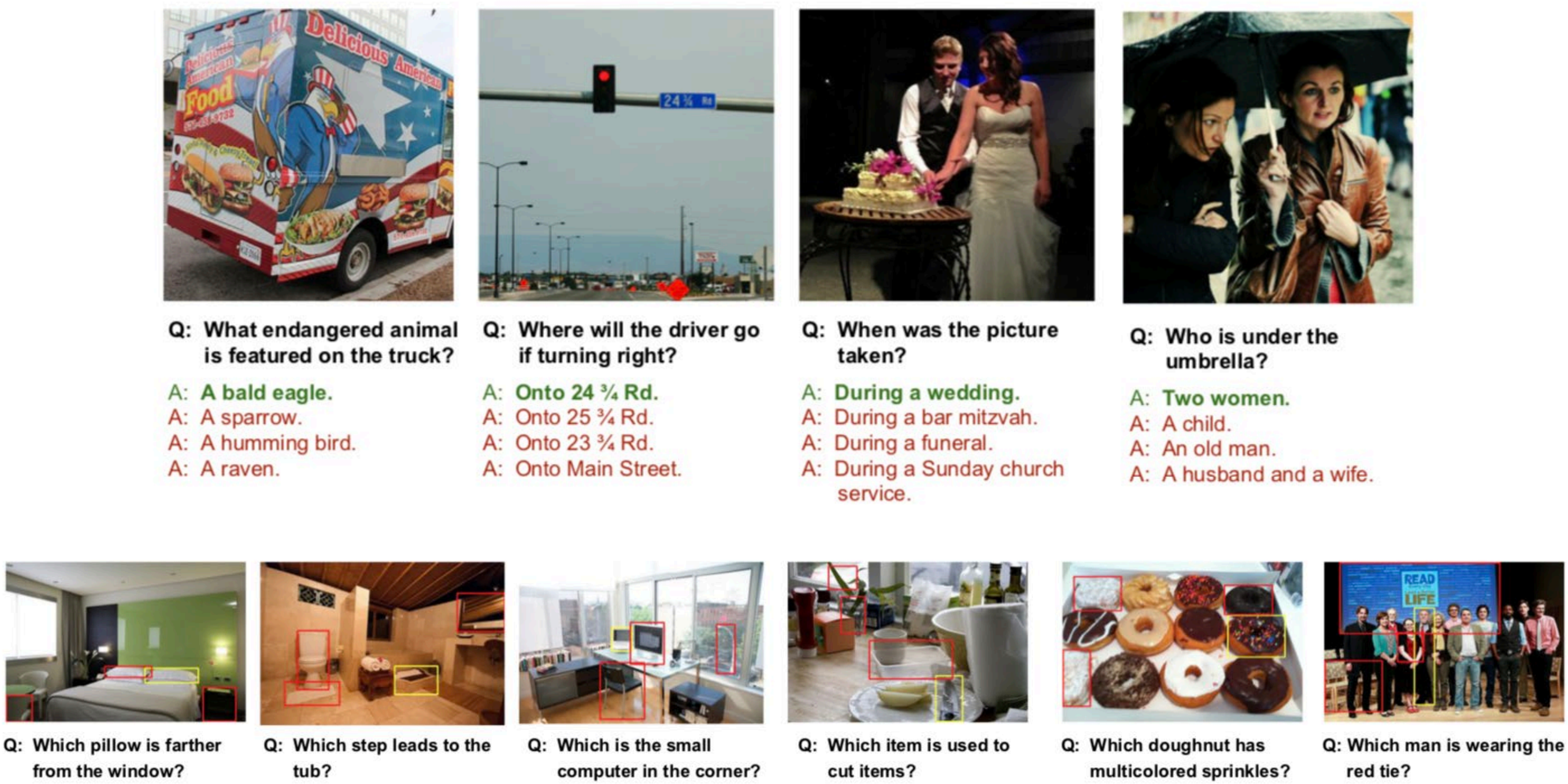
A giraffe standing in a forest with trees in the background.

Selected Attention Models for Image-Based Applications

- Image Captioning
  - Xu et al, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML ’15
- Visual Question Answering
  - Zhu et al, “Visual7W: Grounded Question Answering in Images”, CVPR ’16
- Image Classification
  - Mnih et al, “Recurrent Models of Visual Attention”, NIPS ’14

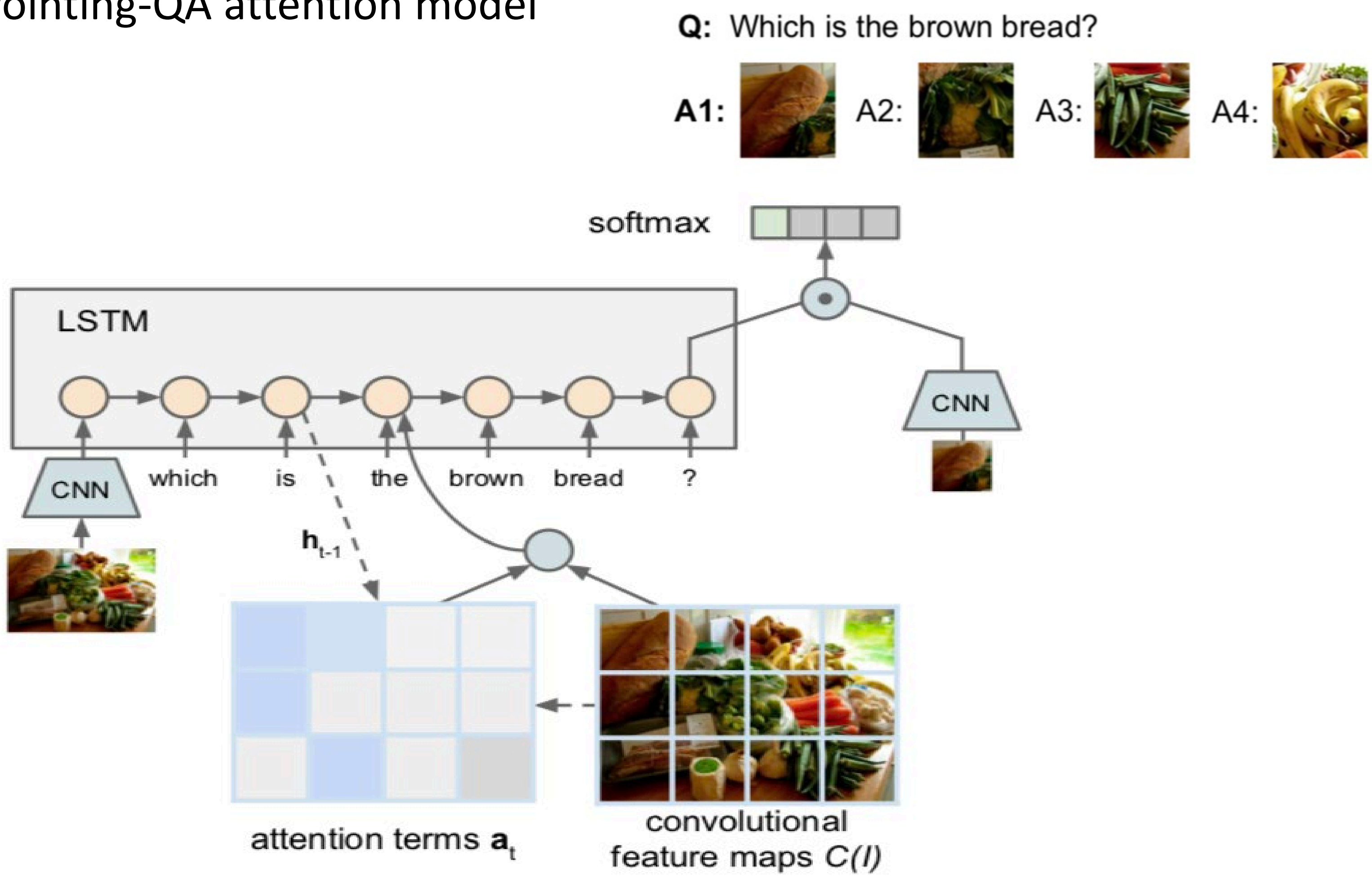
Visual Question Answering

- Examples of multiple-choice QA & pointing QA



Visual Question Answering with Attention

- Pointing-QA attention model



Visual Question Answering with Attention (cont'd)

What kind of animal is in the photo?
A **cat**.

Why is the person holding a knife?
To cut the **cake** with.

Where are the carrots?
At the top.

How many people are there?
Three.

A & B (answers related to physicla objs):
The peaks of the attention maps reside in the bounding boxes of the target objects.

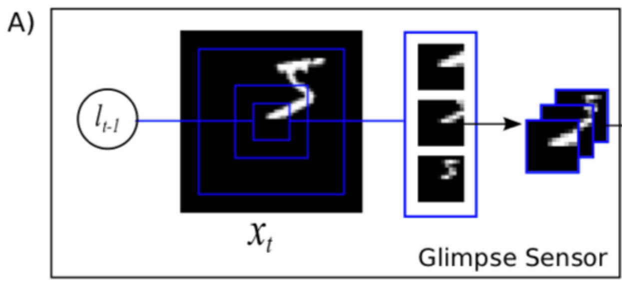
C & D (answers w/ non-physical objs):
The bottom two examples show QA pairs with answers not explicitly containing objects.
The attention heat maps are scattered around the image grids.

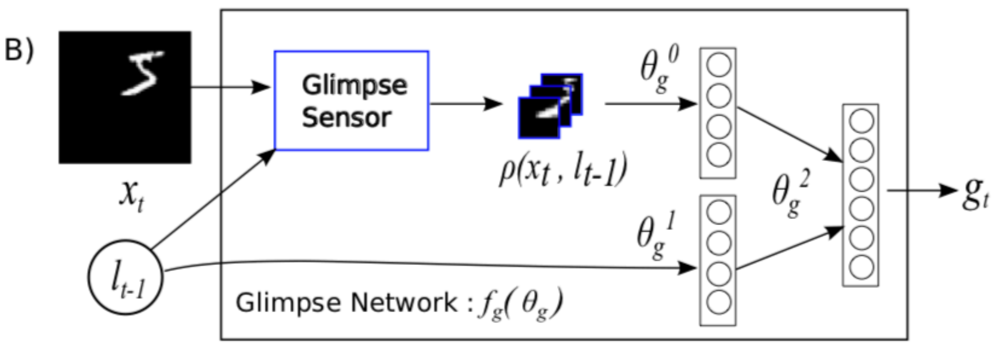
Selected Attention Models for Image-Based Applications

- Image Captioning
  - Xu et al, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, ICML ’15
- Visual Question Answering
  - Zhu et al, “Visual7W: Grounded Question Answering in Images”, CVPR ’16
- Image Classification
  - Mnih et al, “Recurrent Models of Visual Attention”, NIPS ’14

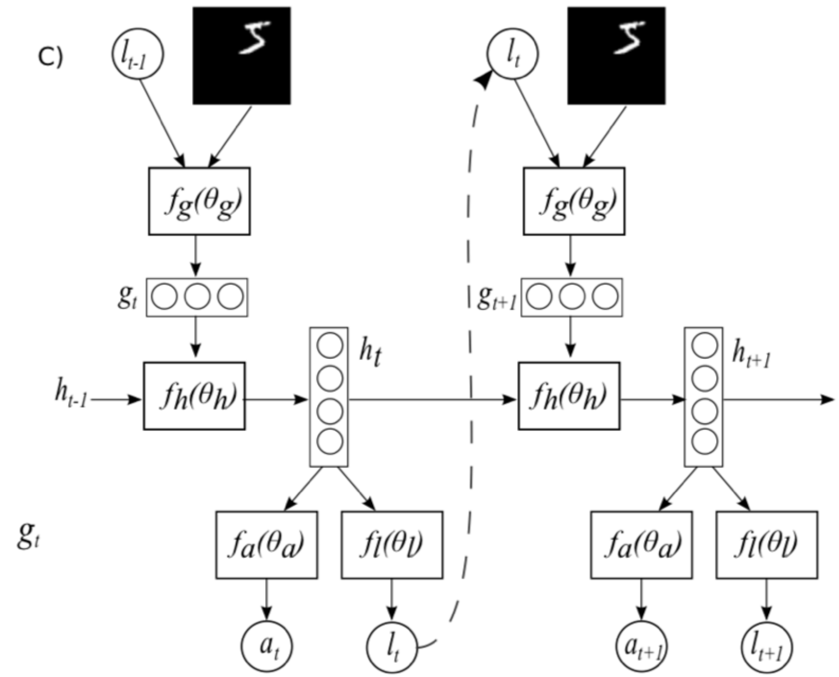
Glimpse Sensor & Glimpse Network

Glimpse sensor: extracts a retina-like representation centered at l_{t-1} that contains multiple resolution patches.





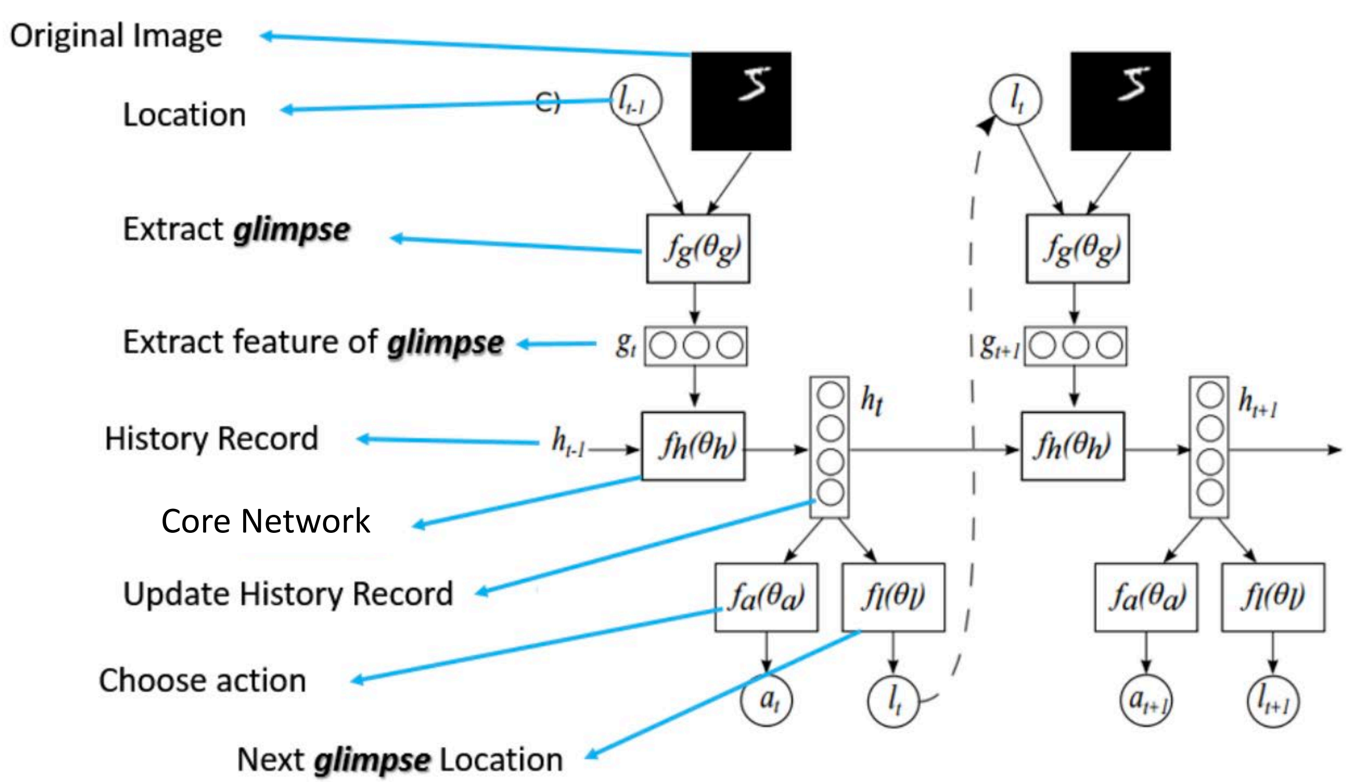
Glimpse network: given location l_{t-1} and image x_t, use the glimpse sensor to extract retina representation, which is mapped into a joint hidden space.



RNN-based model architecture: the core network takes the glimpse representation as input with the hidden state vector from the prevision step, and outputs the new hidden state resulting in **location** and **action** networks to predict the next location to attend and the associated action.

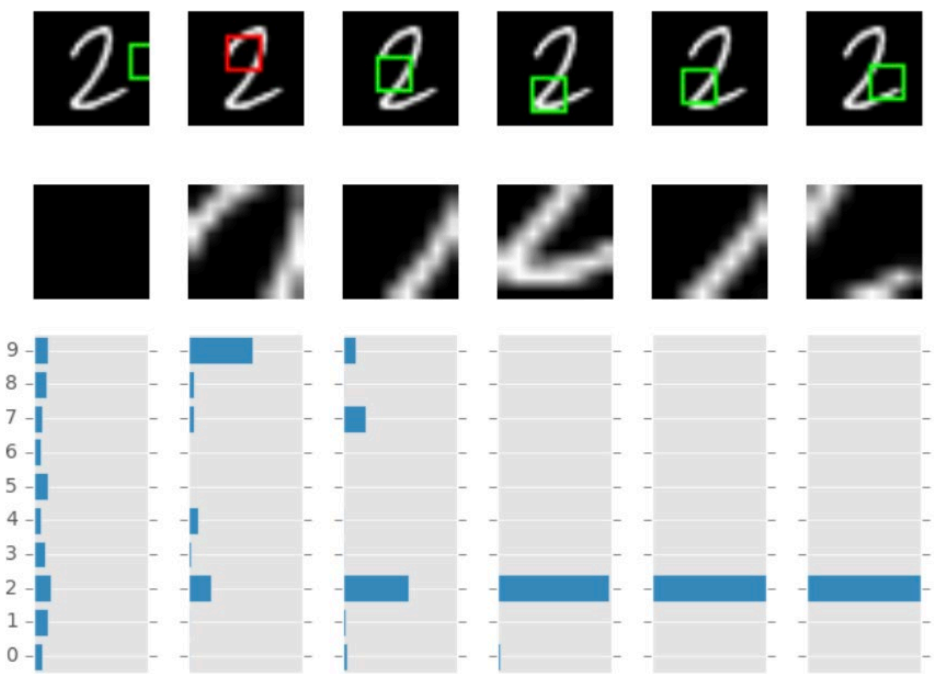
Mnih et al, “Recurrent Models of Visual Attention”, NIPS ‘14

Architecture: RNN with Attention Models

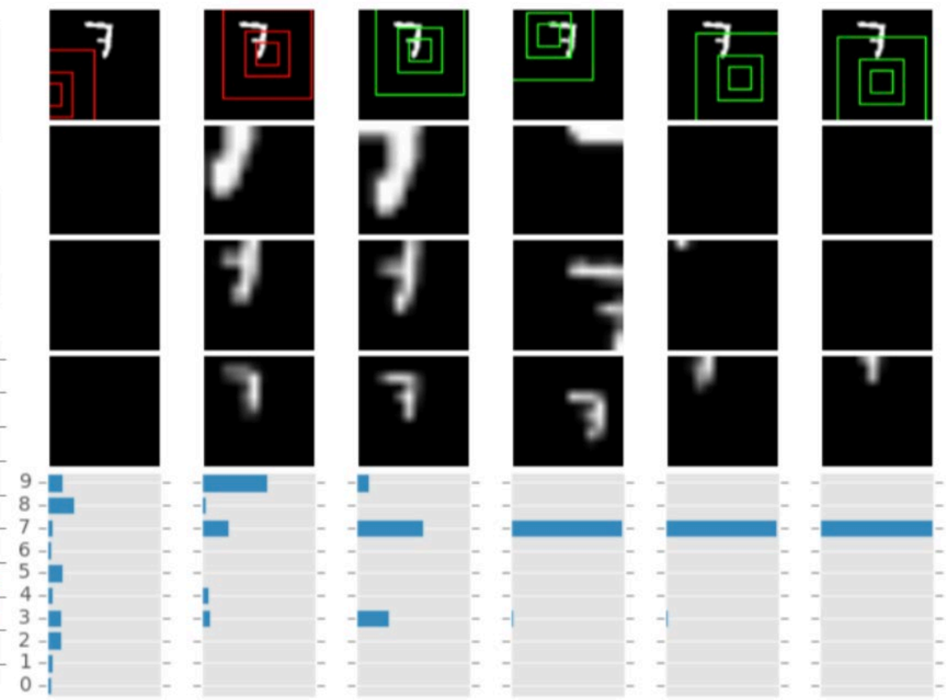


Example Results

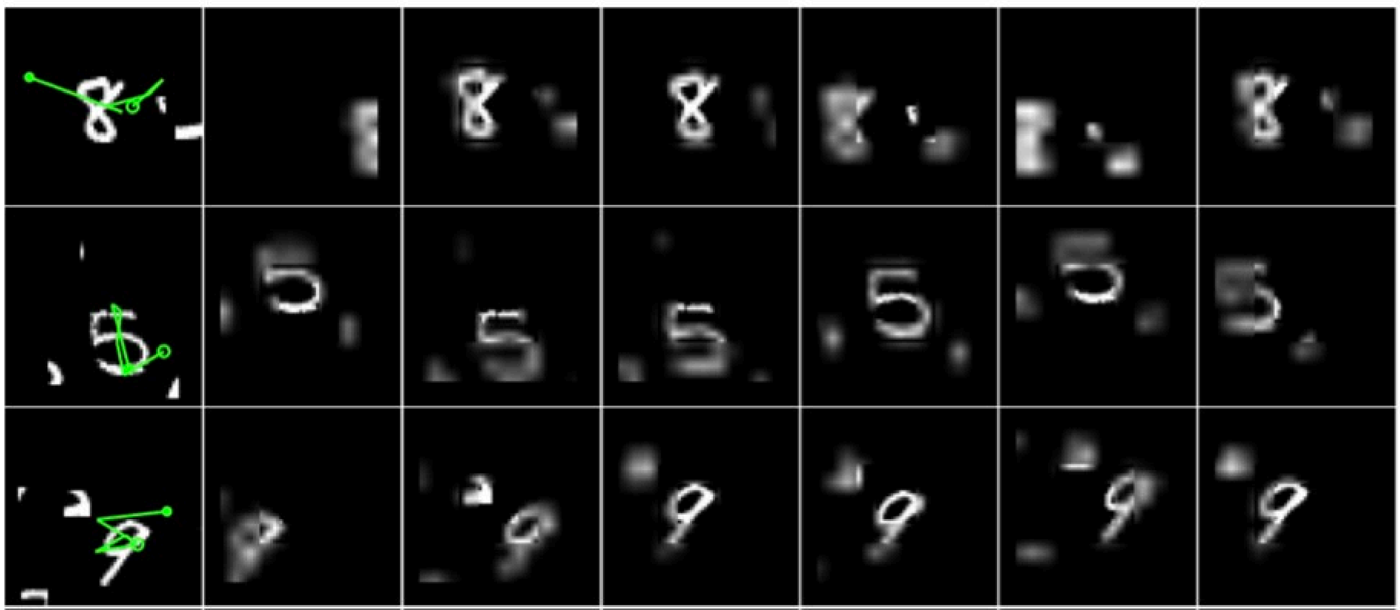
- Original MNIST



- Translated MNIST



Example: Actual Glimpse Path



What to Cover Today…

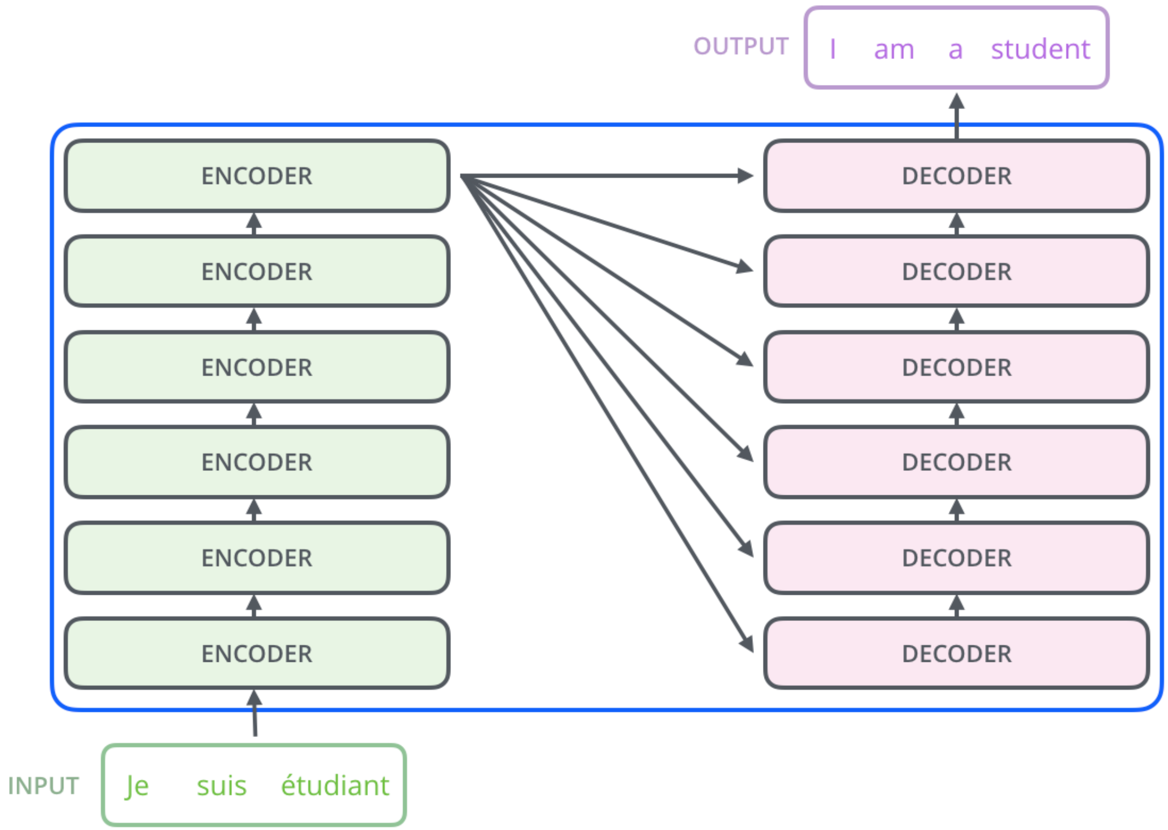
- **Recurrent Neural Network** & **Transformer**
  - Attention in RNN
  - *Attention is All You Need*: Transformer
  - Transformer for Visual Analysis
    - Visual Classification
    - Semantic Segmentation & More
- **Vision & Language**
  - Image Captioning
  - Text-to-Image Synthesis

Teddy bears shopping for groceries in the style of ukiyo-e

"a corgi wearing a bow tie and a birthday hat"

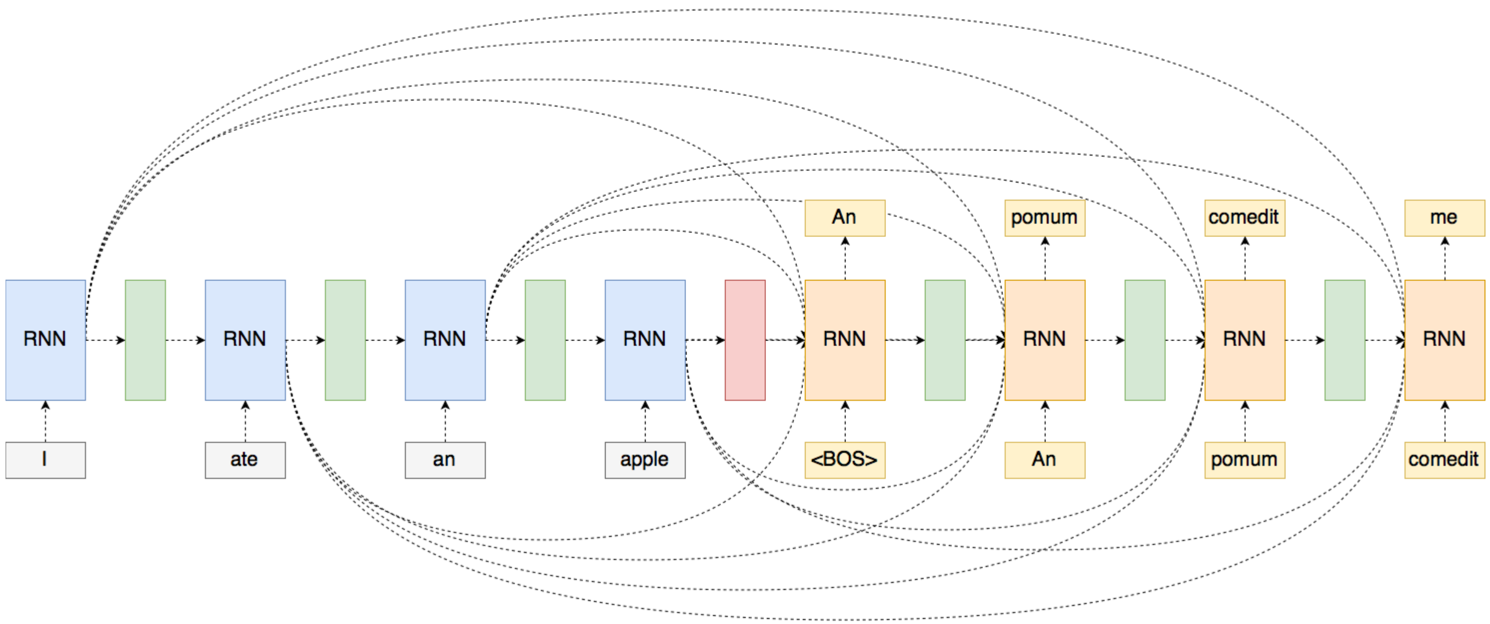
RNN with Attention is Good, But..

- Attention in a pre-defined sequential order
- Information loss due to long sequences…



RNN with Attention is Good, But..

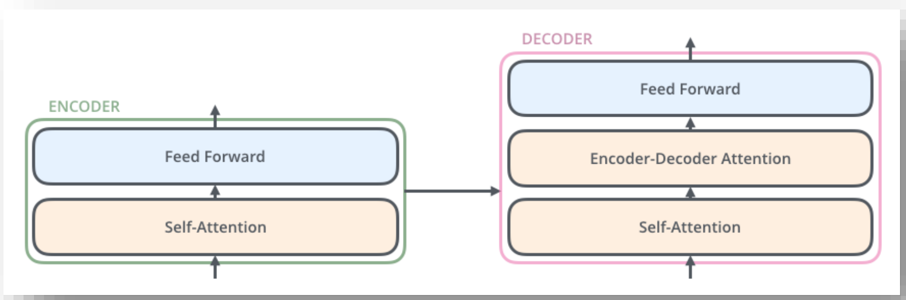
- Connecting every hidden state between encoder and decoder?

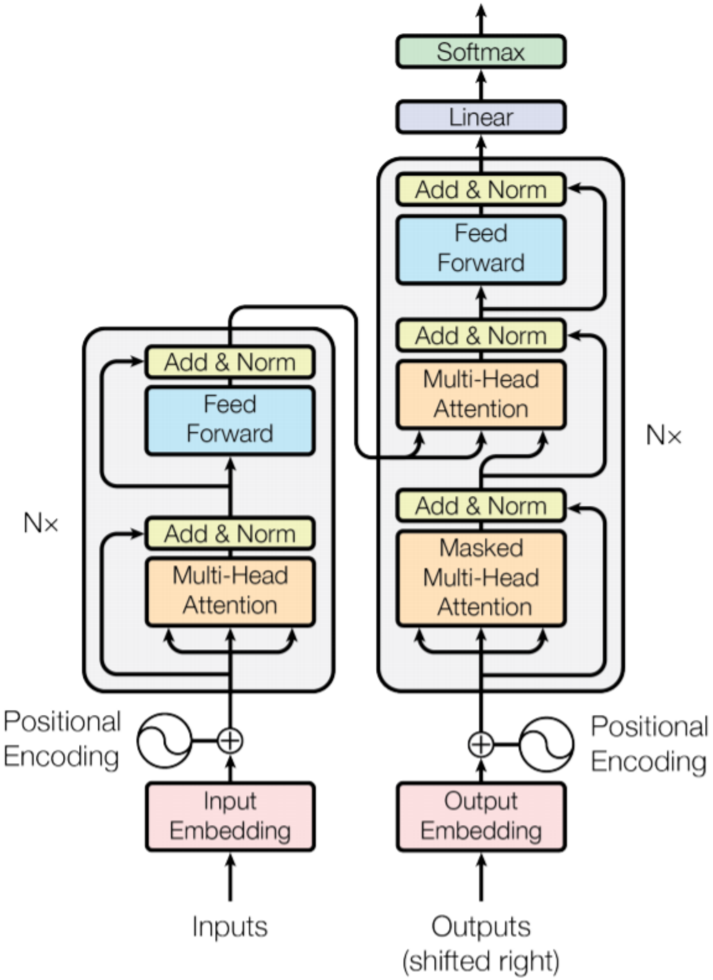


- Infeasible!
  - Both inputs and outputs are with varying sizes.
  - Overparameterized

Solution #2: Transformer

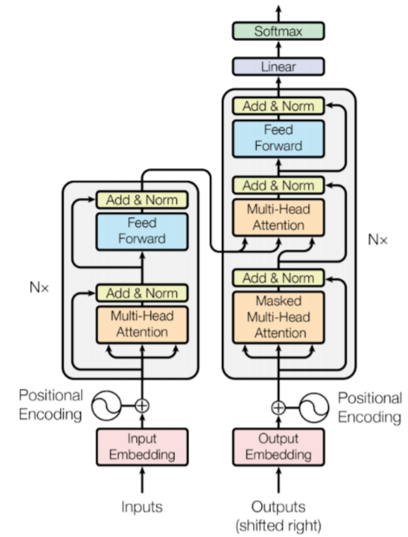
- “Attention is all you need”, NeurIPS 2017
- More details available at: http://jalammar.github.io/illustrated-transformer/

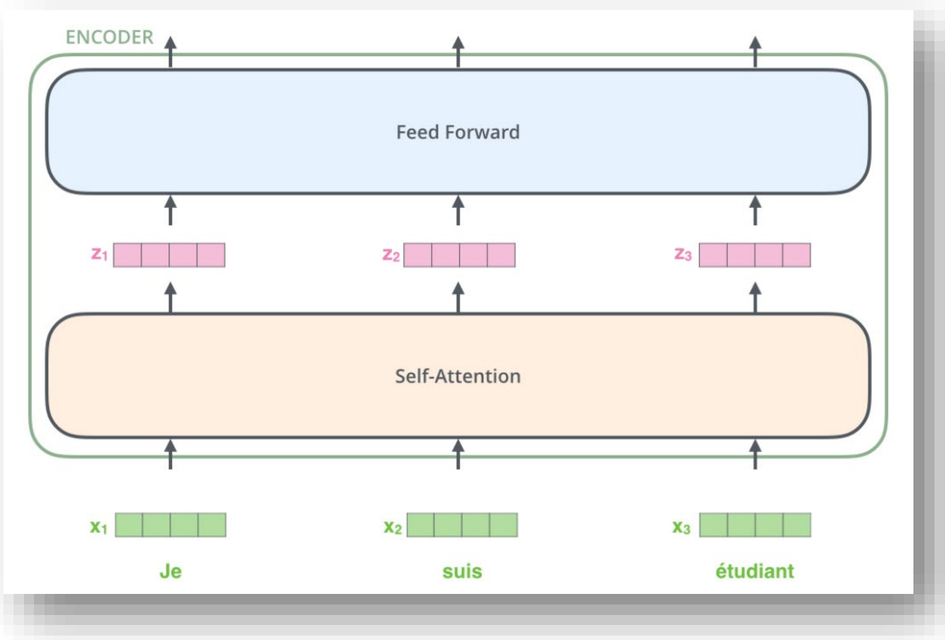


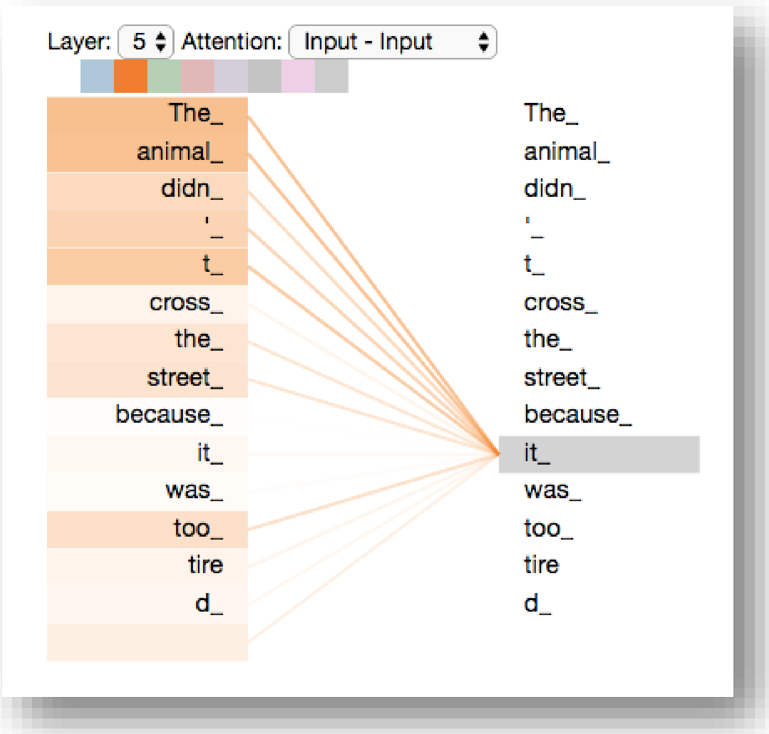


Transformer

- "Attention is all you need", NeurIPS 2017
- Self-attention for text translation



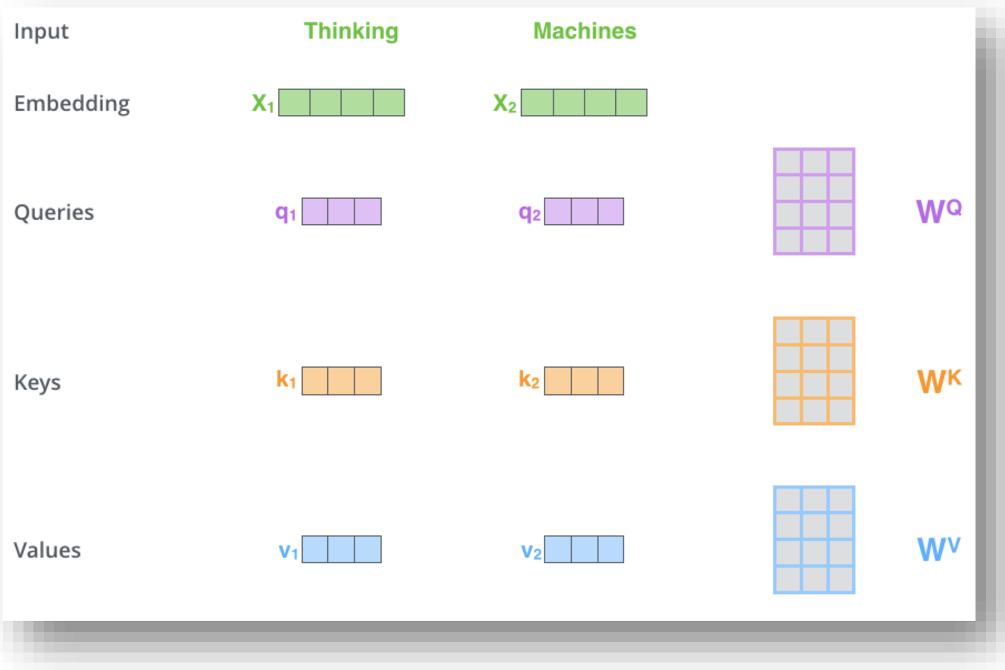


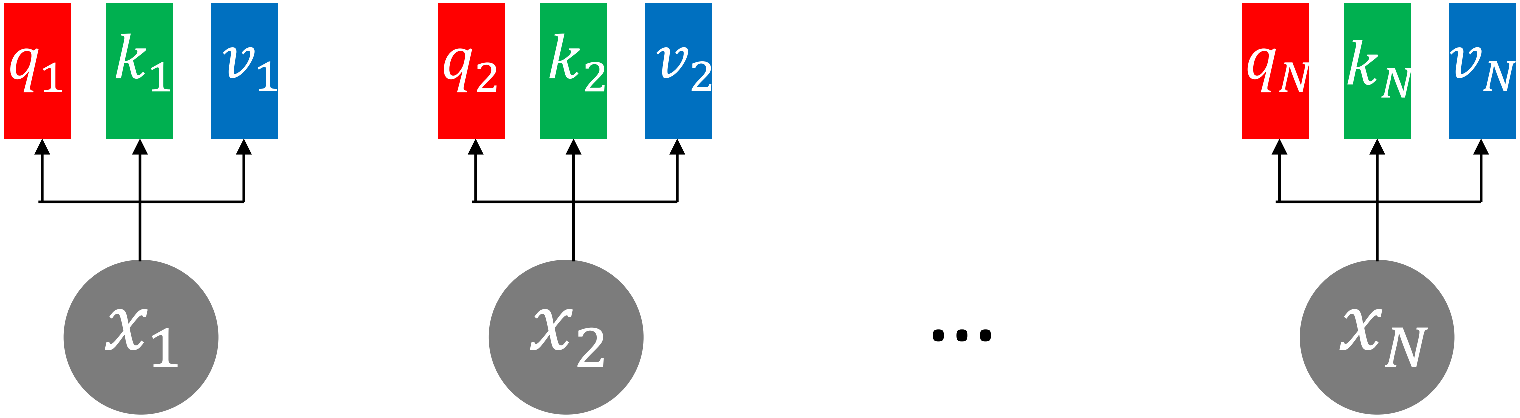


Self-Attention (1/5)

- Query $\boldsymbol{q}$, key $\boldsymbol{k}$, value $\boldsymbol{v}$ vectors are learned from each input $\boldsymbol{x}$

$$q_i = W^Q x_i$$
$$k_i = W^K x_i$$
$$v_i = W^V x_i$$

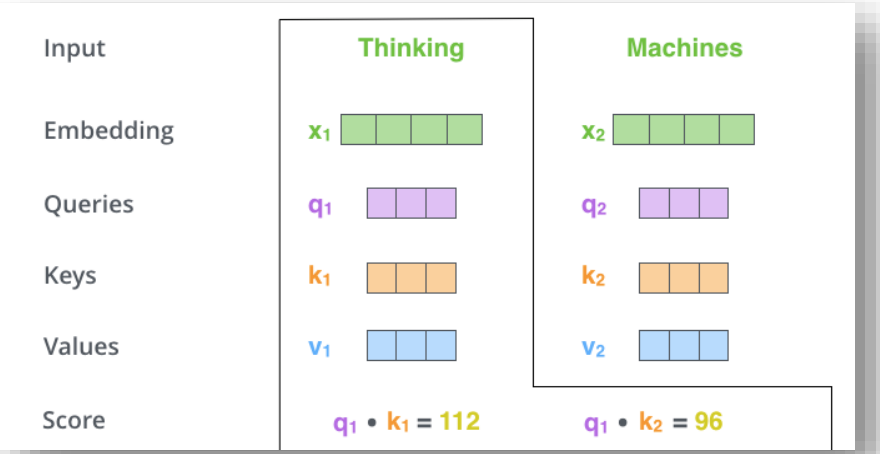


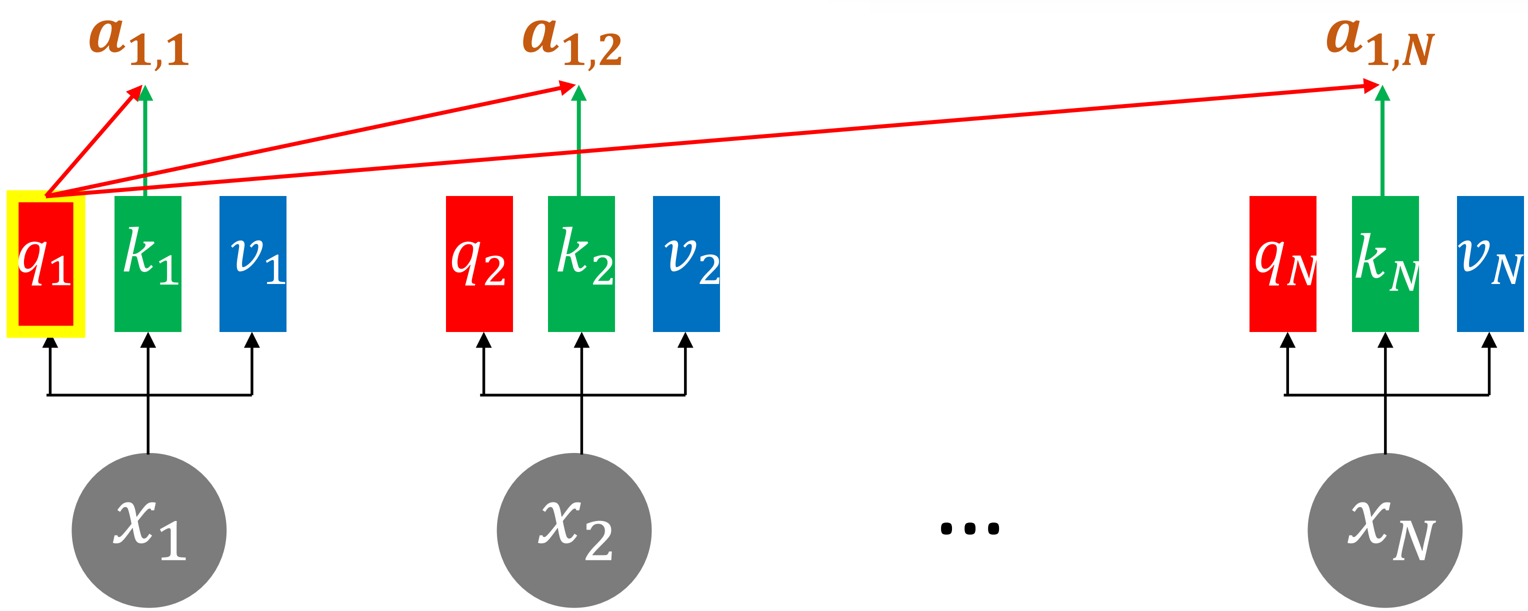


Self-Attention (2/5)

- Relation between each input is modeled by inner-product of query $\boldsymbol{q}$ and key $\boldsymbol{k}$.

$$a_{1,i} = \frac{q_1 \cdot k_i}{\sqrt{d}}, \text{ where } a \in R, q, k \in R^d$$

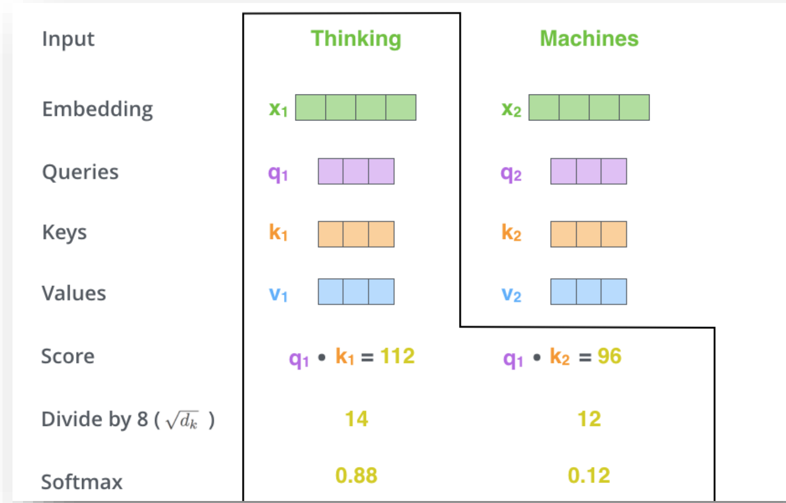


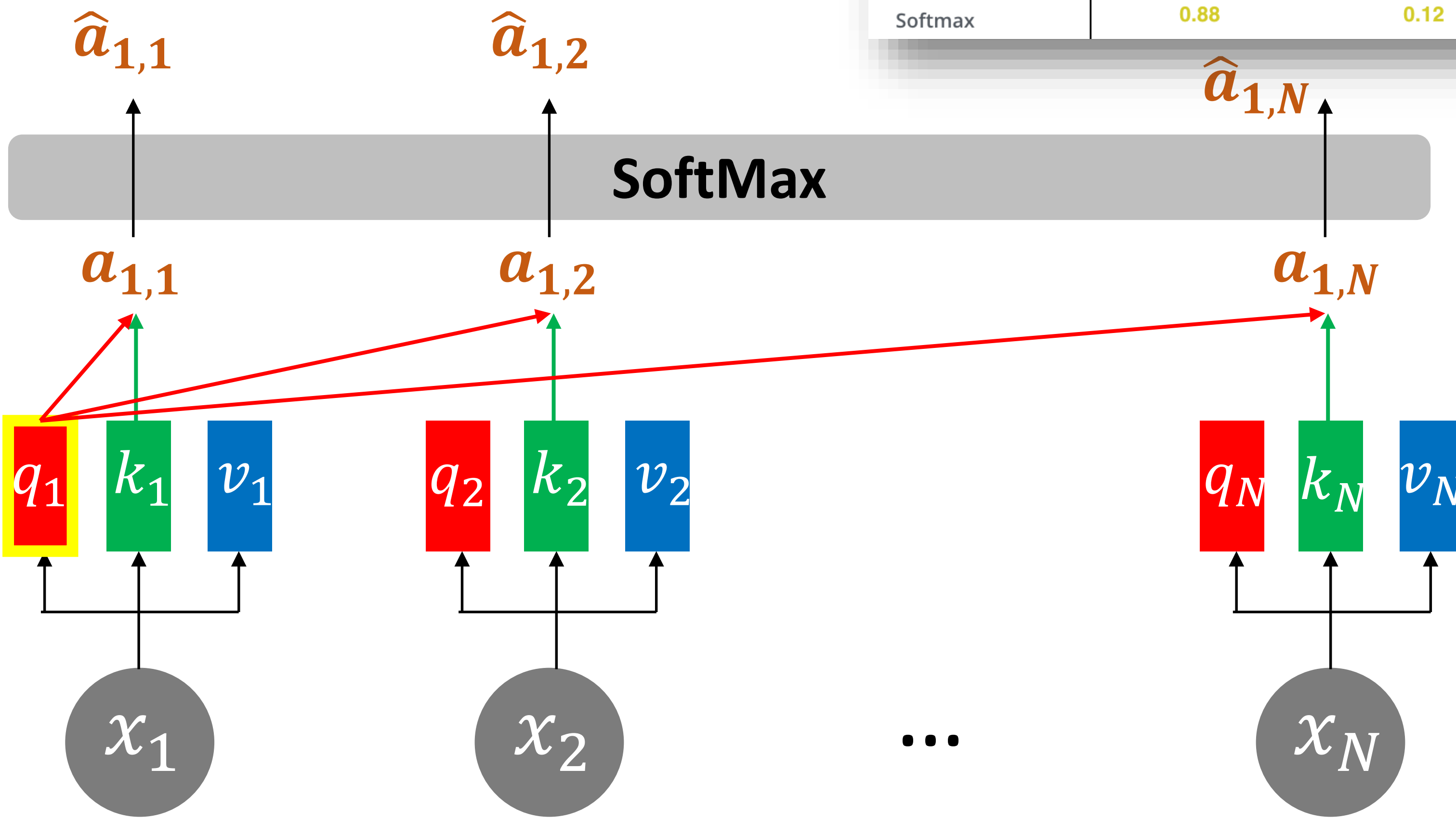


Self-Attention (3/5)

- SoftMax is applied:

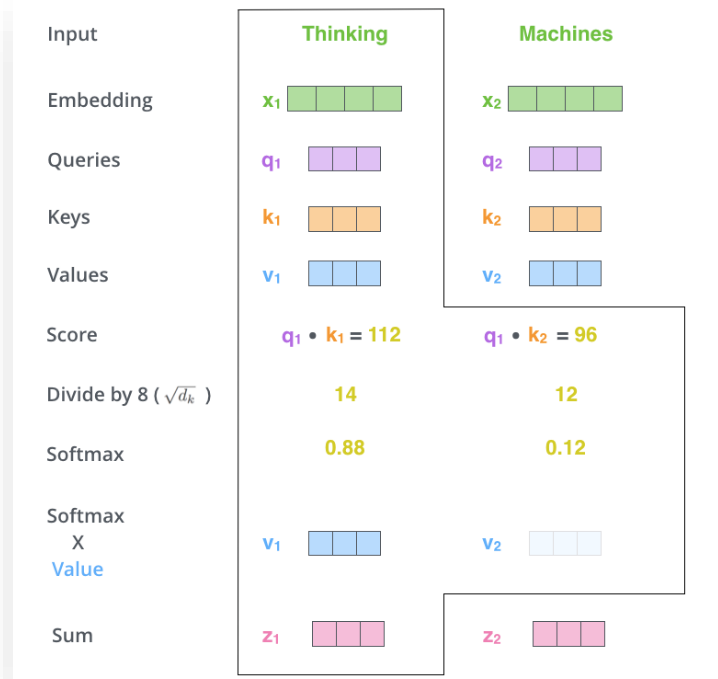
  $0 \le \hat{a}_i = e^{a_i} / \sum_j^N e^{a_j} \le 1$, for I =1, …, N

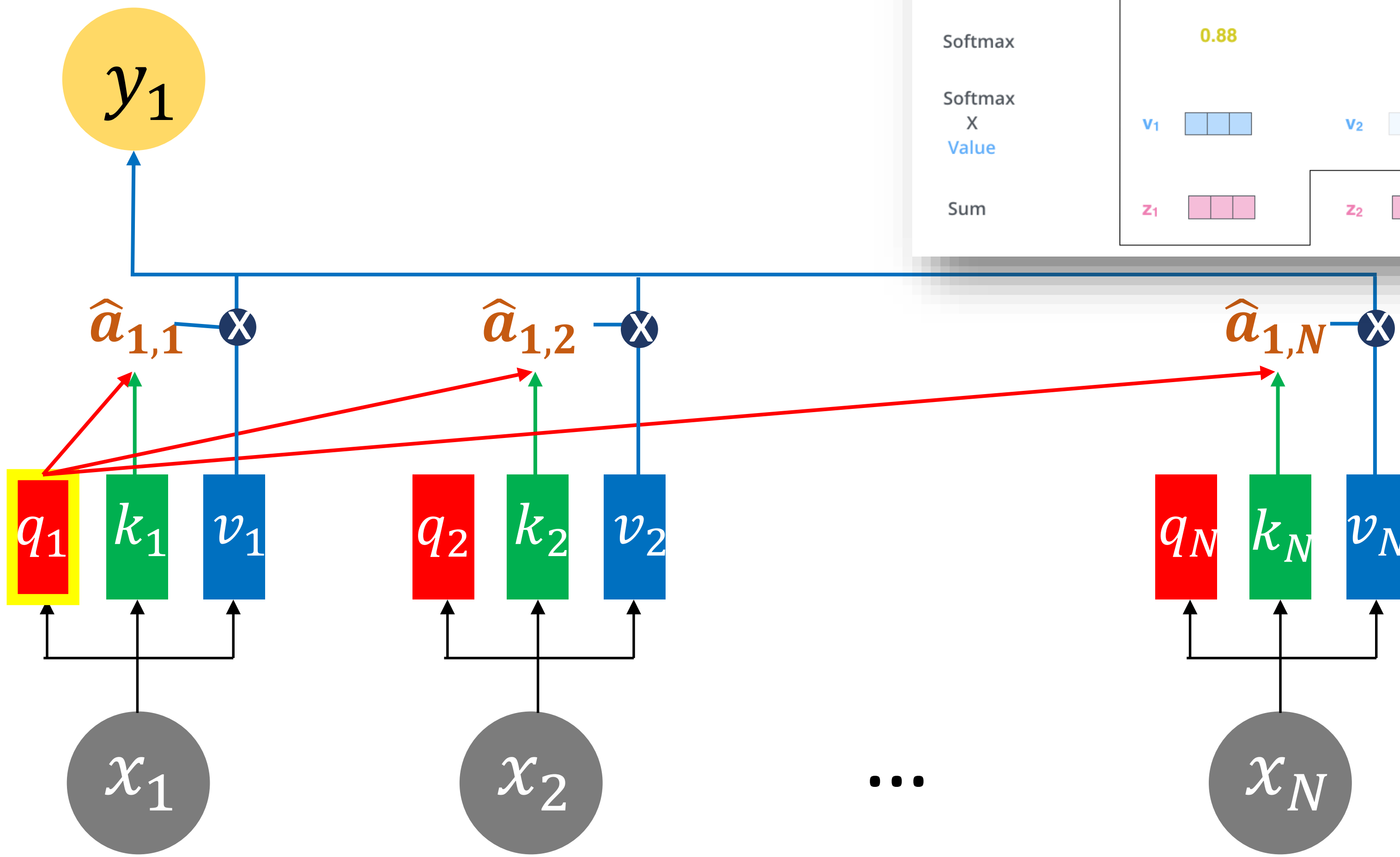




Self-Attention (4/5)

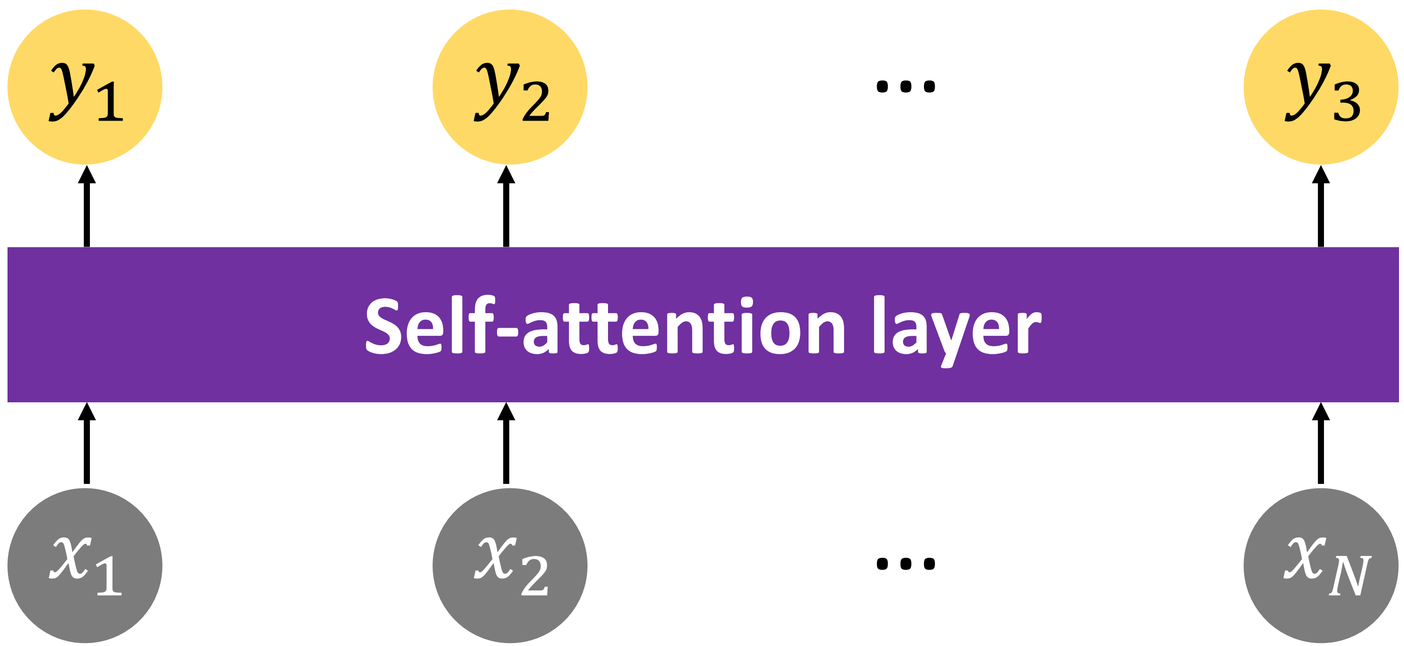
- Value vectors $\boldsymbol{v}$ are aggregated with attention weight $\hat{a}$, i.e., $y_1 = \sum_i^N \hat{a}_i \cdot v_i$





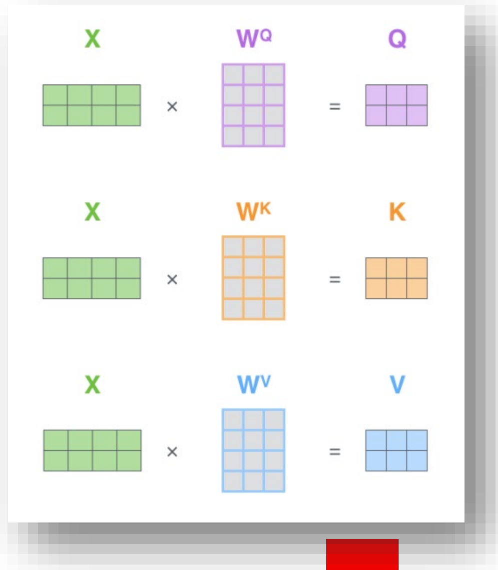
Self-Attention (5/5)

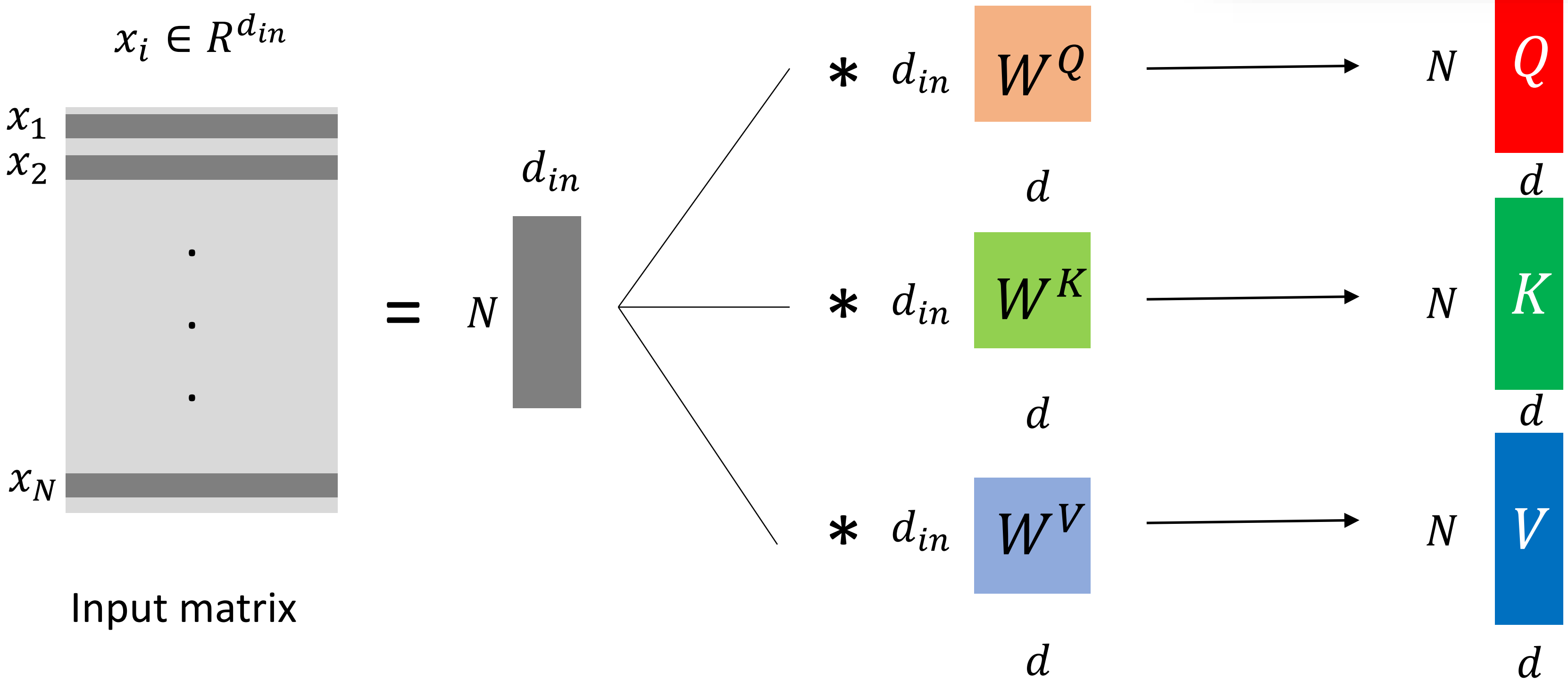
- All y_i can be computed in parallel
- y_i considers $x_1 \sim x_N$, modeling long-distance dependencies.
- Global feature can be obtained by average-pooling over $y_1 \sim y_N$



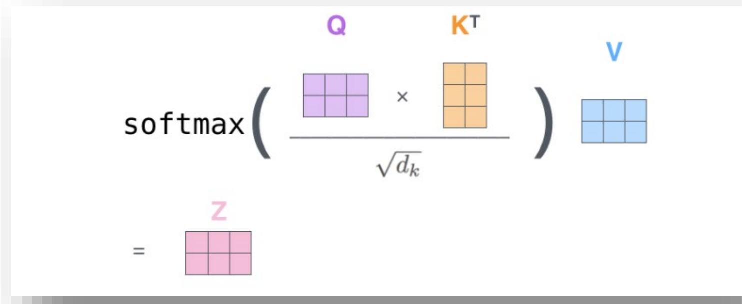
Self-Attention: Implementation

- Input sequence can be represented as a N x d_{in} matrix
- * denotes matrix multiplication

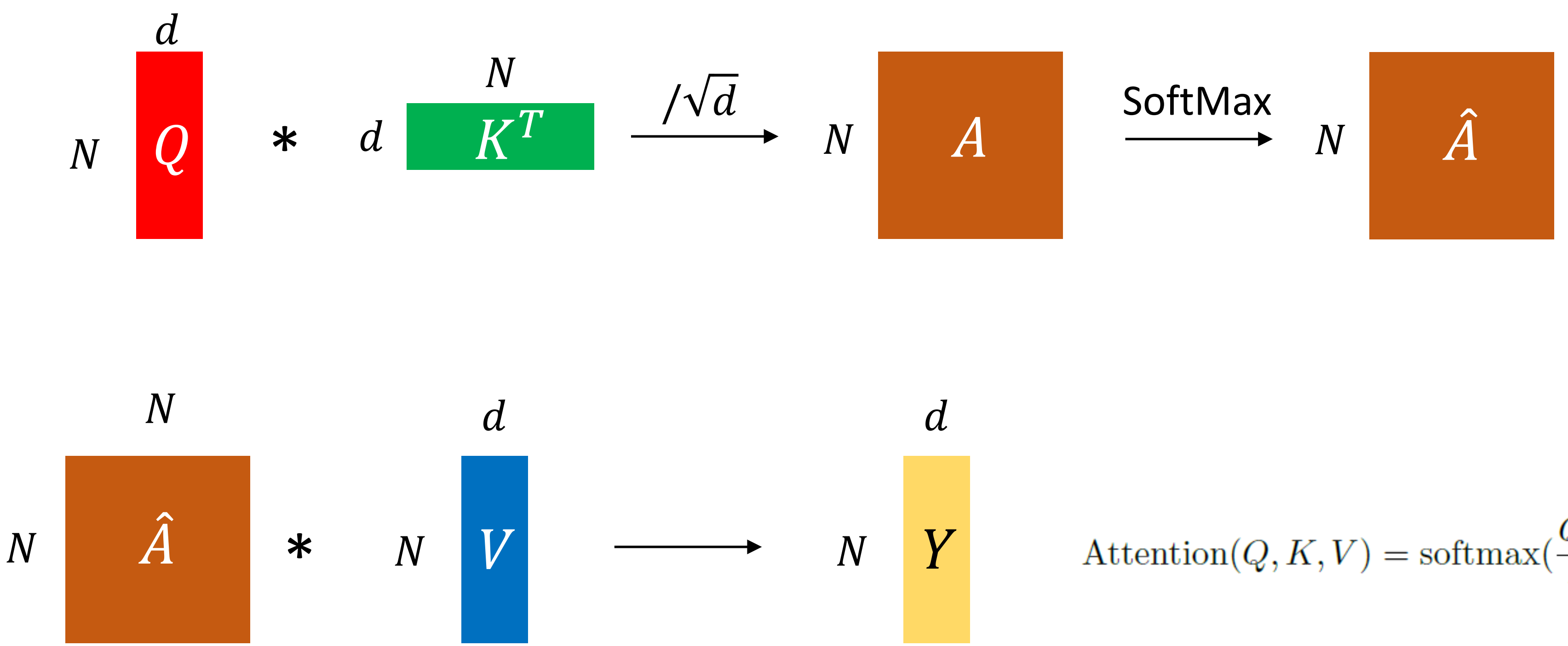




Self-Attention: Implementation



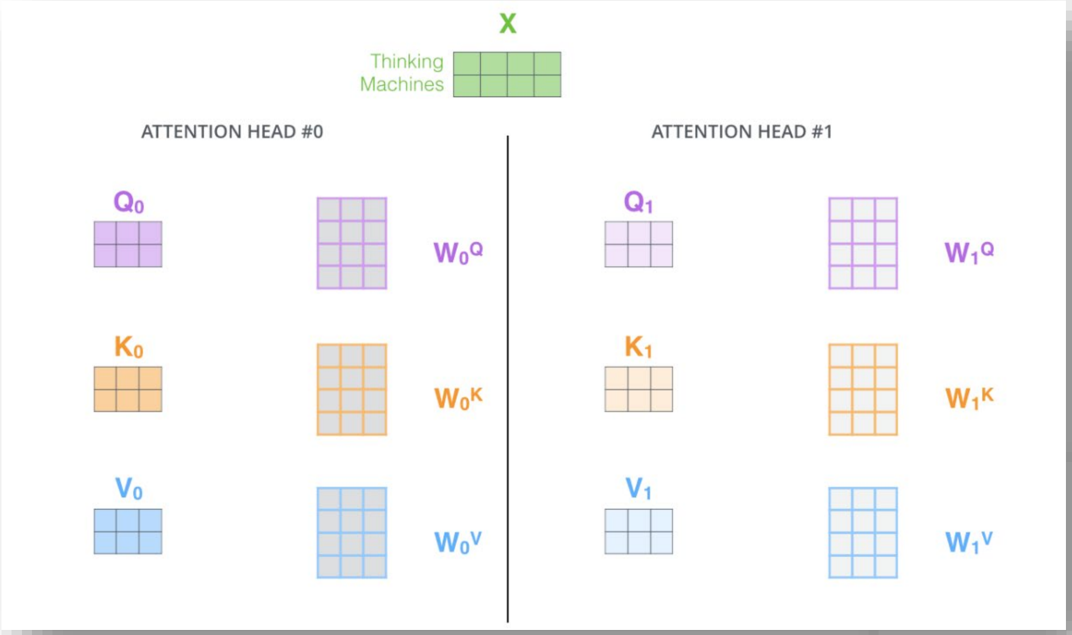
- Output matrix Y
- All operations are **matrix multiplication**, can be parallelized on GPU.

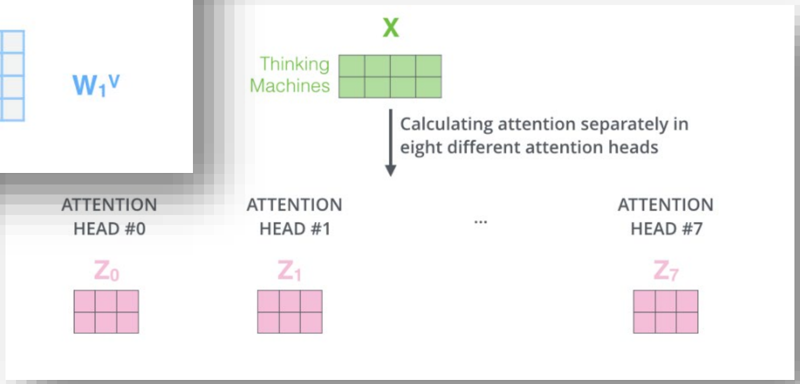


$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multi-Head Self-Attention (1/4)

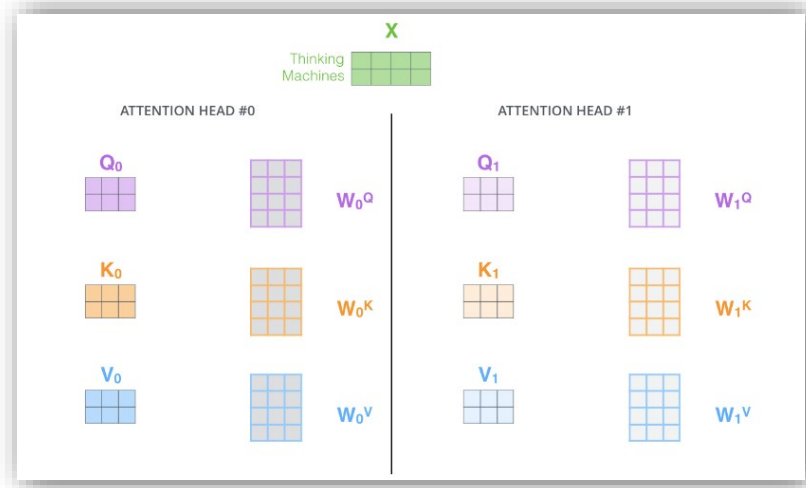
- Perform self-attention at different subspaces, implying attention over different input feature types (e.g., representations, modalities, positions, etc.)

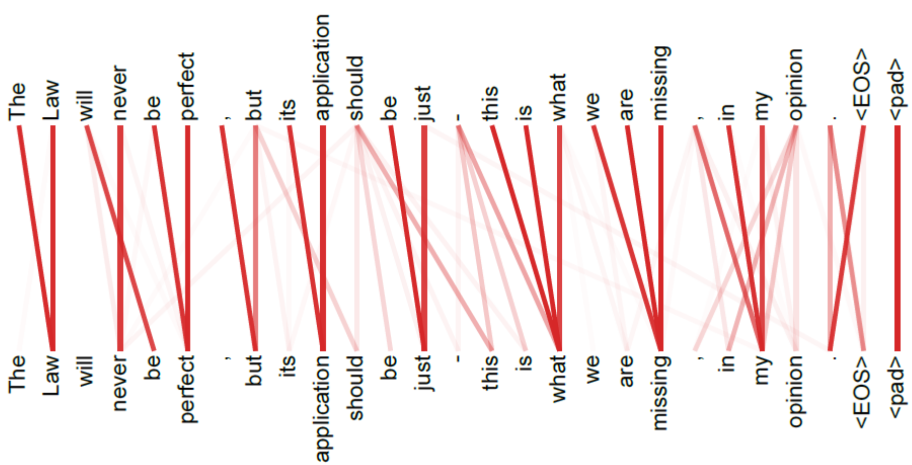




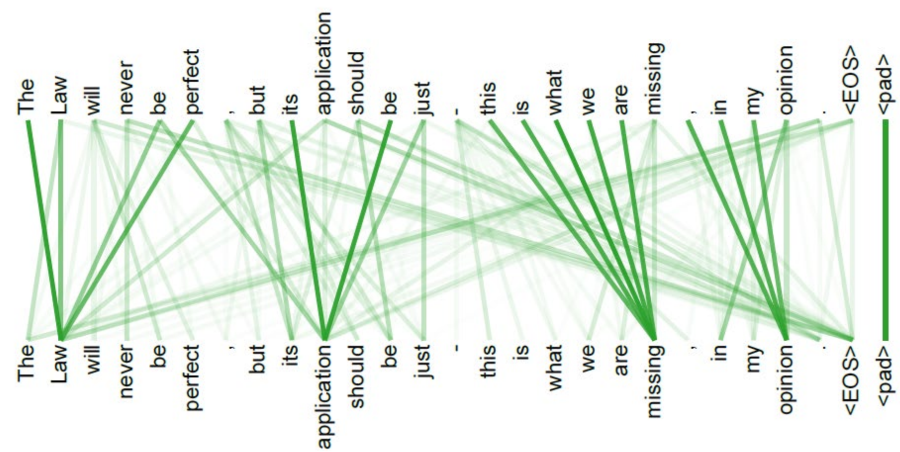
Multi-Head Self-Attention (2/4)

- Perform self-attention at different subspaces, implying attention over different input types
- See example below





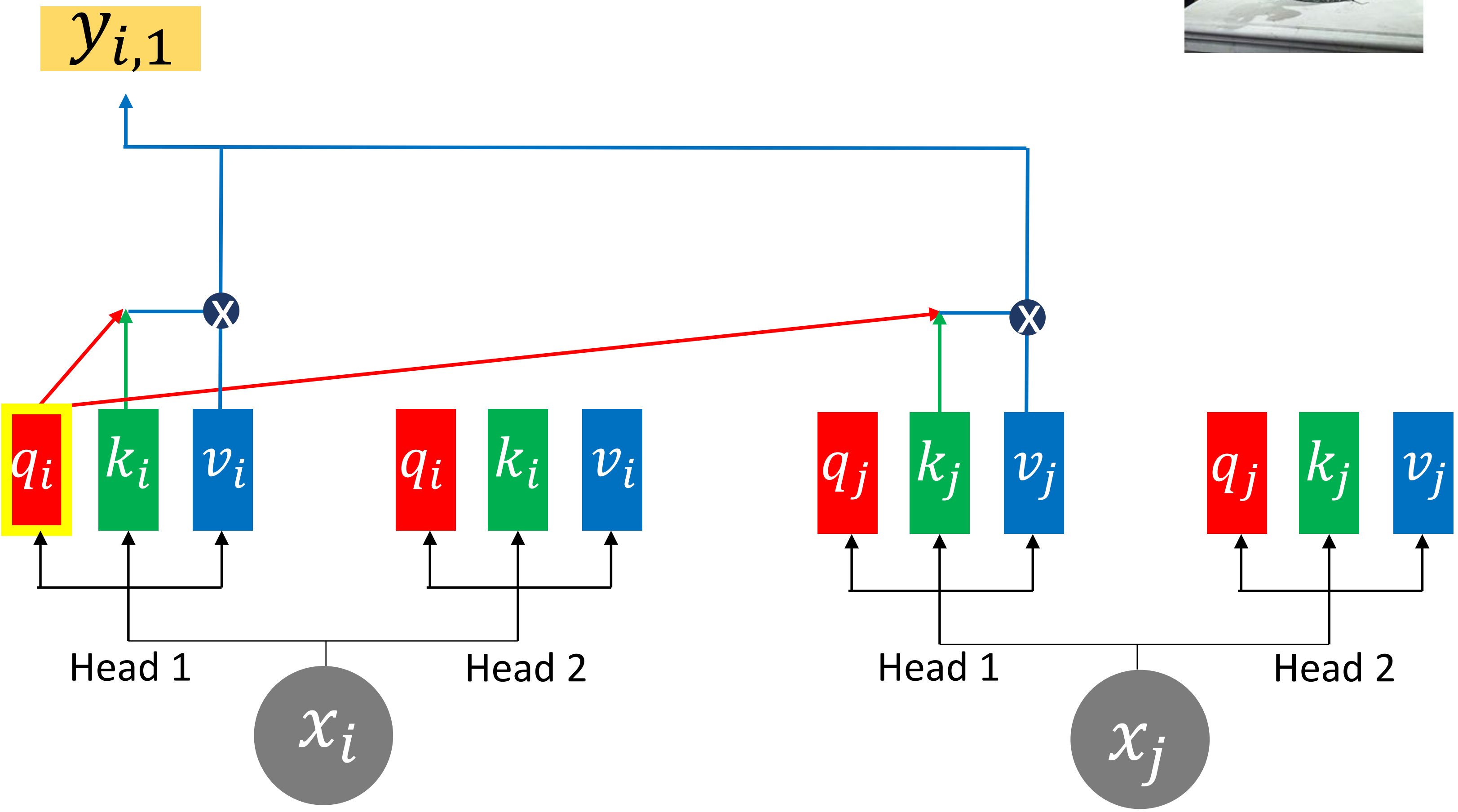
Attention weights of Head 1



Attention weights of Head 2

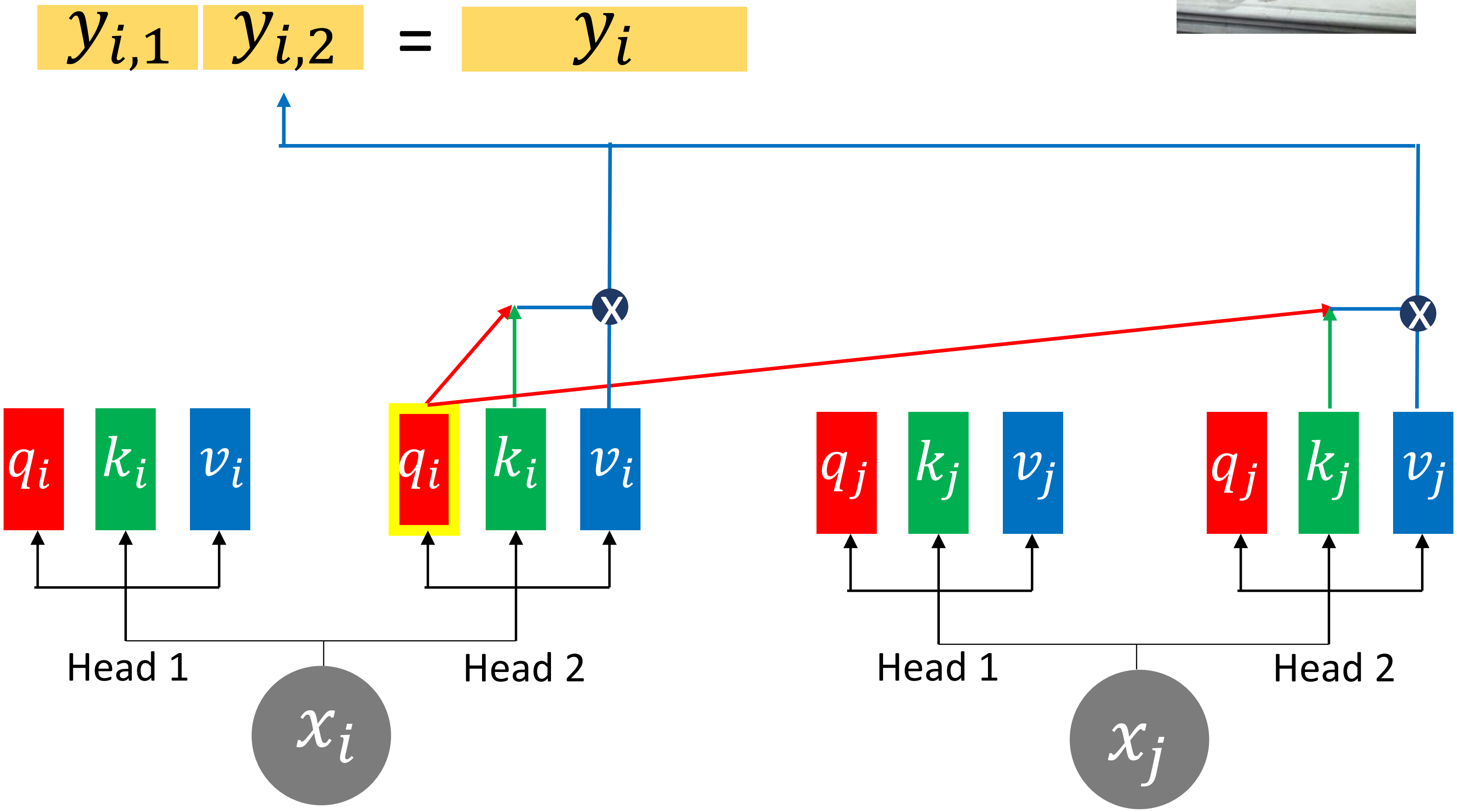
Multi-Head Self-Attention (3/4)

- A 2-head example, output of two heads are concatenated.

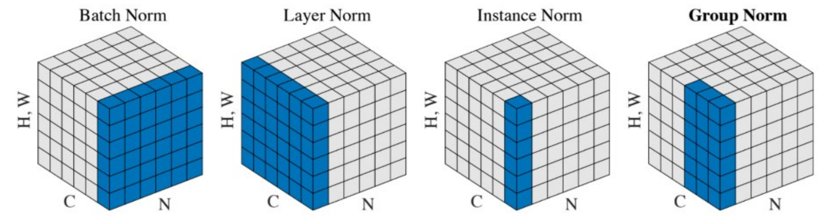


Multi-Head Self-Attention (4/4)

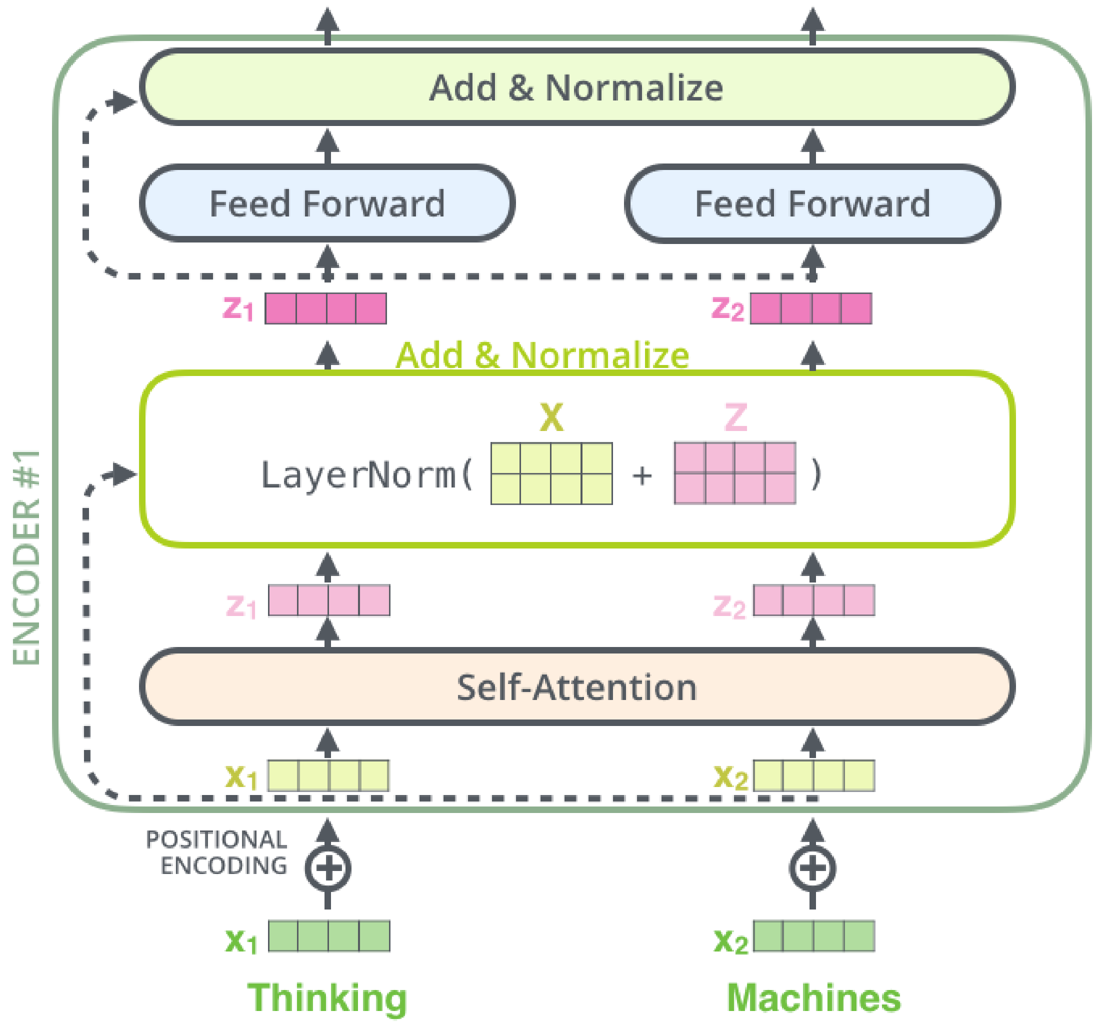
- A 2-head example, output of two heads are concatenated.



The Residuals

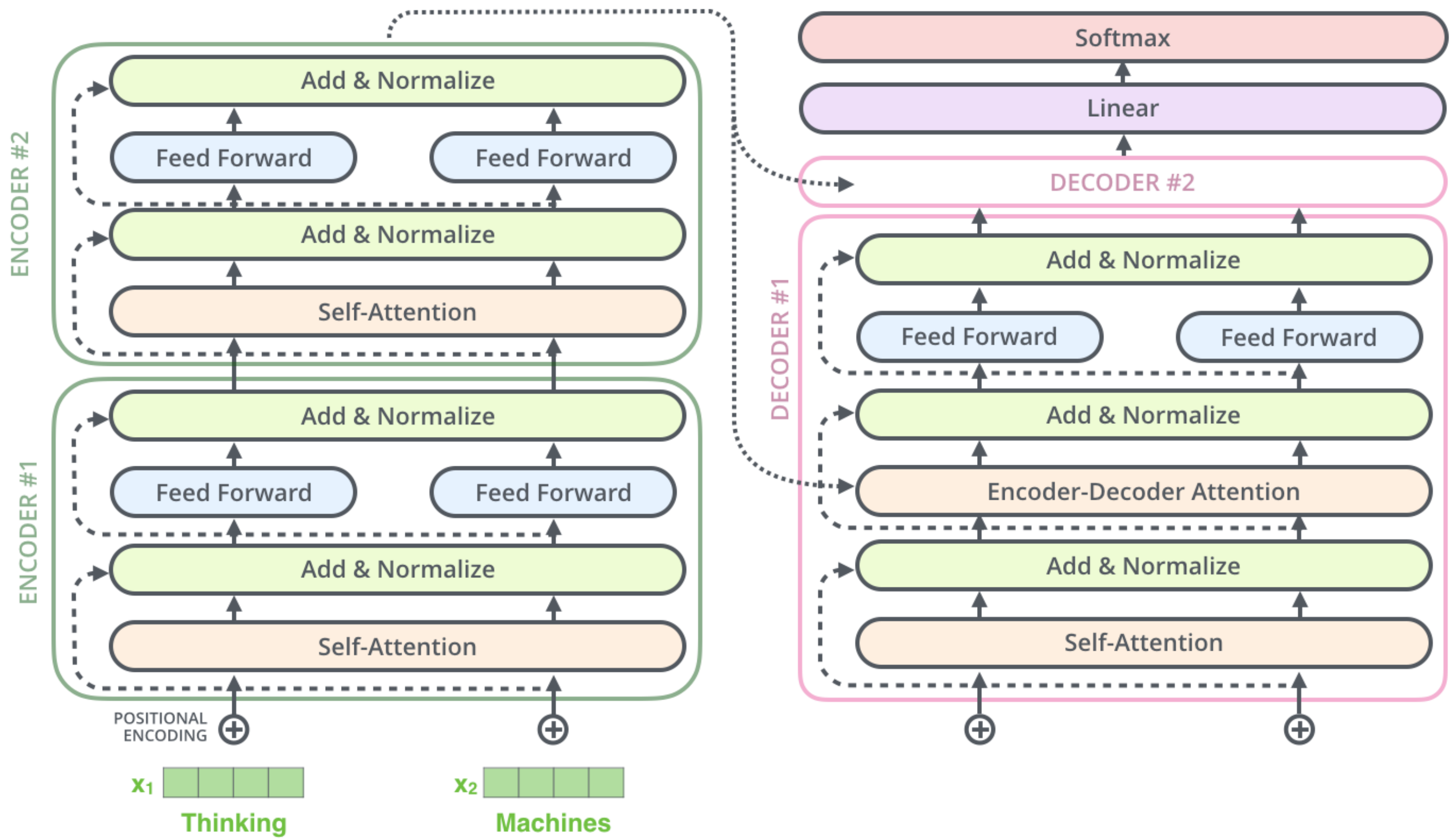


- A residual connection followed by layer normalization



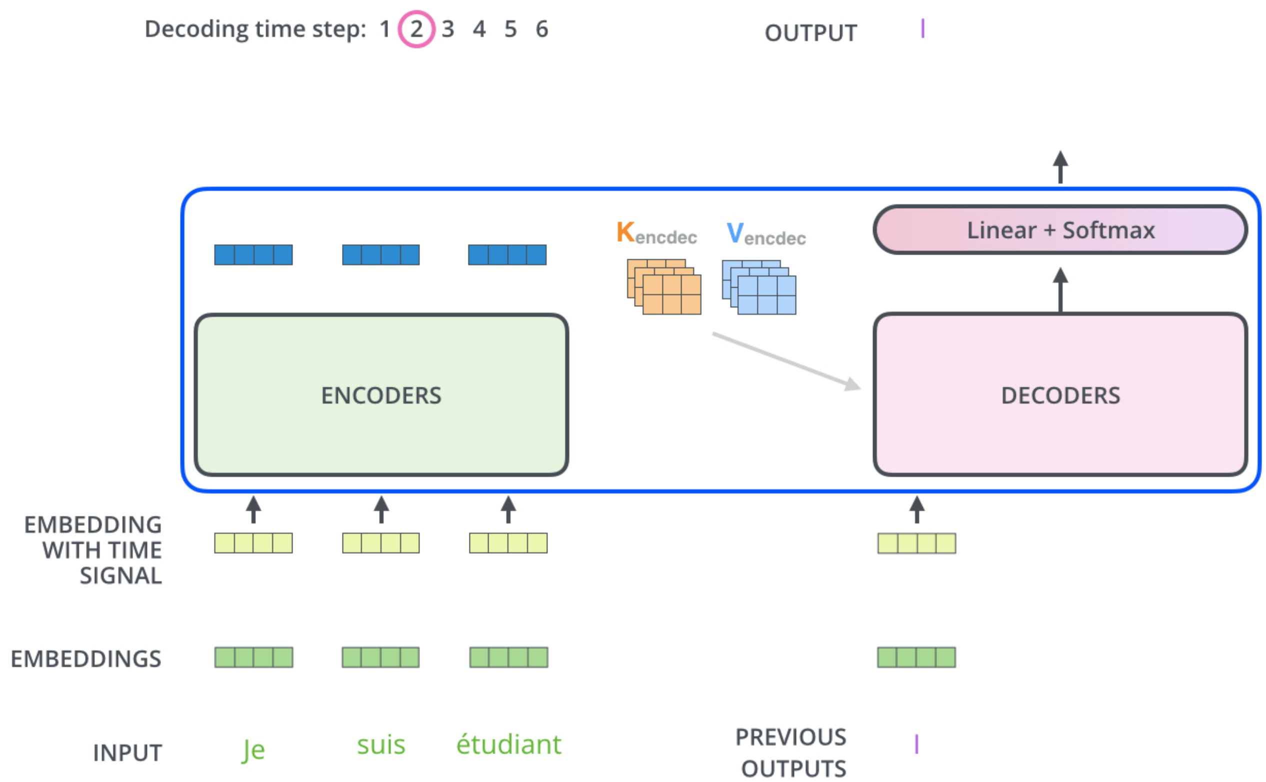
The Decoder in Transformer

- Design similar to that of encoder,
  except the 1st decoder takes additional inputs (of predicted word embeddings).



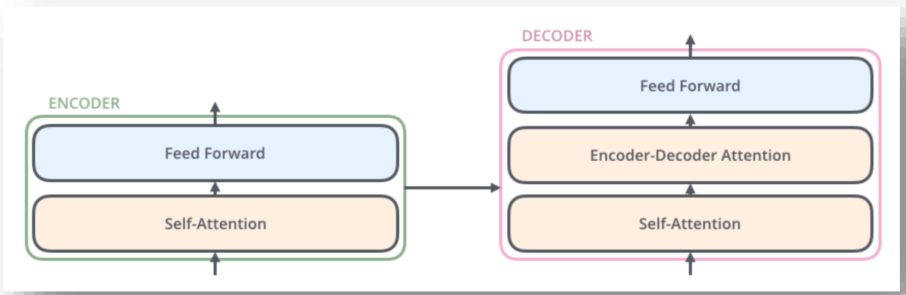
The Decoder in Transformer

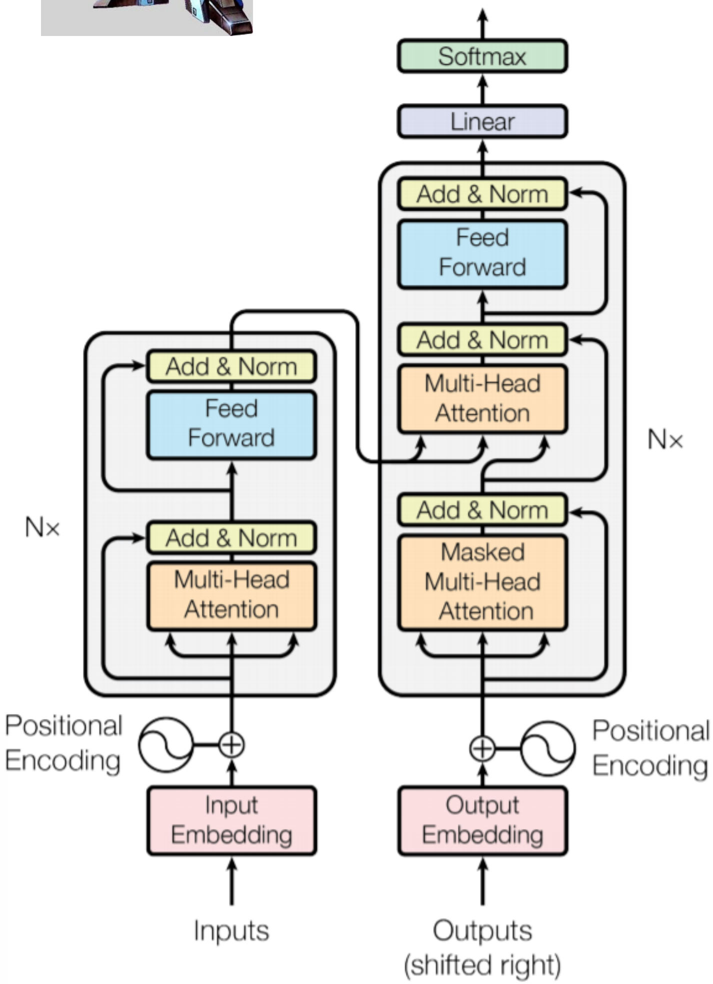
- Design similar to that of encoder,
  except the 1st decoder takes additional inputs (of predicted word embeddings).



Recap: Transformer

- “Attention is all you need”, NeurIPS 2017
- We didn’t cover positional encoding (particularly for language translation)
- More info available at: http://jalammar.github.io/illustrated-transformer/





What to Cover Today…

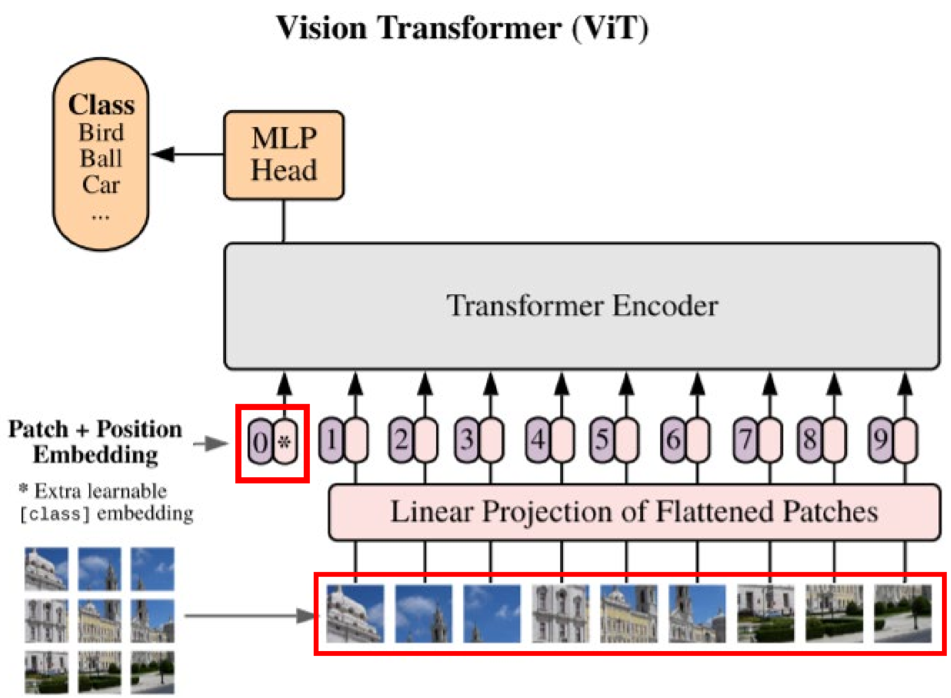
- **Recurrent Neural Network** & **Transformer**
  - Attention in RNN
  - *Attention is All You Need*: Transformer
  - Transformer for Visual Analysis
    - Visual Classification
    - Semantic Segmentation & More
- **Vision & Language**
  - Image Captioning
  - Text-to-Image Synthesis

"a corgi wearing a bow tie and a birthday hat"

Teddy bears shopping for groceries in the style of ukiyo-e

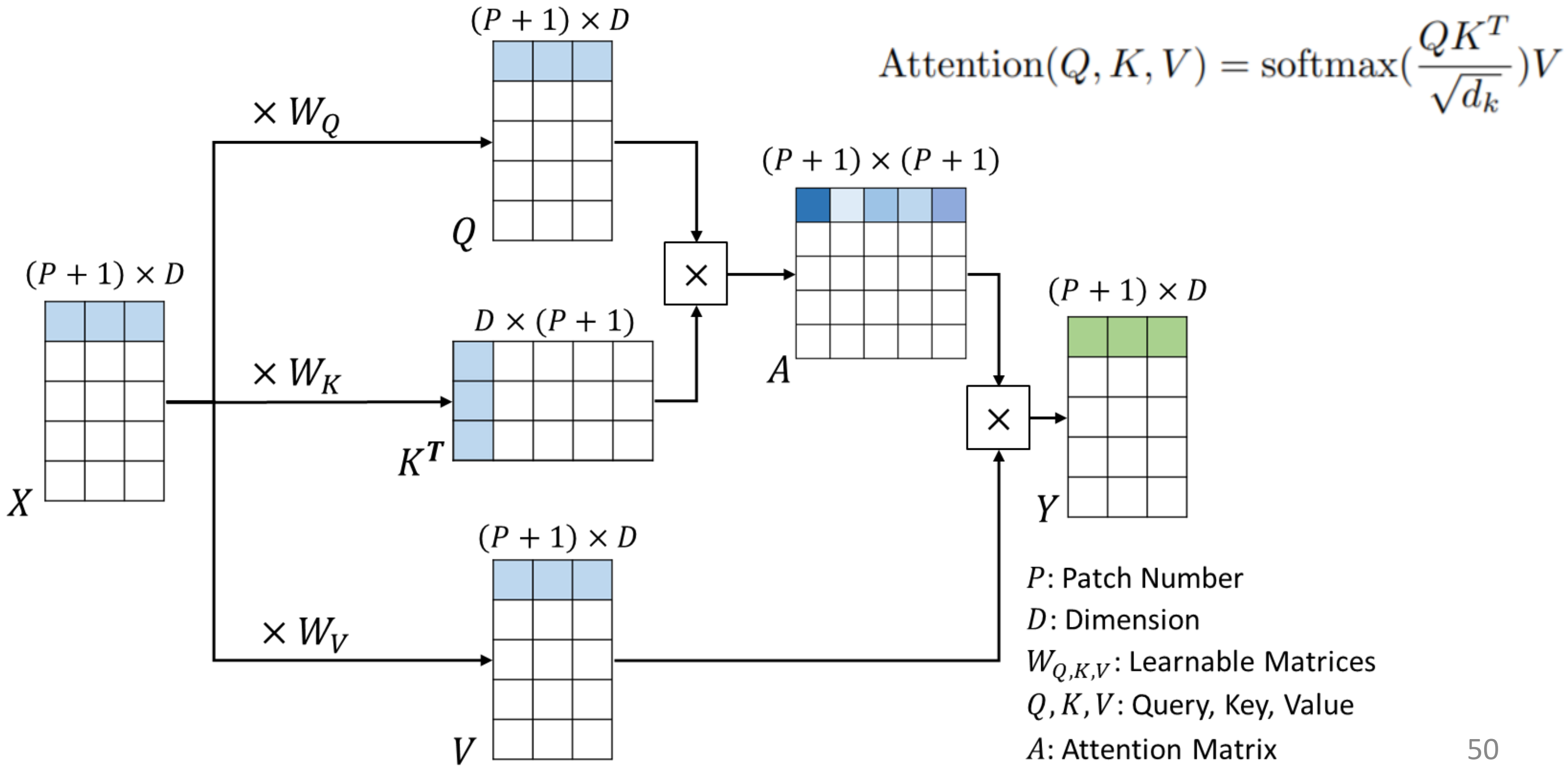
Vision Transformer

- "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR, 2021. (Google Research)
- Partition the input image into a **patch sequence**
- An additional **token** (*) is appended to perform attention on patches
- Both the "*" token and positional embeddings (denoted by 0, 1, 2 …) are **trainable vectors**



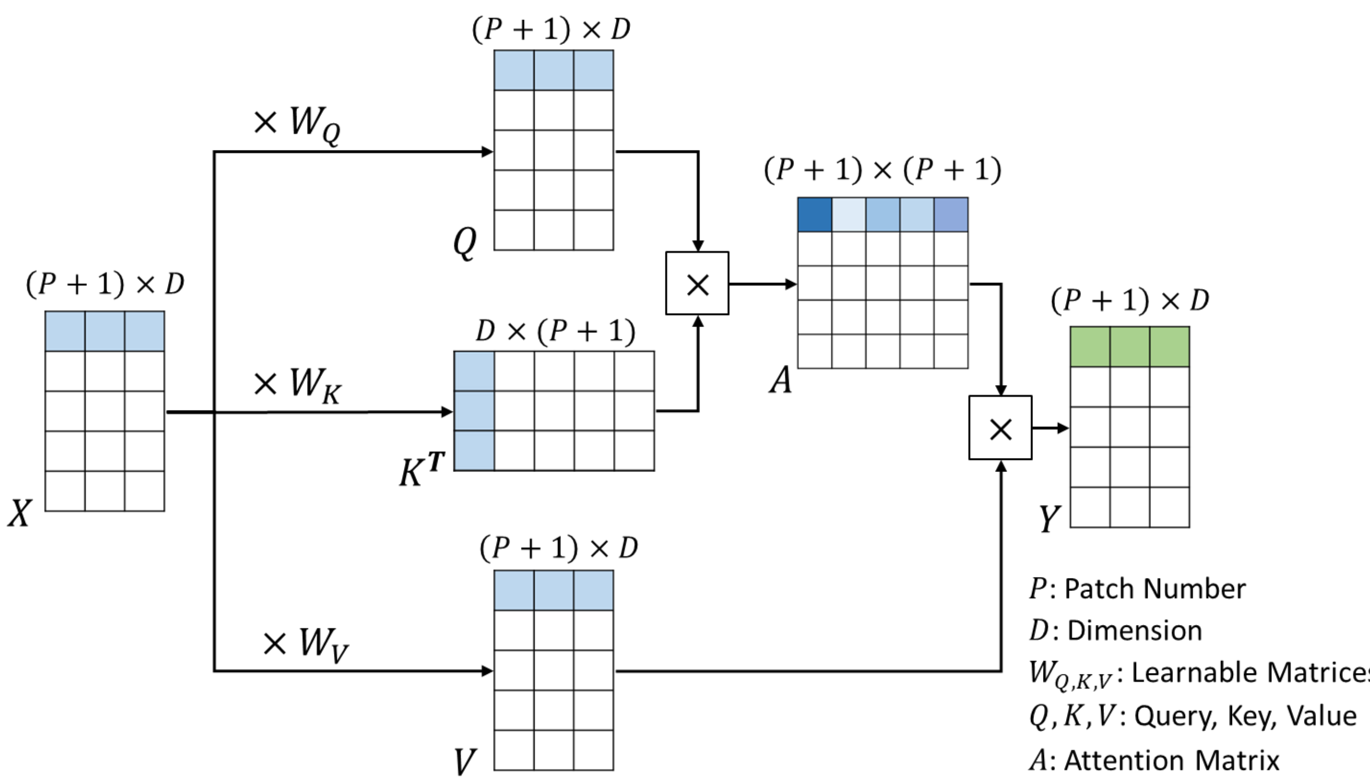
Query-Key-Value Attention in ViT

- Assume that the input is partitioned into 4 patches and the feature dimension is 3, that is, P=4 and D=3
- Note that there are (P+1) rows since we have an additional token



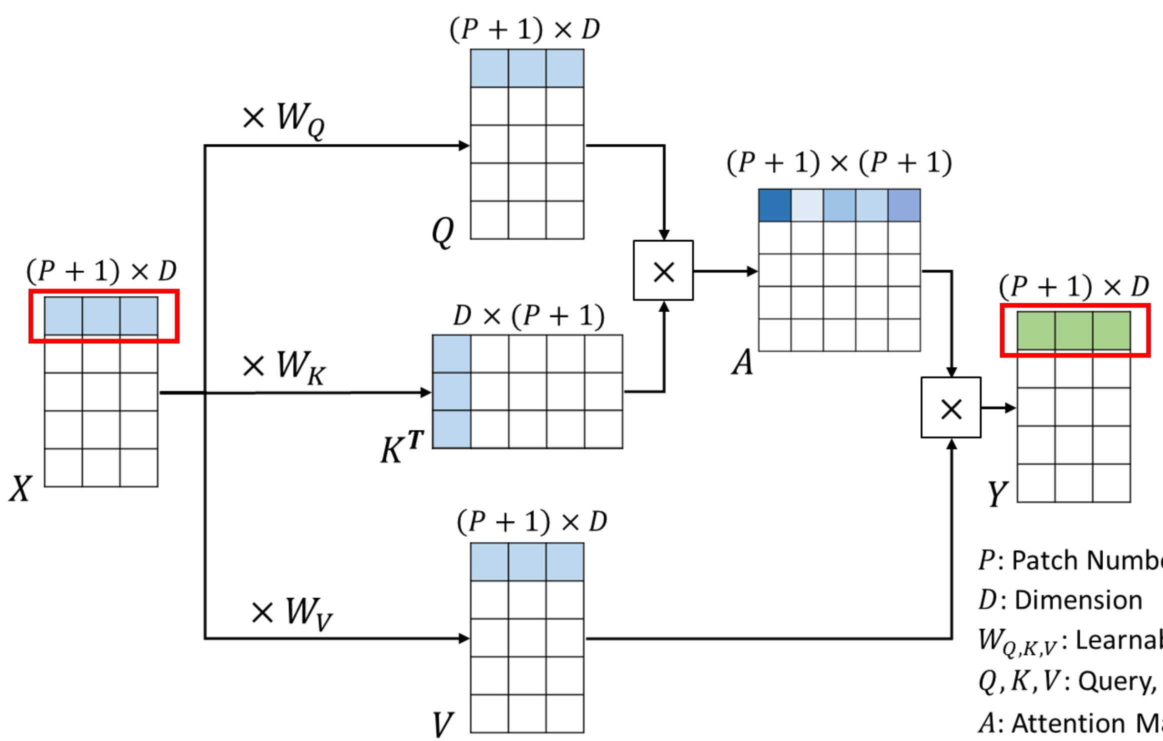
Query-Key-Value Attention in ViT

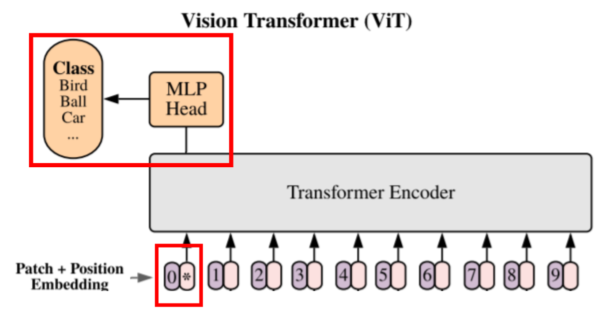
- By performing attention, the input sequence X (of length P+1) is "transformed" into another sequence Y with the same length
- That is why it is called "**Transformer**" and how it is a **seq2seq** model



Query-Key-Value Attention in ViT

- In standard vision transformer, we only take the **first output token** of the output sequence (the **first row** of Y) for classification purposes
- This corresponds to the output when **token “0”** serves as query

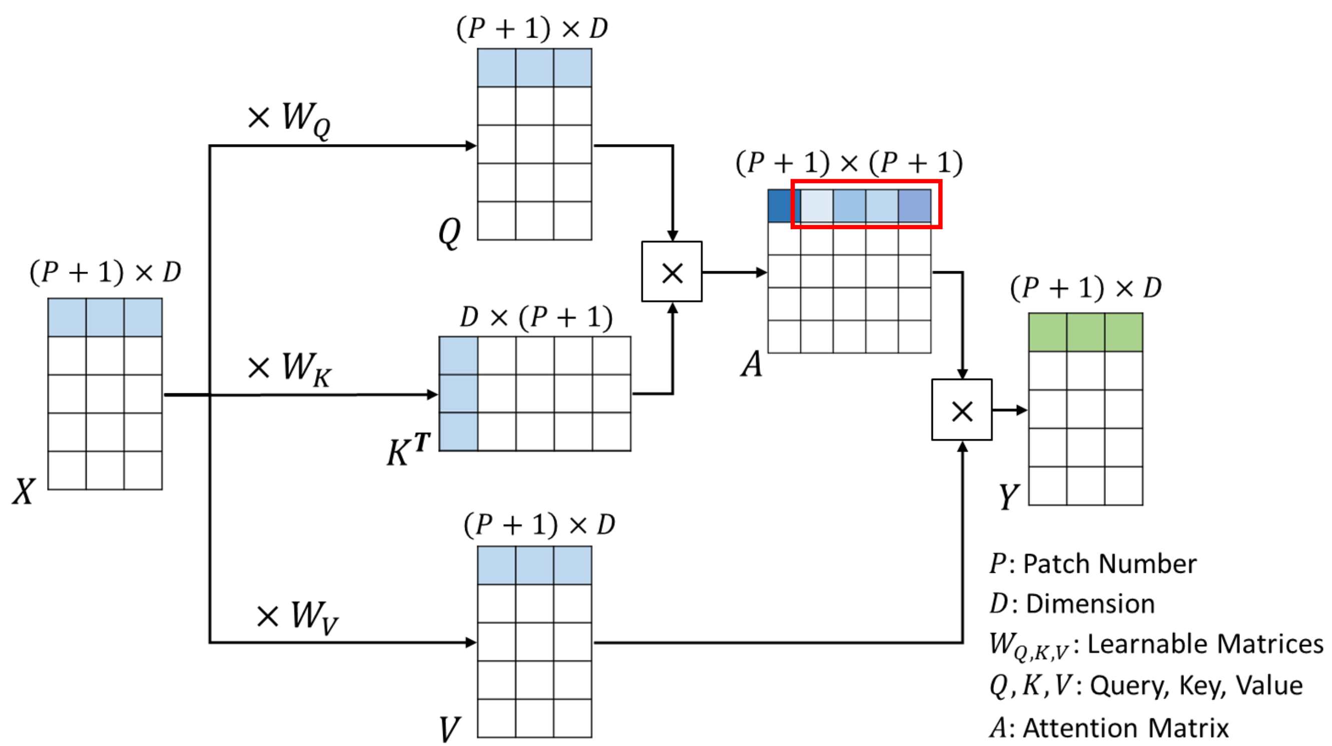




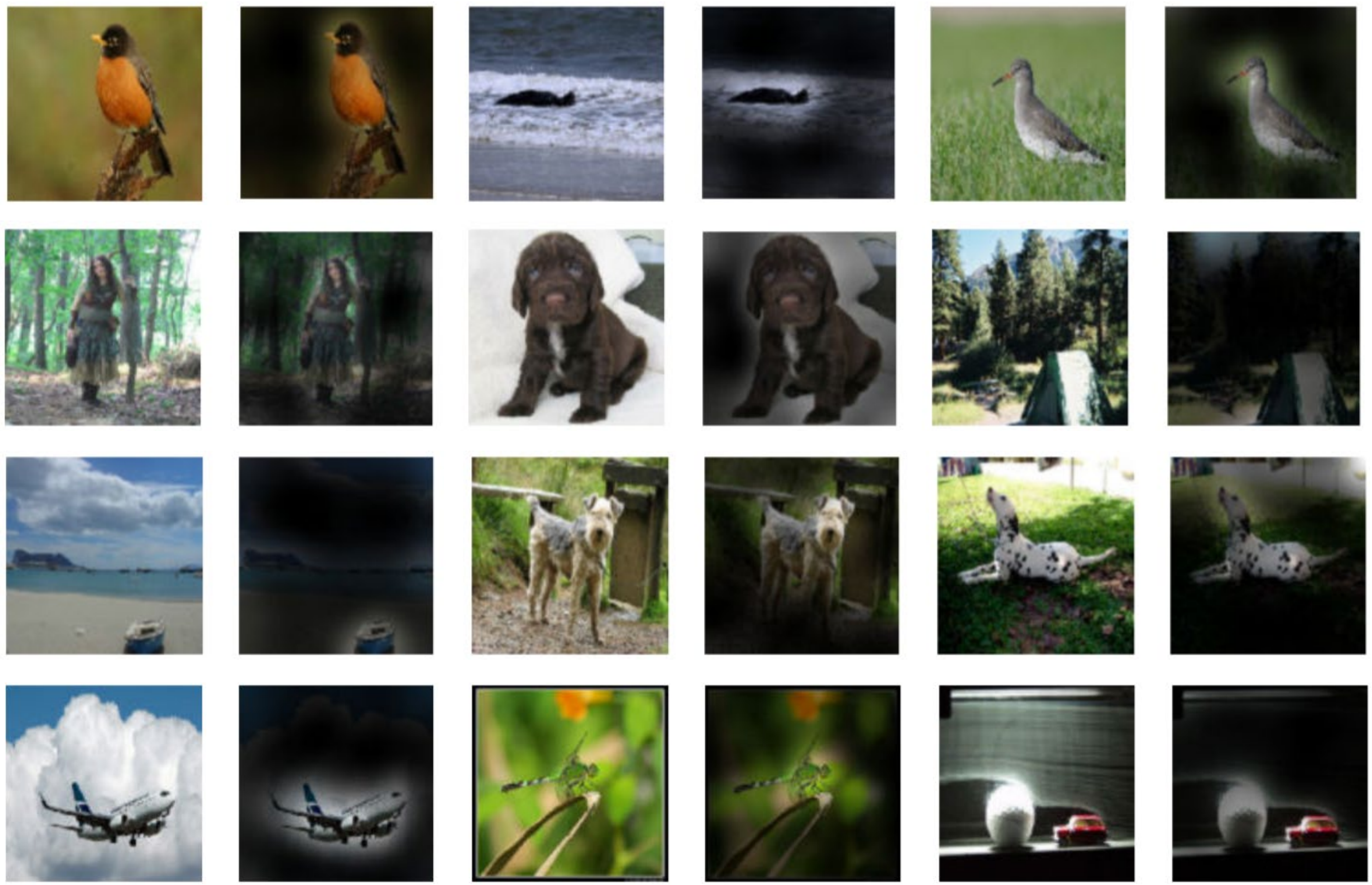
P: Patch Number
D: Dimension
$W_{Q,K,V}$: Learnable Matrices
Q, K, V: Query, Key, Value
A: Attention Matrix

Visualization

- To visualize the attention maps, we take the attention scores from the **first row** of A (when token "0" serves as query)
- Note the first element is excluded, and thus there are **P scores** corresponding to the P image patches

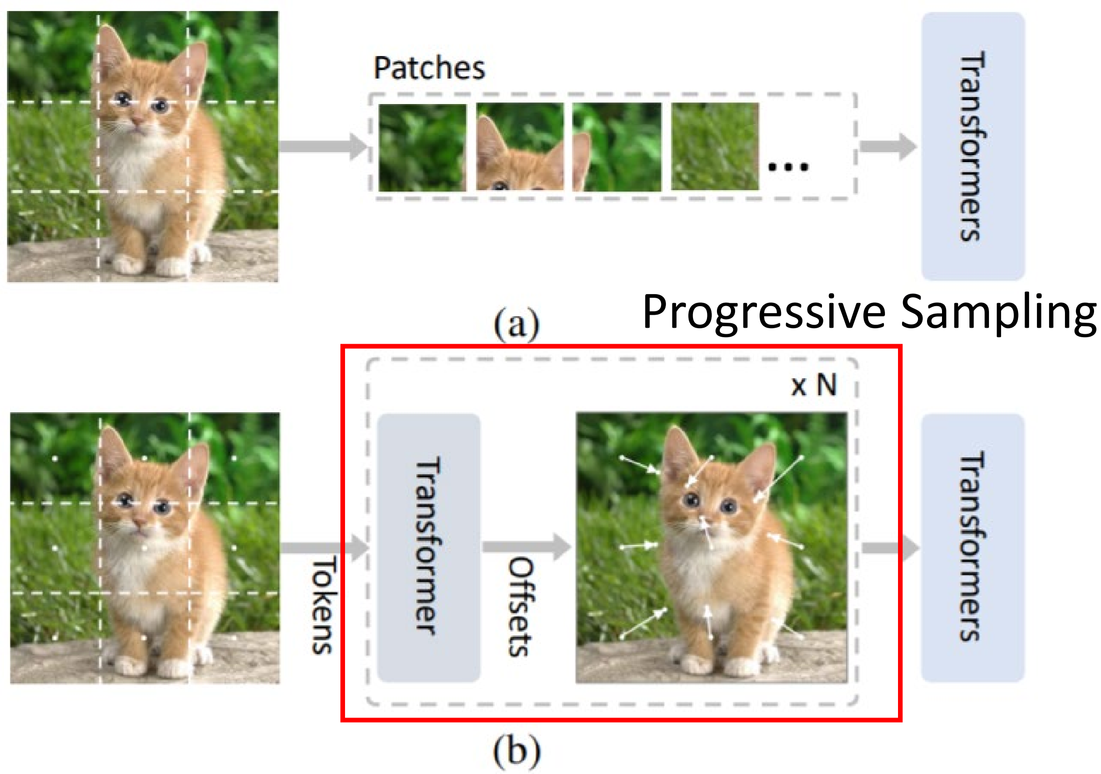


Example Visualization



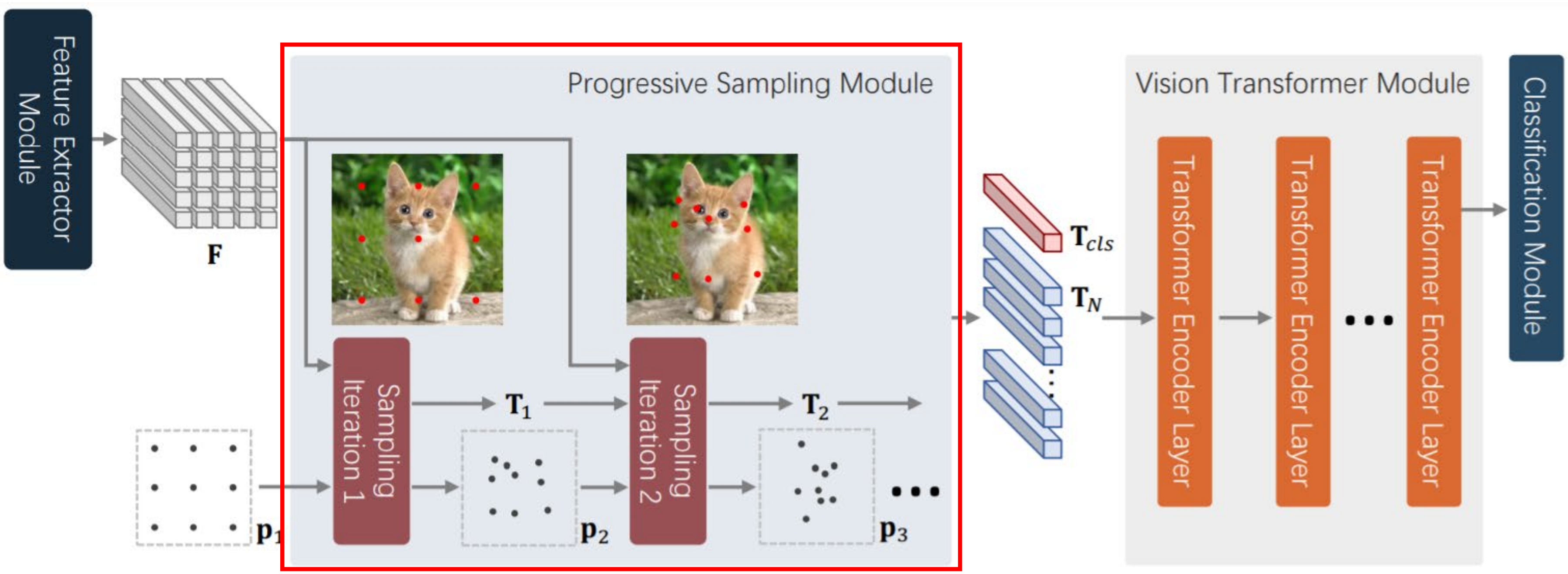
PS-ViT

- Vision Transformers with Progressive Sampling
- Progressively select important patches by shifting patch centers

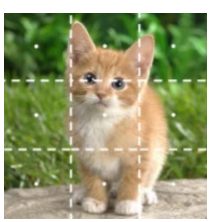


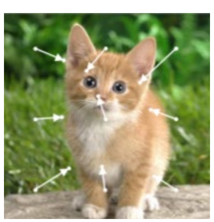
Yue et al. "Vision transformer with progressive sampling." ICCV 2021

PS-ViT (cont'd)



Feature Extractor Module
F
Progressive Sampling Module
Sampling iteration 1
T_1
Sampling iteration 2
T_2
p_1
p_2
p_3
T_{cls}
T_N
Vision Transformer Module
Transformer Encoder Layer
Transformer Encoder Layer
Transformer Encoder Layer
Classification Module

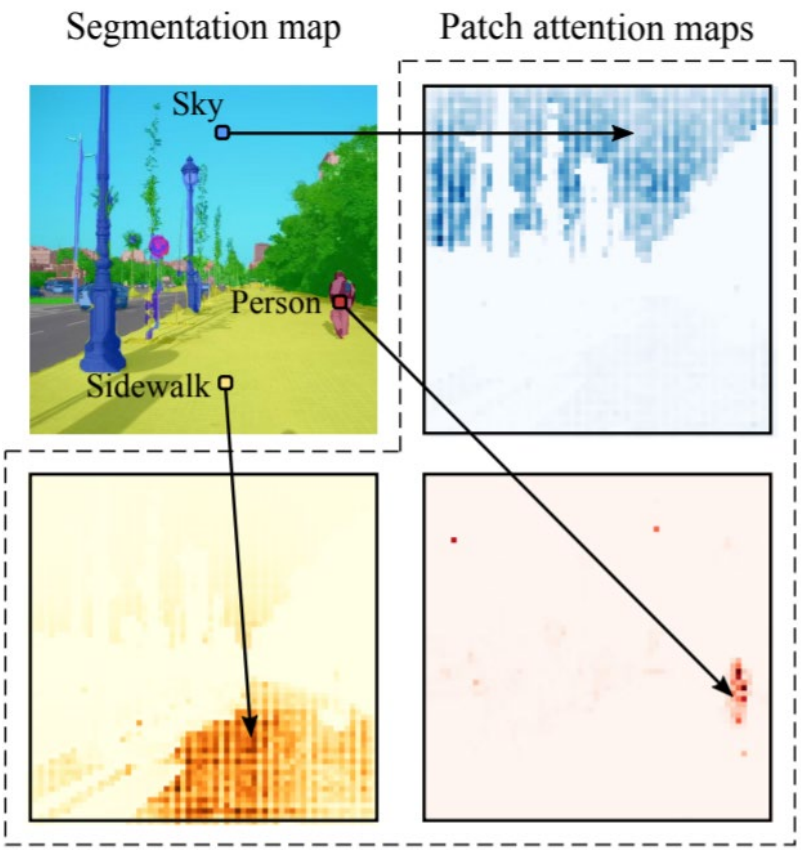




Example Visualization

Transformer for Semantic Segmentation

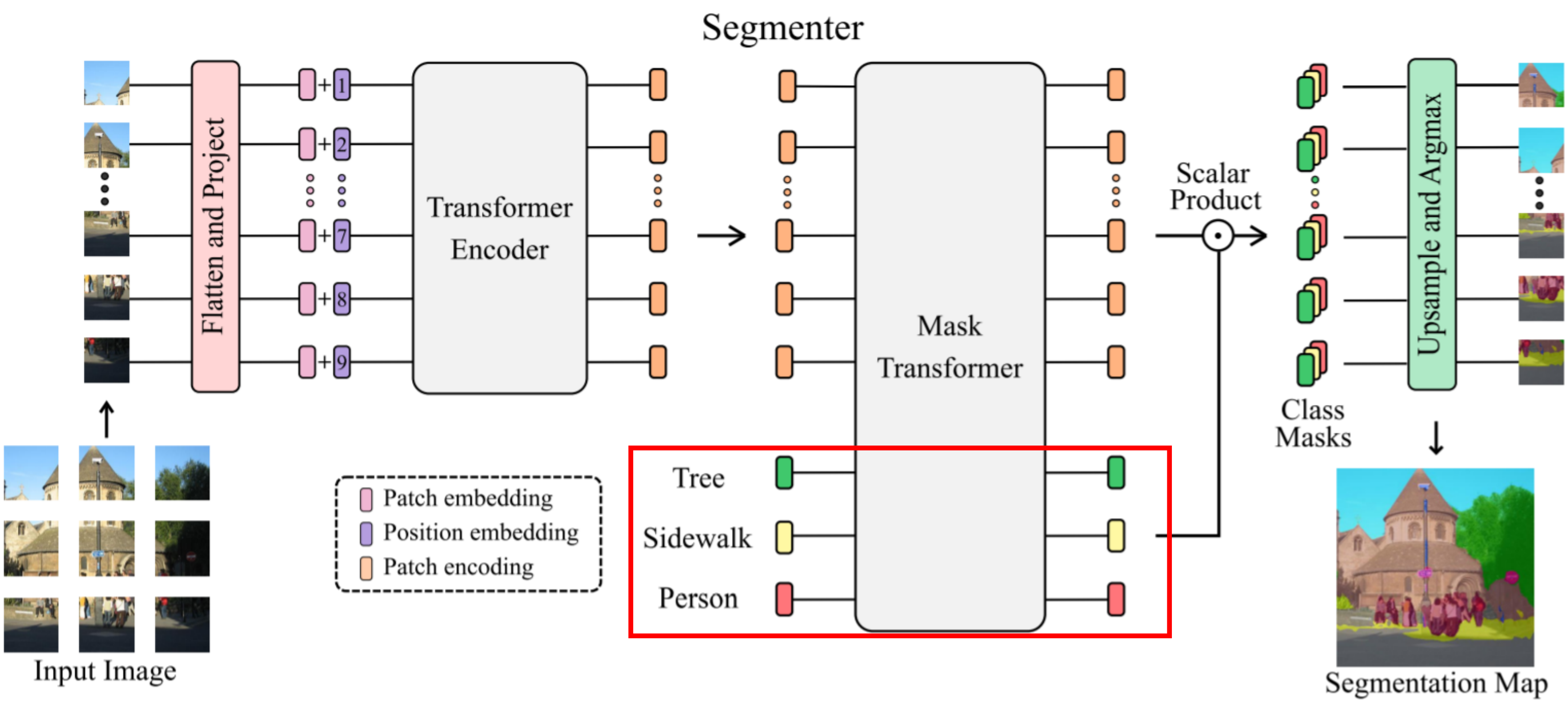
- Segmentation via attention



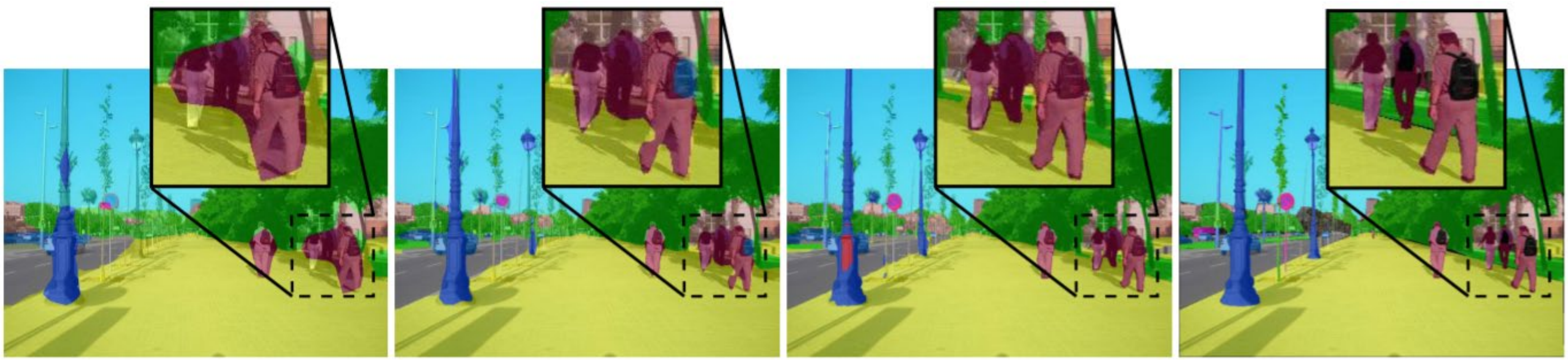
Strudel et al. "Segmenter: Transformer for Semantic Segmentation." ICCV 2021

Transformer for Semantic Segmentation

- Using different class tokens (“Tree”, “Sidewalk”, “Person”, …) as queries



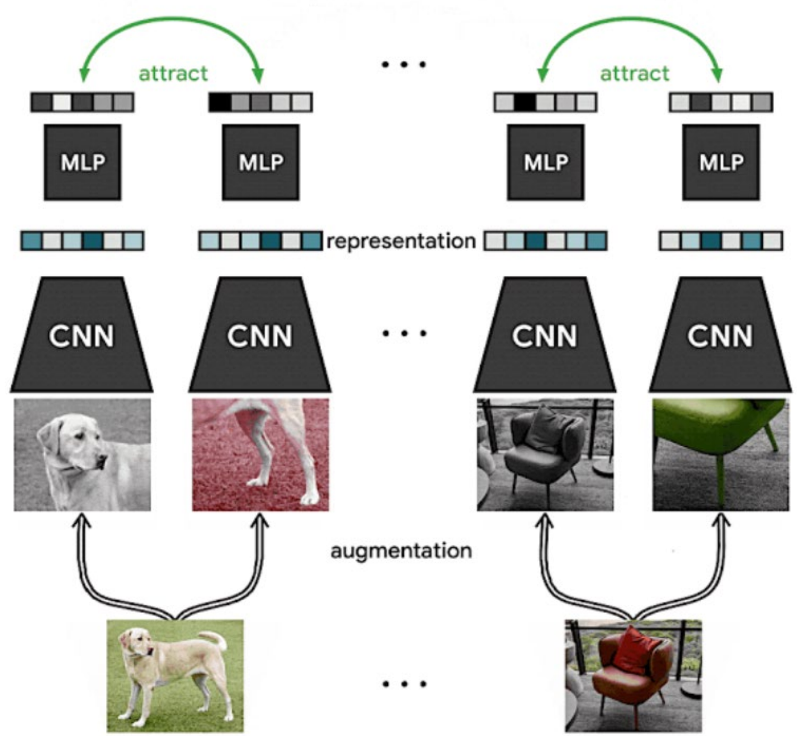
Example Visualization



(a) Patch size 32 × 32 (b) Patch size 16 × 16 (c) Patch size 8 × 8 (d) Ground Truth

Self-Supervised Learning (SSL) for Transformer

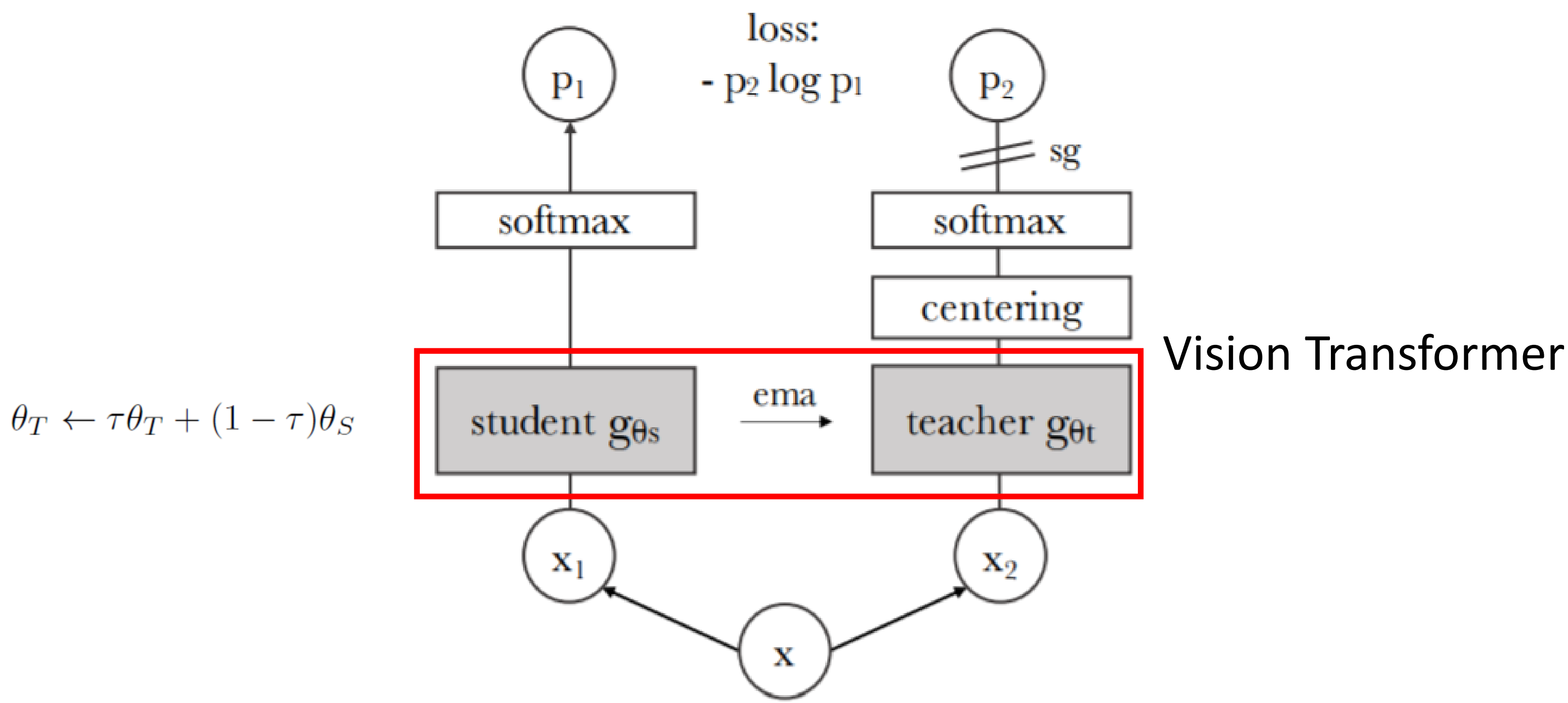
- Learning discriminative representations from **unlabeled** data
- Create self-supervised tasks via **data augmentation**



Chen et al. "A simple framework for contrastive learning of visual representations." ICML 2020

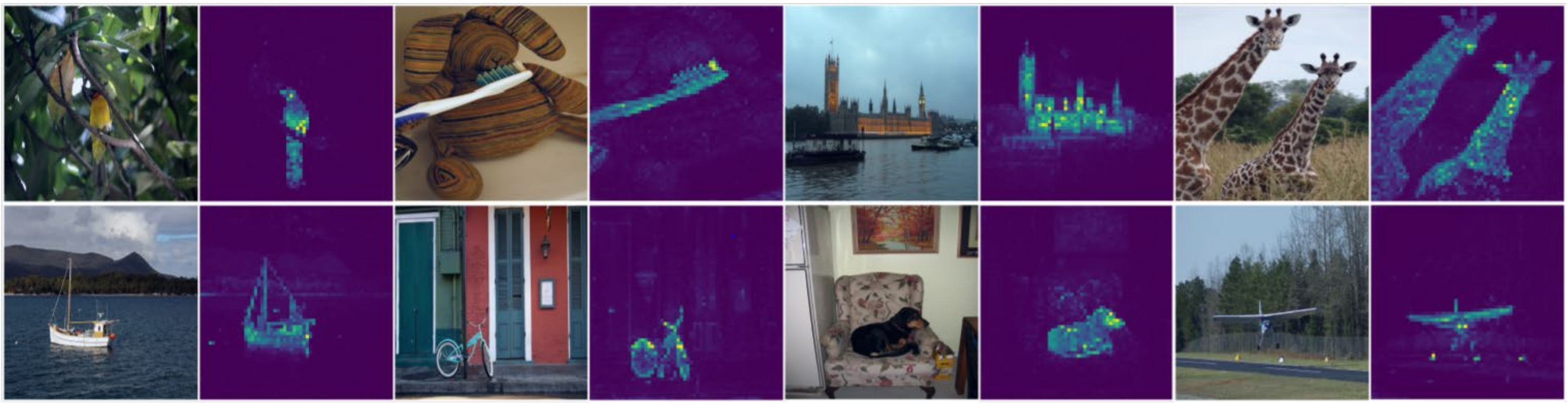
Self-Supervised Transformer

- Vision Transformer + **SSL**
- Maximize the similarity between the augmented version and itself
- Avoid collapse with **student-teacher** network



Caron et al. “Emerging properties in self-supervised vision transformers.” ICCV 2021

Qualitative & Quantitative Results



Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**

What to Cover Today…

- **Recurrent Neural Network** & **Transformer**
  - Attention in RNN
  - *Attention is All You Need*: Transformer
  - Transformer for Visual Analysis
    - Visual Classification
    - Semantic Segmentation & More
- **Vision & Language**
  - Image Captioning
  - Text-to-Image Synthesis

"a corgi wearing a bow tie and a birthday hat"

Teddy bears shopping for groceries in the style of ukiyo-e

A picture is worth a thousand words…
Is it that simple?

Vision + Language → ?

- Image Captioning
- Image Manipulation/Completion
- Composed Image Retrieval
- Visual Question Answering (VQA)
  and many more…

Image Captioning

Applications: semantics understanding, image-text retrieval, medical AI, etc.

Image Captioning (cont'd)

- Training a captioning model requires a large amount of image-caption data pairs
- Image captioning in the wild:
  - Describing images with novel content during inference
  - For example, COCO dataset has 80 object categories.
    How to generalize captioning models to Open Image (w/ 600 classes)?
- Domain-specific image captioning:
  - From general-purpose captioning to task-oriented captioning

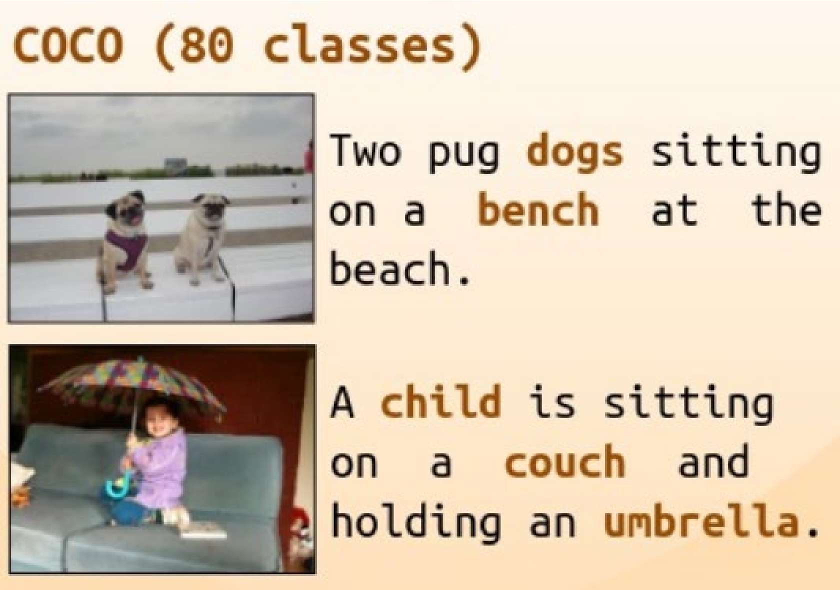
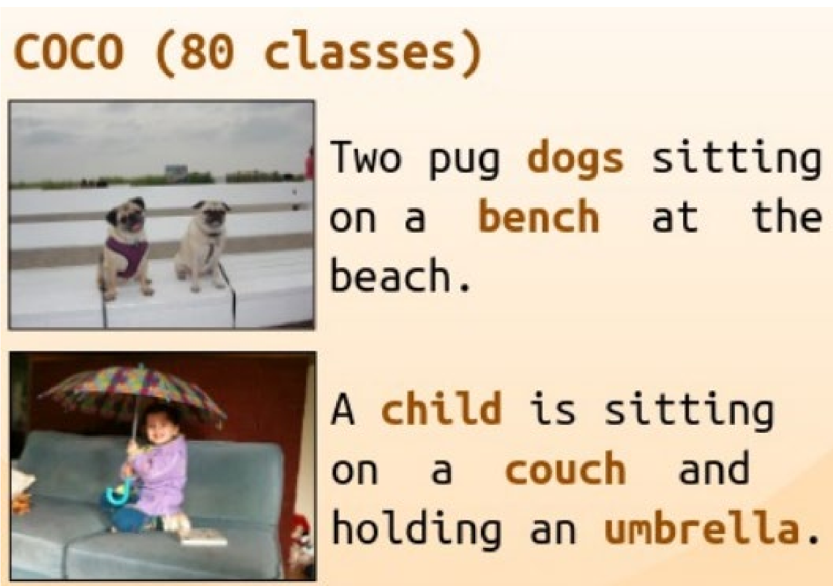


Image Captioning *in the Wild*

- **Novel Object Captioning (NOC)**
  - Training with captioned and uncaptioned data
    captioned data: labeled image data with captions (e.g., COCO)
    uncaptioned data: only labels of novel classes available (e.g., Open Images)

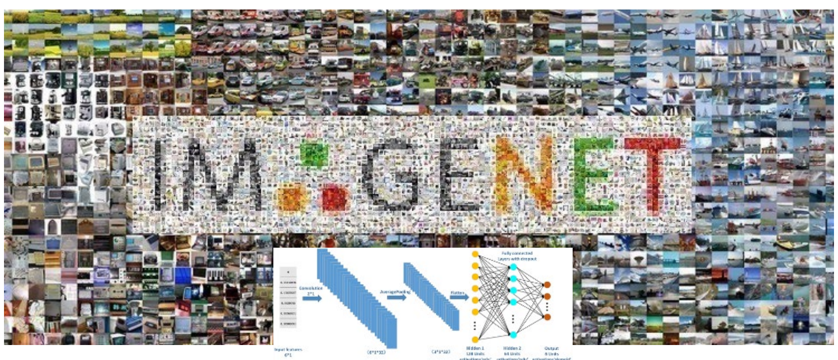


We have captioning data

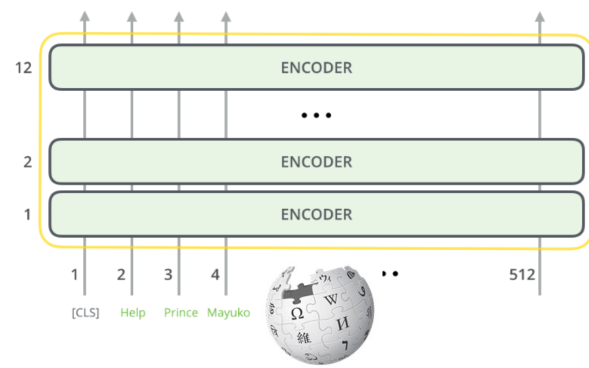
Data with labels for novel objects but w/o captions

Novel Object Captioning

- **VIVO**: **Vi**sual **Vo**cabulary Pre-Training for Novel Object Caption Captioning (AAAI'21)
  - Pre-training a cross-modality Transformer for vision & language tasks
  - Pre-training really matters, since it's been observed in
    - Computer Vision (e.g., models pre-trained on ImageNet)
    - Natural Language Processing (e.g., BERT pre-trained on Wikipedia)



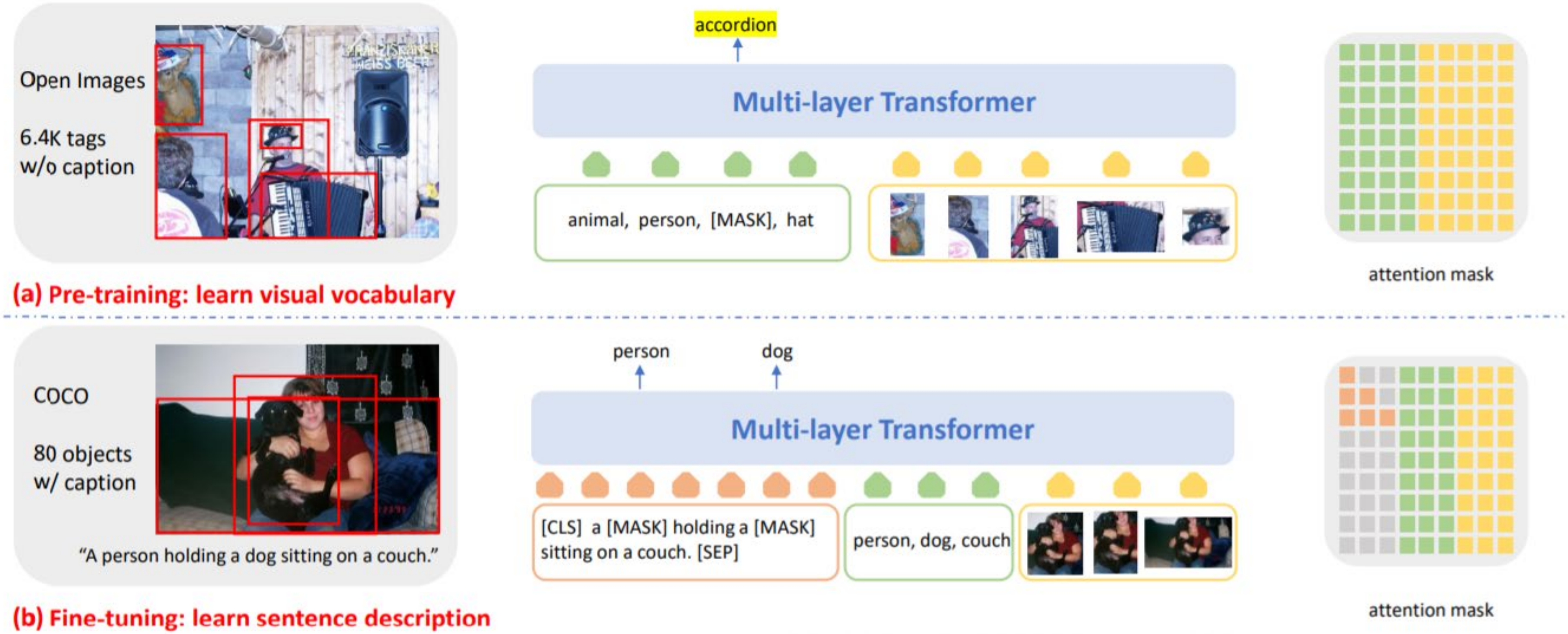
Object detection,
semantic segmentation, etc.



Question answering,
Sentence classification, etc.

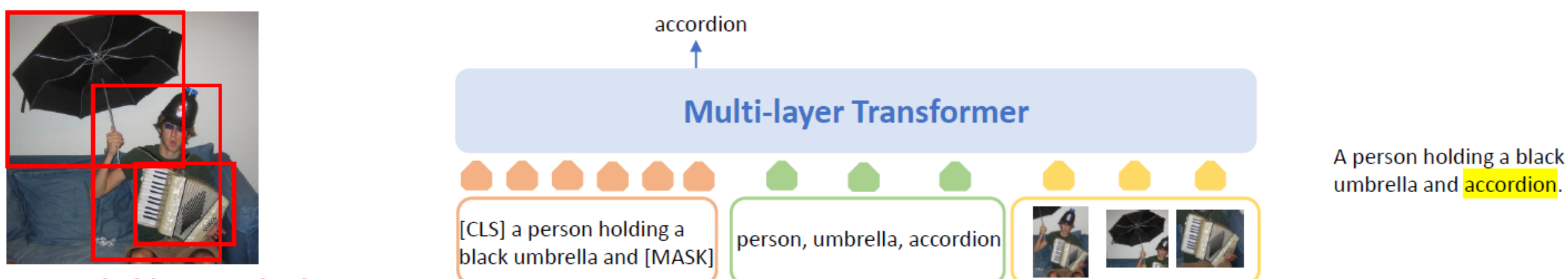
Novel Object Captioning (cont'd)

- **VIVO**: **Vi**sual **Vo**cabulary Pre-Training for Novel Object Caption Captioning
  - Pre-training: uncaptioned image data containing novel class labels
  - Fine-tuning: (a limited amount of) image data with class labels & descriptions



Novel Object Captioning (cont'd)

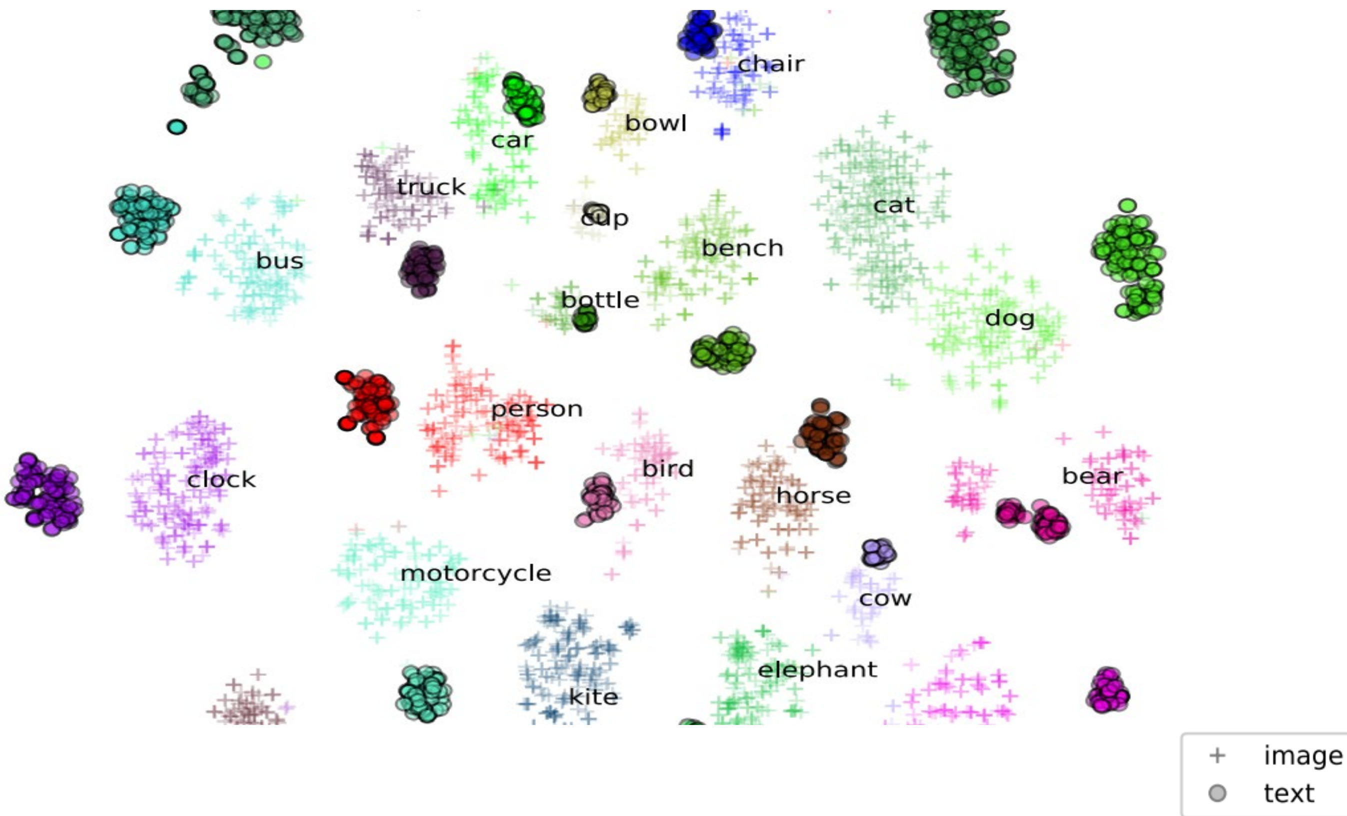
- **VIVO**: **Vi**sual **Vo**cabulary Pre-Training for Novel Object Caption Captioning
  - Pre-training: uncaptioned image data containing novel class labels
  - Fine-tuning: (a limited amount of) image data with class labels & descriptions
  - Inference:
    - Inputs: image (with region features & tags) & [CLS]
    - Output: caption



(c) Inference: novel object captioning

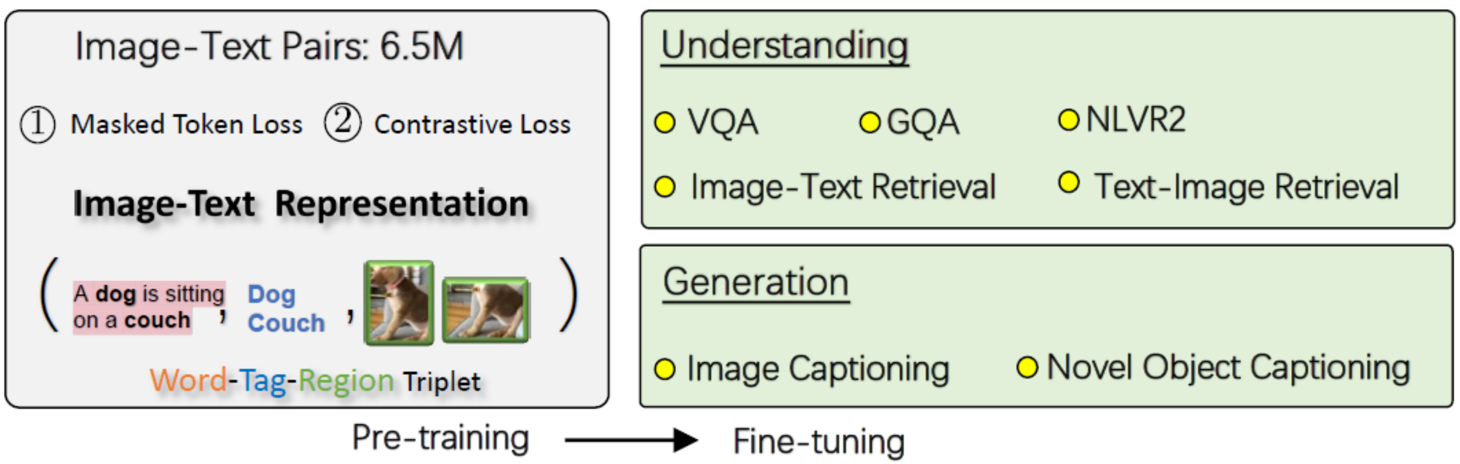
Novel Object Captioning (cont'd)

- VIVO: Visual Vocabulary Pre-Training for Novel Object Caption Captioning
  - Properly aligned image and text data for captioning



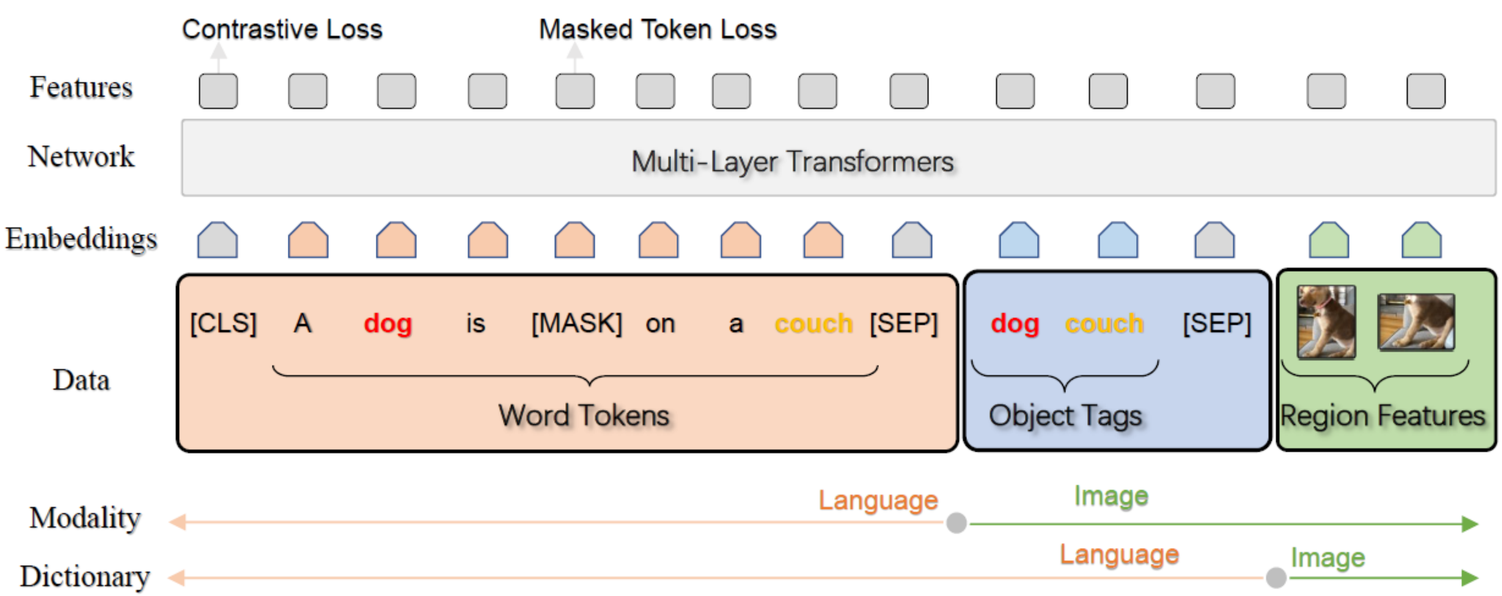
Beyond Image Captioning: Unified Vision & Language Model

- **Oscar**: **O**bject-**S**emanti**c**s **A**ligned P**r**e-training for Vision-Language Tasks (ECCV'20)
  - Training data:
    triplets of caption-tag-region
  - Objectives:
    1. Masked token loss for words & tags
    2. Contrastive loss tags and others
  - Fine-tuning:
    5 vision & language tasks (VQA, image-text retrieval, image captioning, NOC, etc.)



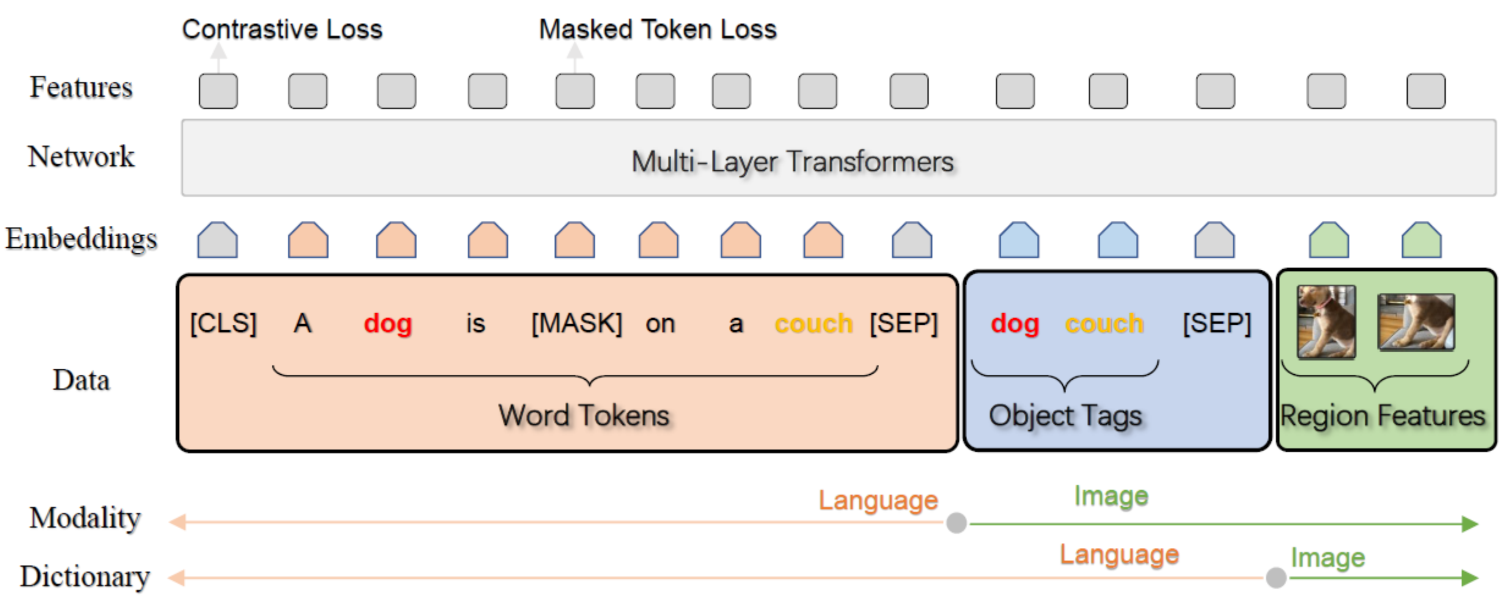
Semantics-Aligned Pre-training for V+L Tasks

- **Oscar**: **O**bject-**S**emanti**c**s **A**ligned P**r**e-training for Vision-Language Tasks (ECCV'20)
  - Training:
    - Inputs: triplets of caption-tag-region
    - Objectives: Masked token loss for words & tags + Contrastive loss tags and others
  - Fine-tuning:
    5 vision & language tasks (image captioning, NOC, VQA, image-text retrieval, etc.)



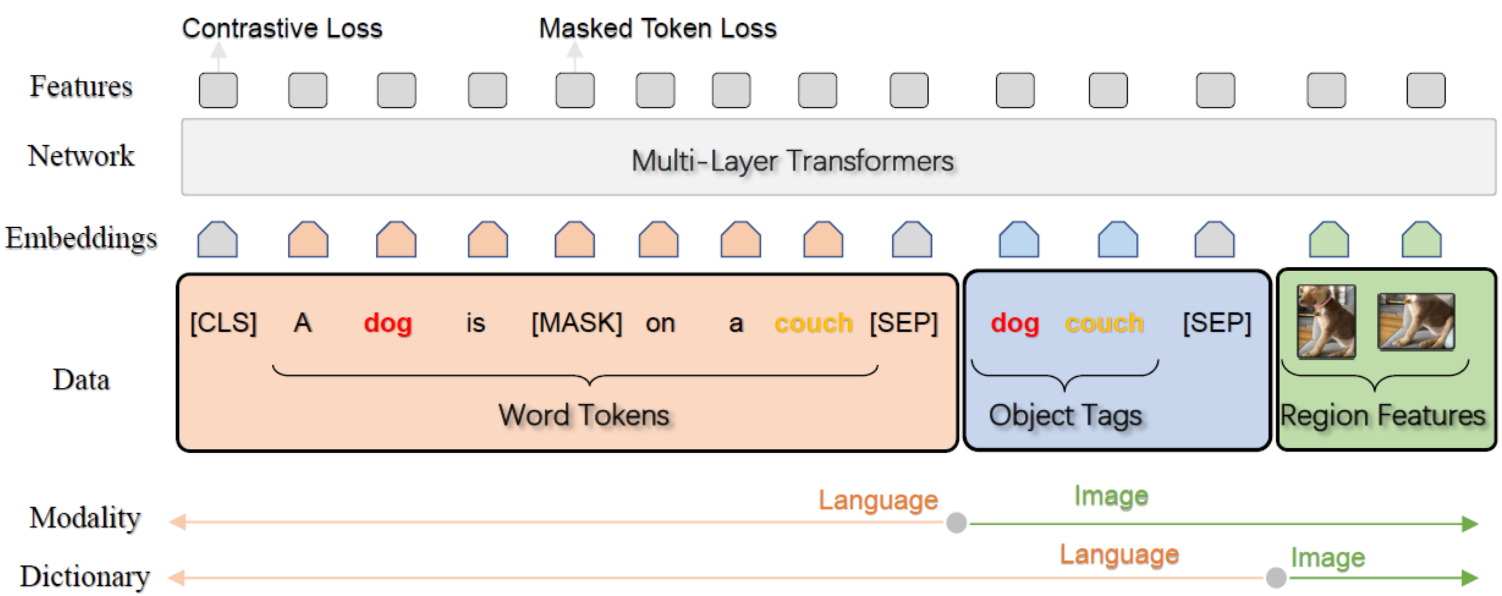
Semantics-Aligned Pre-training for V+L Tasks (cont'd)

- **Oscar**: **O**bject-**S**emanti**c**s **A**ligned P**r**e-training for Vision-Language Tasks (ECCV'20)
  - Training:
    - Inputs: triplets of word-tag-region
    - Objectives: Masked token loss for words & tags + Contrastive loss tags and others
  - Fine-tuning:
    - 5 vision & language tasks (image captioning, NOC, VQA, image-text retrieval, etc.)



Semantics-Aligned Pre-training for V+L Tasks (cont'd)

- **Oscar**: **O**bject-**S**emanti**c**s **A**ligned P**r**e-training for Vision-Language Tasks (ECCV'20)
  - Fine-tuning:
    5 vision & language tasks (image captioning, NOC, VQA, image-text retrieval, etc.)
  - Take **image captioning** as an example
    - Training: triplets of image regions features + object tags + captions as inputs;
      caption tokens with full attention on image regions/tags but not the other way around
    - Inference: image regions, tags and **[CLS]** as inputs,
      with **[MASK]** tokens sequentially added/predicted



Holding an apple ⟷  or 

- **Oscar** (cont'd)
  - Fine-tuning:
    5 vision & language tasks (image captioning, NOC, VQA, image-text retrieval, etc.)
  - Take **image-text retrieval** as an example
    - Training: aligned/mis-aligned image-text pairs as positive/negative input pairs, with **[CLS]** for binary classification (1/0)
    - Inference: for either image or text retrieval, calculate classification score of **[CLS]** for the input query

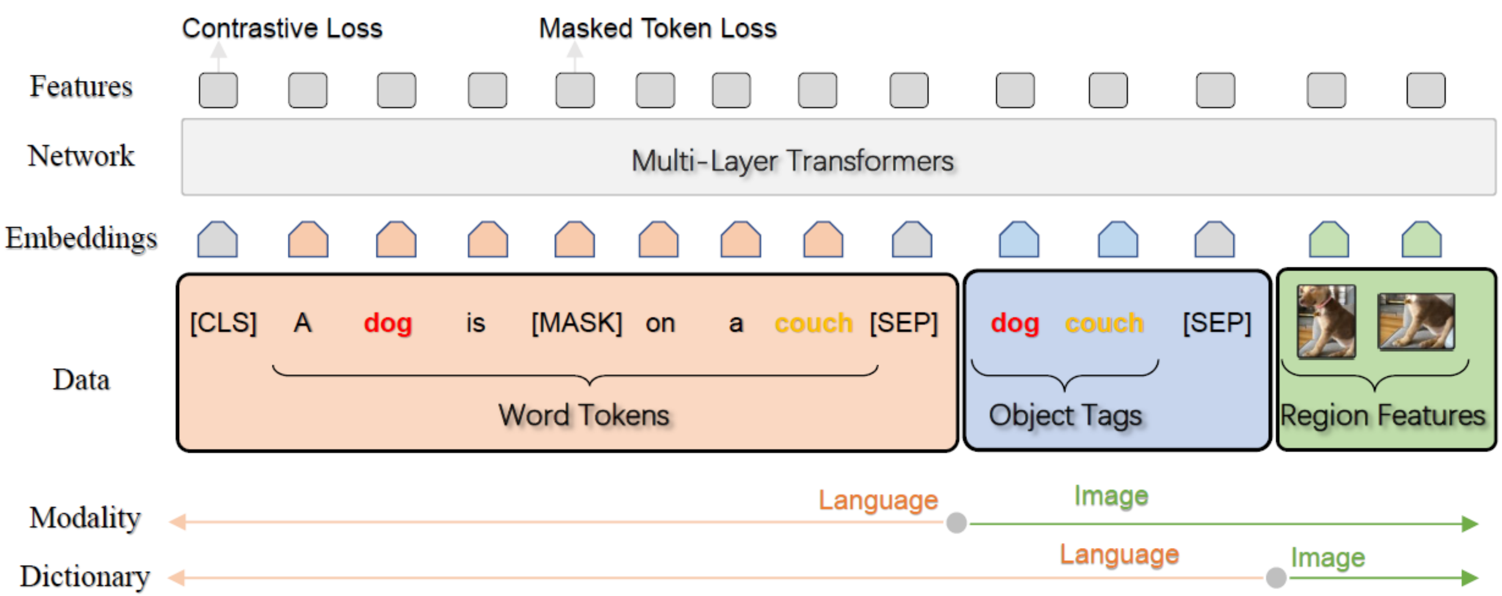
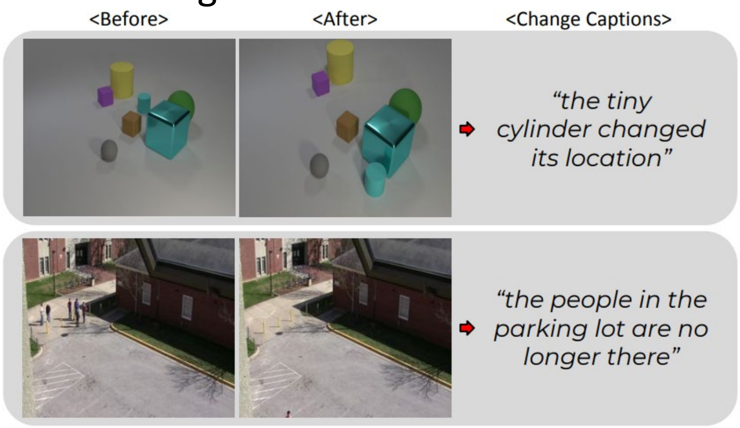


Image Change Captioning

- **Goal: Caption the difference(s) between input images**
  - Inputs: images with difference(s) + ground truth caption for the difference(s)
  - For image pair with one change



  - For image pair with multiple changes (Yue et al., ICCV'21)

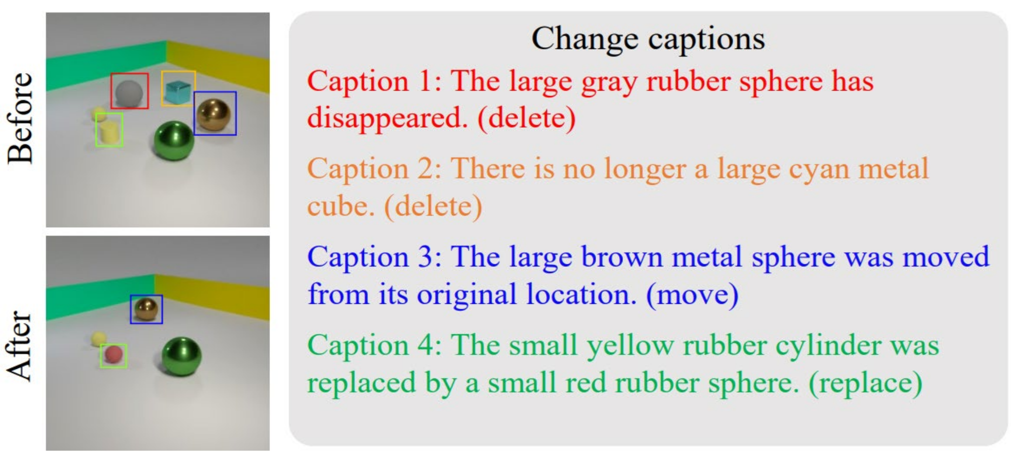
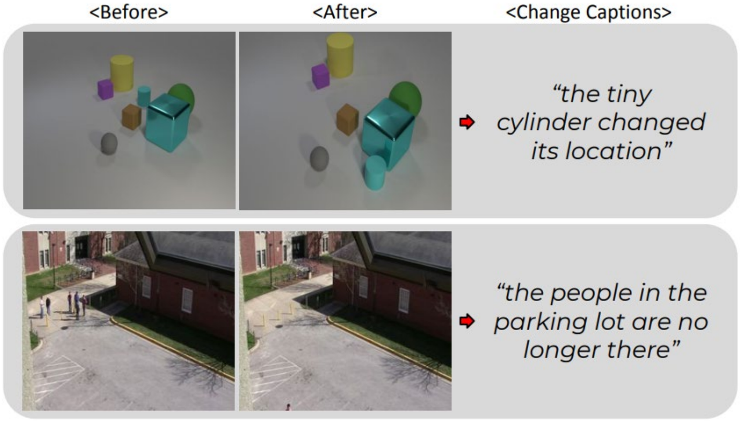
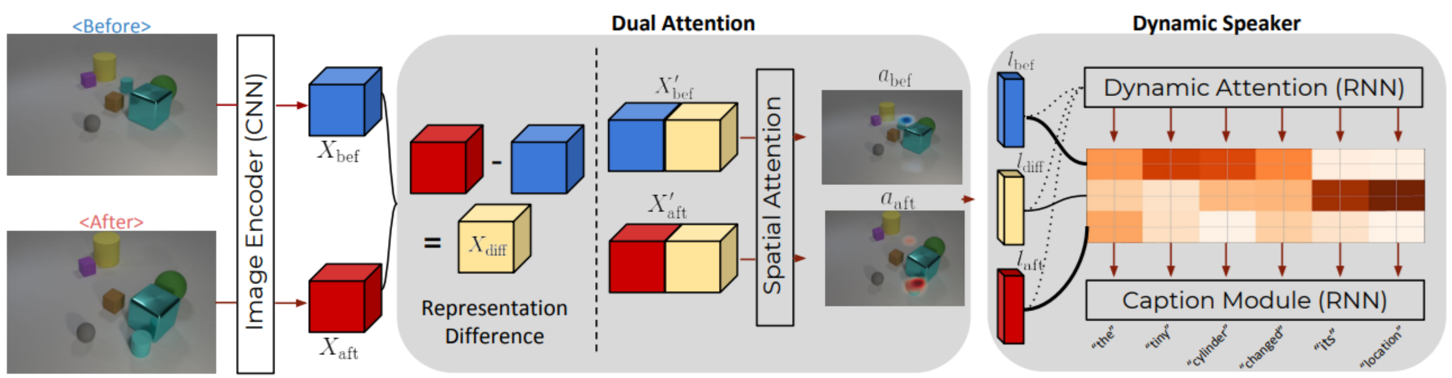


Image Change Captioning

- **Goal: Caption the difference(s) between input images**
  - Inputs: images with difference(s) + ground truth caption for the difference(s)
  - For image pair with one change



  - E.g., Robust Image Change Captioning, Dong et al., ICCV'19



What to Cover Today…

- **Recurrent Neural Network** & **Transformer**
  - Attention in RNN
  - *Attention is All You Need*: Transformer
  - Transformer for Visual Analysis
    - Visual Classification
    - Semantic Segmentation & More
- **Vision & Language**
  - Image Captioning
  - Text-to-Image Synthesis

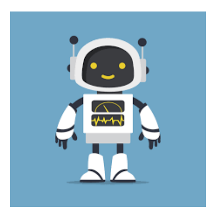
"a corgi wearing a bow tie and a birthday hat"

Teddy bears shopping for groceries in the style of ukiyo-e

Image Manipulation

- Text-to-Image Synthesis & Manipulation
  - Task #1: Text-to-image generation
    - Produce images based on their descriptions
    - Training: image-caption pairs
    - Recent works: Show & Tell (CVPR’15), StackGAN (ICCV’17), DALL-E (OpenAI)
    - Example:

Teddy bears shopping for groceries in the style of ukiyo-e



DALL-E

- Text-to-Image Synthesis & Manipulation (cont'd)
  - Text-to-image generation
  - Task #2: Image manipulation by text instruction
    - Allow users to edit an image with complex instructions (e.g., add, remove, etc.)
    - Training: reference image & instruction as inputs; target image as output
    - E.g., GeNeVa-GAN (ICCV'19), TIM-GAN (MM'21)
  - Task #3: Text/caption-guided image manipulation
    - Edit image regions to match image descriptions
    - Training: image-caption pairs
    - E.g., GLIDE (OpenAI'21), Tedi-GAN (CVPR'21), ManiTrans (CVPR'22)

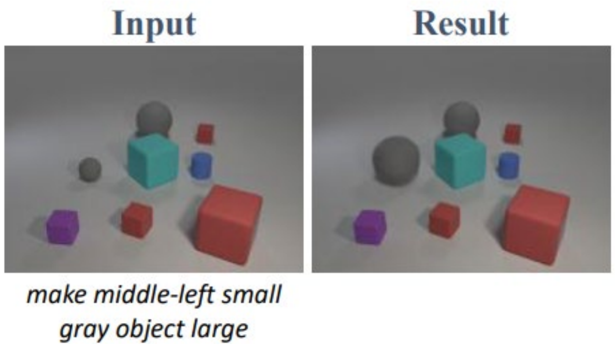


Fig. 1 Example of image manipulation by text instruction

Fig. 2 Example of text (caption)-guided image manipulation

Challenges in Text-Guided Image Manipulation

- Localization
  - Needs to identify objects in an image, locate the target location or objects of interest
  - Requires image understanding (with both semantics & spatial info)
- Manipulation
  - Needs to understand the input caption/instruction for manipulating images
  - Preserves object interaction and style to alleviate possible mismatch after manipulation

Input **Localization** **Manipulation**

a fire in the background

Text-Guided Image Manipulation (cont'd)

- Remarks & Opportunities
  - Not easy to collect training data with full supervision
  - Large-scale V&L pre-training models available (e.g., CLIP)
  - **Task #2** (manipulate by instruction) vs**. Task #3** (manipulate by text guidance)

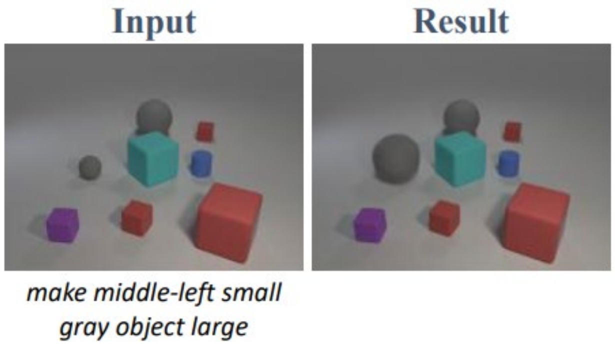


Fig. 1 Example of image manipulation by text instruction

Fig. 2 Example of text (caption)-guided image manipulation

  - Can scale up to industrial level with paired training data available

Selected Work on Text-Guided Image Manipulation

- GLIDE
  - Developed by OpenAI in 2021
  - Training:
    - Image-caption pairs and randomly generated masks
    - Learns to recover the missing part based on the caption

  - Testing: image, caption, and mask annotated by user
  - Later extended by a recent CVPR'22 work (DiffusionCLIP) for semantics improvements

"a corgi wearing a bow tie and a birthday hat"

"only one cloud in the sky today"

Composed Image Retrieval

- Goal
  - Given a reference image and its modification text (i.e., a cross-modal query), retrieve the target image from the database
  - Very different from image-text or text-image retrieval!

I want to change it to longer sleeves and yellow in color.

Reference Image

Modification Text

Target Image

Composed Image Retrieval with Pre-trained V&L Models

- Composed Image Retrieval using Pretrained LANguage Transformers (CIRPLANT)
  - Extract image features by a pre-trained ResNet
  - Aggregate information from modification text and reference image by a pre-trained **OSCAR**
  - Instead of use of output token [CLS], the derived output image feature ϕ is used for retrieval

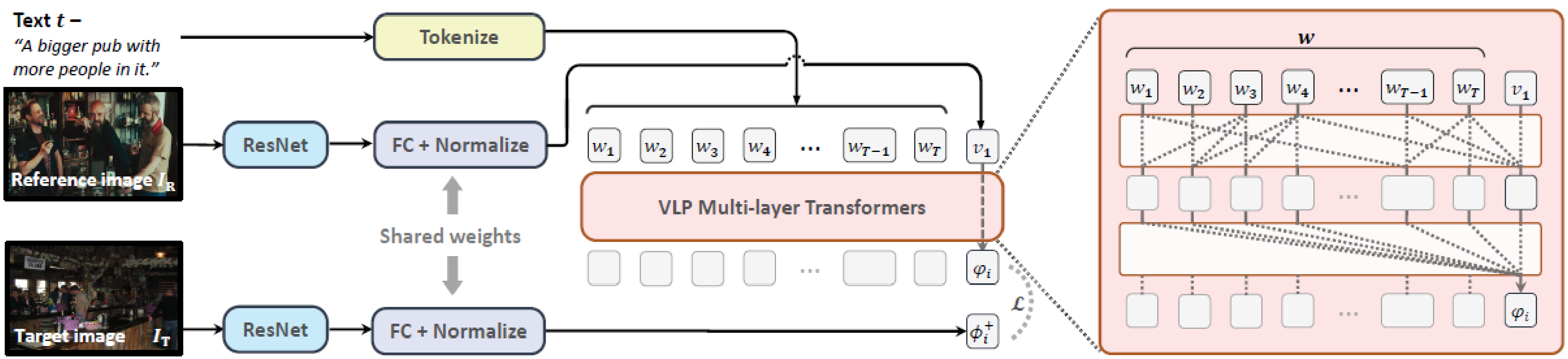
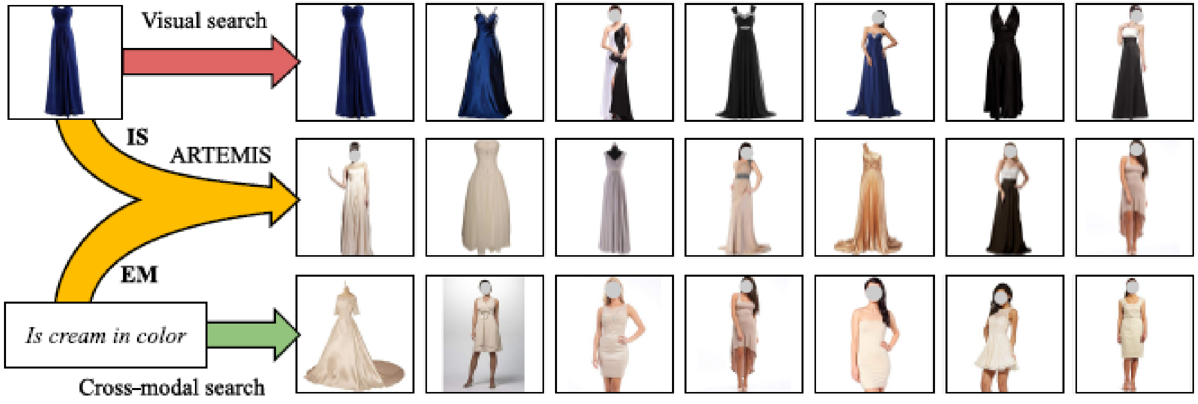


Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. *Zheyuan Liu et al*. ICCV 2021

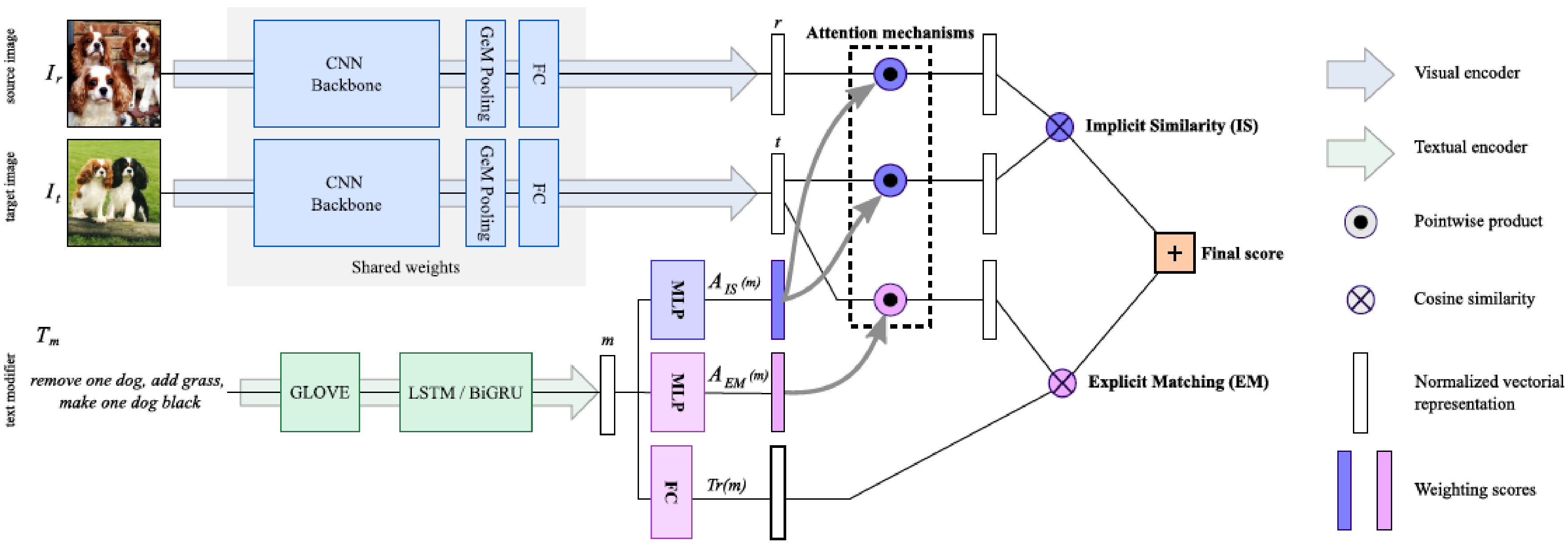
Retrieval with Text-Explicit Matching & Implicit Similarity

- **A**ttention-based **R**etrieval with **T**ext-**E**xplicit **M**atching and **I**mplicit **S**imilarity (**ARTEMIS**)
  - Image search with free-form text modifier
  - Cross-modal learning and visual retrieval
    - Text-guided attention is introduced ARTEMIS

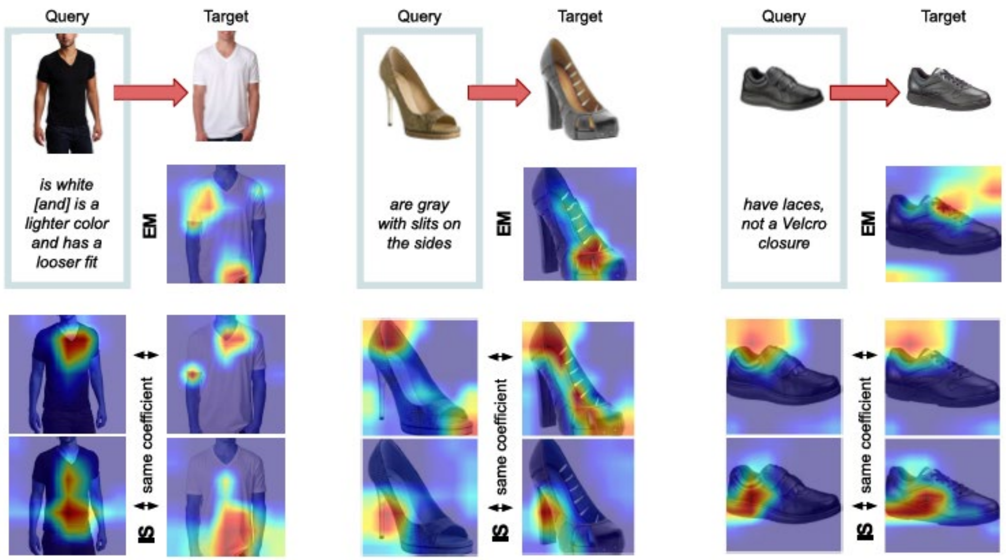


ARTEMIS: Attention-based Retrieval with Text-Explicit Matching & Implicit Similarity. *Ginger Delmas et al*. ICLR 2022

- **A**ttention-based **R**etrieval with **T**ext-**E**xplicit **M**atching and **I**mplicit **S**imilarity (**ARTEMIS**) (cont'd)
  - Implicit Similarity (IS):
    attention mechanism focusing on what's not mentioned by text and should be preserved
  - Explicit Matching (EM):
    attention mechanism focusing on what's mentioned by text and should be changed.



- **A**ttention-based **R**etrieval with **T**ext-**E**xplicit **M**atching and **I**mplicit **S**imilarity (**ARTEMIS**) (cont'd)
  - Example Results & Extension



What to Cover Today…

- **Recurrent Neural Network** & **Transformer**
  - Attention in RNN
  - *Attention is All You Need*: Transformer
  - Transformer for Visual Analysis
    - Visual Classification
    - Semantic Segmentation & More
- **Vision & Language**
  - Image Captioning
  - Text-to-Image Synthesis
- **HW #3 is out!**

"a corgi wearing a bow tie and a birthday hat"

Teddy bears shopping for groceries in the style of ukiyo-e