CSCI 566: Deep Learning and Its Applications

Jesse Thomason

Lecture 8: Multimodal Deep Learning

CSCI 566 Roadmap

Module 1:
Neural Network Basics

Module 2:
Deep Learning Applications

Module 3:
Advanced Topics in Deep Learning

January 2023

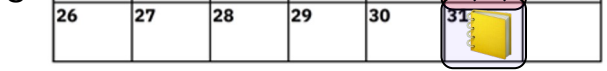
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

February 2023

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28				

March 2023

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

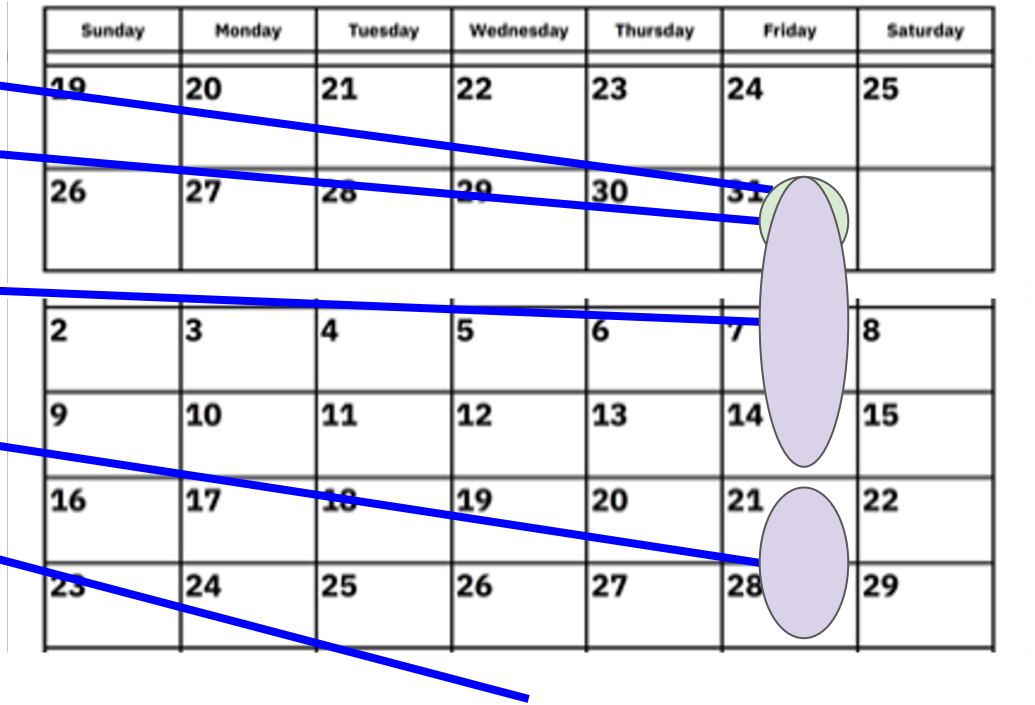


April 2023

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

[Mar 24] Endgame Deliverables

Assignment 2 Due	Mar 31
Project Midterm Report	Mar 31
Paper Roleplaying Breakout Sessions	Mar 31, April 7, April 14
Final Project Presentations	April 21, April 28
Project Final Report	May 3



Project Midterm Reports Due Mar 31st, 11:59pm

- Details about the set of initial experiments you have done for the project
  - Can involve a short abstract/introduction
- Should involve reproducing the results of a state-of-the-art baseline model for the task of interest with code that you have implemented, or a pilot version of your proposed approach
- Talk about the specific dataset(s) you choose to use, the evaluation protocol and metric you decide to have, and the experiment settings.
- Perform an analysis of what remaining errors your mid-report model makes and describe how you will approach these by the final report
- Format: a 3-page report + contributions page
- Syllabus provides suggested format outline and grading breakdown

Paper Roleplaying Breakouts: March 31, April 7, April 14

- On these days, we will have ~90 minutes of lecture followed by ~90 minutes for paper roleplaying breakout sessions
- Upshot: You will take on a “role” that defines your relationship with a paper to be discussed
  - You will deep dive into the paper with respect to your role
  - You will produce a 2 page document + participate in the detailed, 90-minute discussion with your breakout group
- You are expected to play *one role* for *one session* across these three days; you’re free to participate as an observer on days for which you are not scheduled to be in a session

Paper Roleplaying Breakouts: March 31, April 7, April 14

ROLE	Description
PRESENTER	🗣️ Presenter. Give a 10 minute or less oral description of the highlights of the paper. Some papers may require more technical background/discussion than others. You don't have to cover every last detail. You can't use slides, but you may bring (or digitally distribute) copies of a 2 page or less handout with bullet points, key figures/tables, etc.
REVIEWER	🕵️ Scientific Peer Reviewer. The paper has not been published yet and is currently submitted to a top conference where you've been assigned as a peer reviewer. Complete a full review of the paper answering all prompts of the official review form of ACL Rolling Review (https://aclrollingreview.org/reviewform). Your review should include recommending whether to accept or reject the paper.
TEACHER	🧑‍🏫 Teacher. Creating an exercise to engage with some aspect of the paper. For example, you might run a hypothetical scenario of involving applications of the paper to real data. You could also create an activity like a quiz or discussion question centered around understanding a particular result/figure/equation.
ARCHAEOLOGIST	🏺 Archaeologist. This paper was found buried underground in the desert. You're an archeologist who must determine where this paper sits in the context of previous and subsequent work. Find and report on one older paper cited within the current paper that substantially influenced the current paper and one newer paper that cites this current paper.
RESEARCHER	🧫 Academic Researcher. You're a researcher who is working on a new project in this area. Propose an imaginary follow-up project not just based on the current but only possible due to the existence and success of the current paper.
INDUSTRY	💰 Industry Practitioner. You work at a company or organization developing an application or product of your choice (that has not already been suggested in a prior session). Bring a convincing pitch for why you should be paid to implement the method in the paper, and discuss at least one positive and negative impact of this application.
HACKER	👾 Hacker. You're a hacker who needs to implement the approach described in the paper as faithfully as possible. Are the methods and experimental conditions described in sufficient detail to be replicated? Were usable resources (code, data, etc.) released? If so, spend 15 min. looking through them to see how well documented they are and how easy they will be to use.
PRIVATE INVESTIGATOR	🕵️ Private Investigator. You are a detective who needs to run a background check on one of the paper's authors. Where have they worked? What did they study? What previous projects might have led to working on this one? What motivated them to work on this project? Feel free to contact the authors, but remember to be courteous, polite, and on-topic.
SOCIAL	🌎 Social Impact Assessor. Identify how this paper self-assesses its (likely positive) impact on the world. Have any additional positive social impacts left out? What are possible negative social impacts that were overlooked or omitted?

Paper Roleplaying Breakouts: March 31, April 7, April 14

Date	March 31					April 7					April 14				
Paper	On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜	Neural Word Embedding as Implicit Matrix Factorization	CLIPort: What and Where Pathways for Robotic Manipulation	Learning Transferable Visual Models From Natural Language Supervision	Auto-Encoding Variational Bayes	High-Resolution Image Synthesis with Latent Diffusion Models	Deep Residual Learning for Image Recognition	U-Net: Convolutional Networks for Biomedical Image Segmentation	Language Model Cascades	The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks	Emergence of Maps in the Memories of Blind Navigation Agents	Learning to Explore using Active Neural SLAM	Attention Is All You Need	Playing Atari with Deep Reinforcement Learning	The Hardware Lottery
PRESENTER															
REVIEWER 1															
REVIEWER 2															
REVIEWER 3															
TEACHER 1															
TEACHER 2															
ARCHAEOLOGIST															
RESEARCHER															
INDUSTRY															
HACKER 1															
HACKER 2															
PI 1															
PI 2															
SOCIAL 1															
SOCIAL 2															

- Self sign-up; don't overwrite or move other people!
- Self sign-up sheet

Project Final Presentations: April 21, April 28

- Note that project final presentations will take place over two different classes [team sign up sheet]
- You should sign up your team for a slot on one of these days
- We will discuss further details about what should be contained in the final talk closer to the due dates; you will be expected to attend these final presentations
    - Your group will submit to us a multiple choice question about your final presentation, which we'll use to build quizzes that summarize each day's presentations.

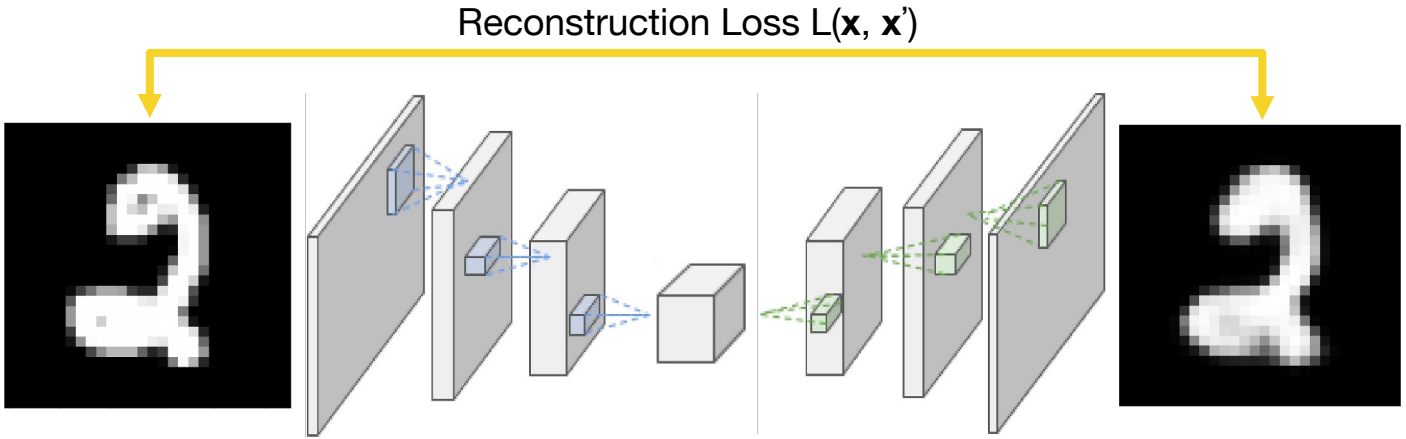
Overview of Today's Plan

- ~~Course organization and deliverables~~
  - Any questions before we move on?
- Lecture 7 Recap
- Multimodal Deep Learning

Autoencoder

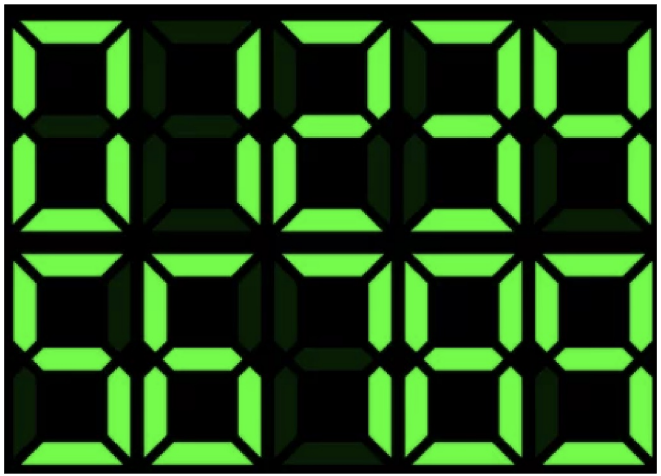
- That word embeddings are a learned compression of co-occurrence matrices for tokens is part of a larger trend
- One can formulate deep learning models to be *learned compression algorithms* for a given input space
- Autoencoding is *unsupervised*:
  - Given input of size X, shrink to size x<X with learned function (encoder) and unshrink it back to X with a another learned function (decoder)
- Autoencoders learn a *lower dimensional manifold* for input

CNN Autoencoder

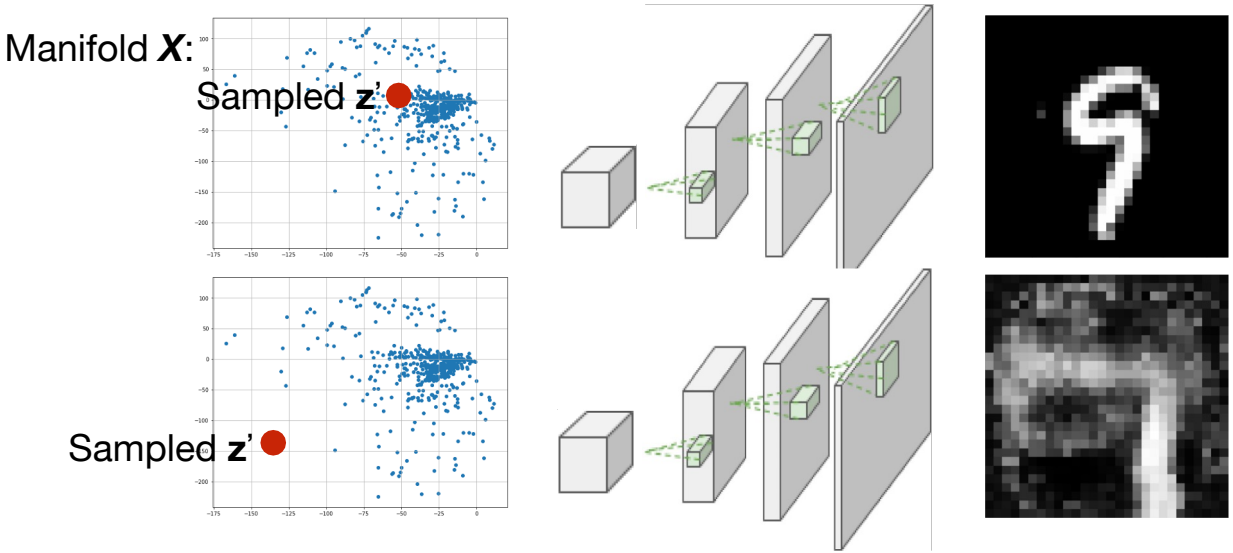


What Might The Latent Space “Look Like”?

- Let’s imagine for ourselves: you are compressing digits and using a cross-entropy loss (rather than MSE) to judge whether the reconstructions are accurately classified as class 0…9
- Let’s say the latent space is a single, binary vector of dimension seven
- What representation could emerge?

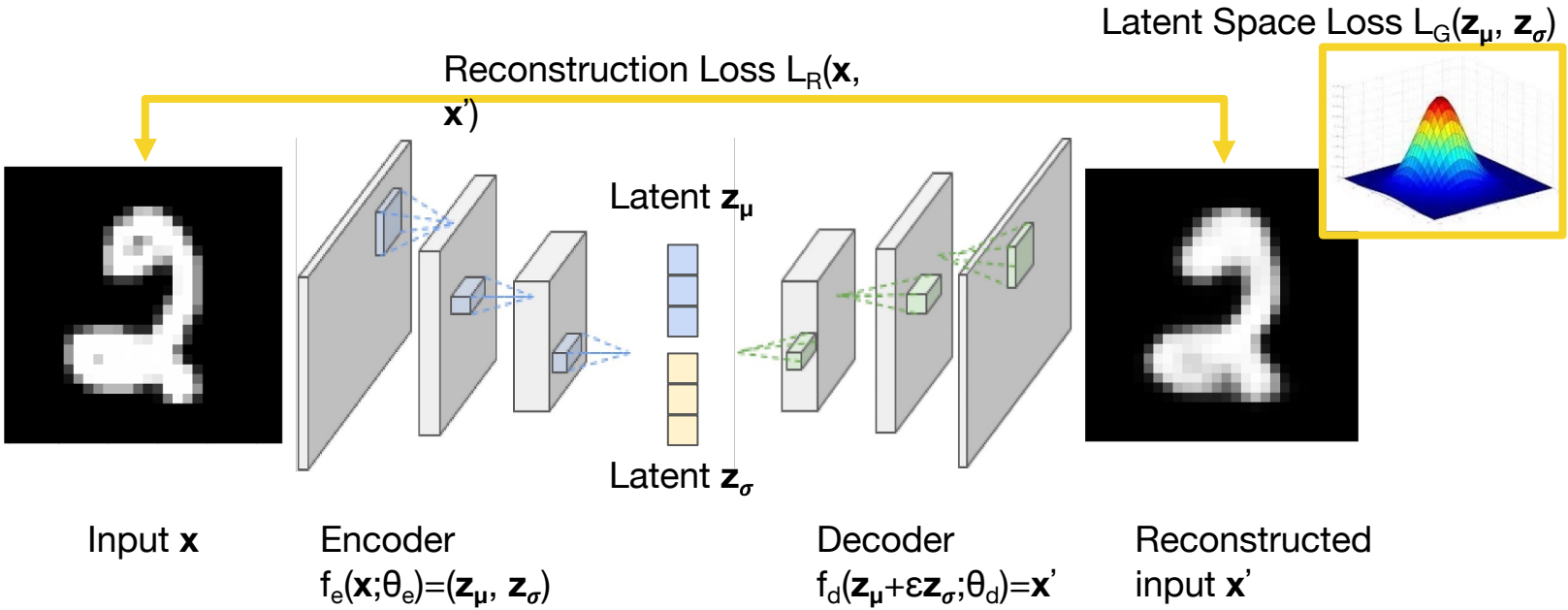


Autoencoder Decoder As Data Generator



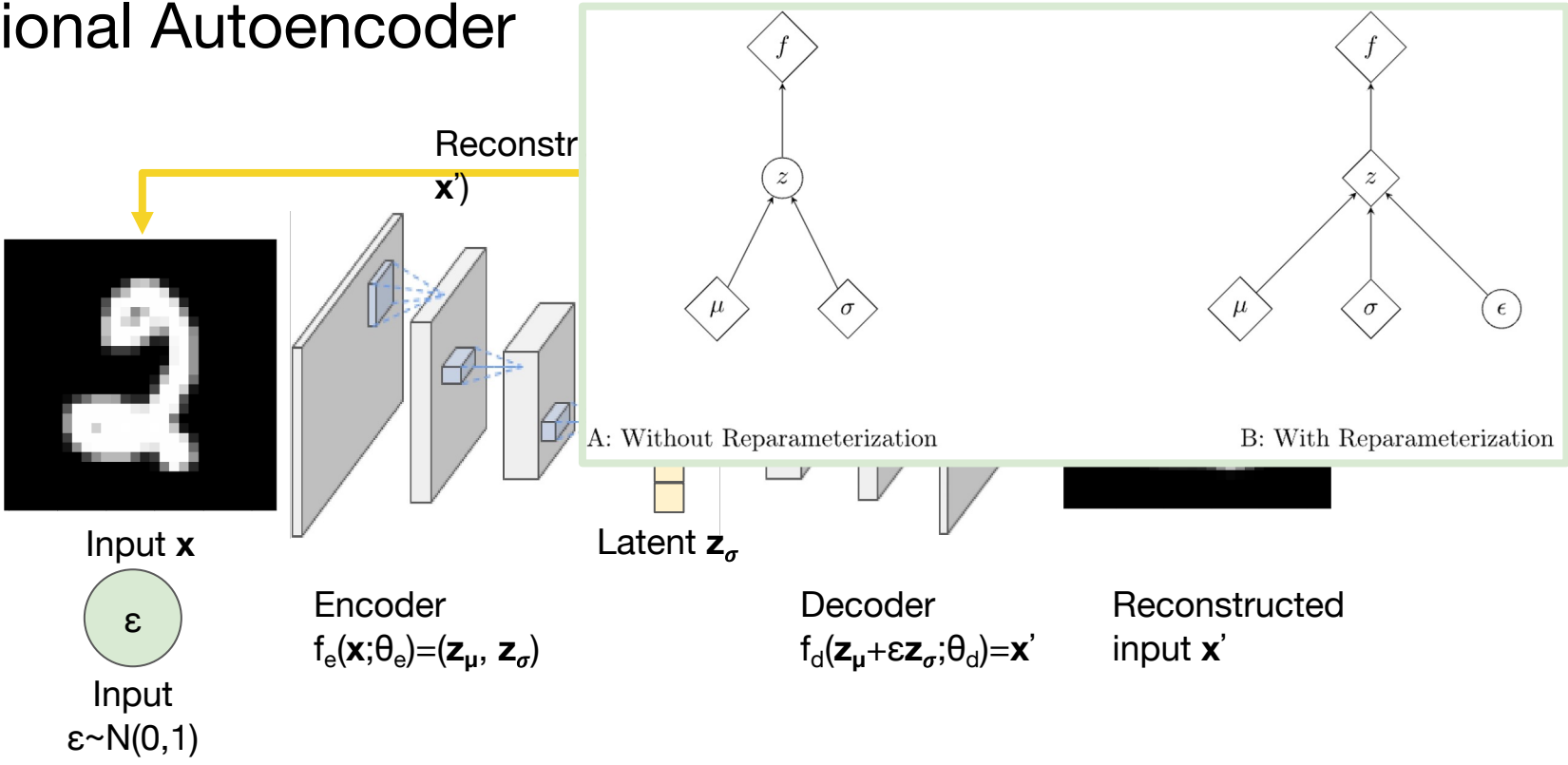
- What if we sample vector ***z'*** as draws from a Gaussian?
- Will we achieve reconstructions that look like our input data *X*?
  - No! Our data *X* encodes to a particular manifold in $R^{|z|}$

Variational Autoencoder



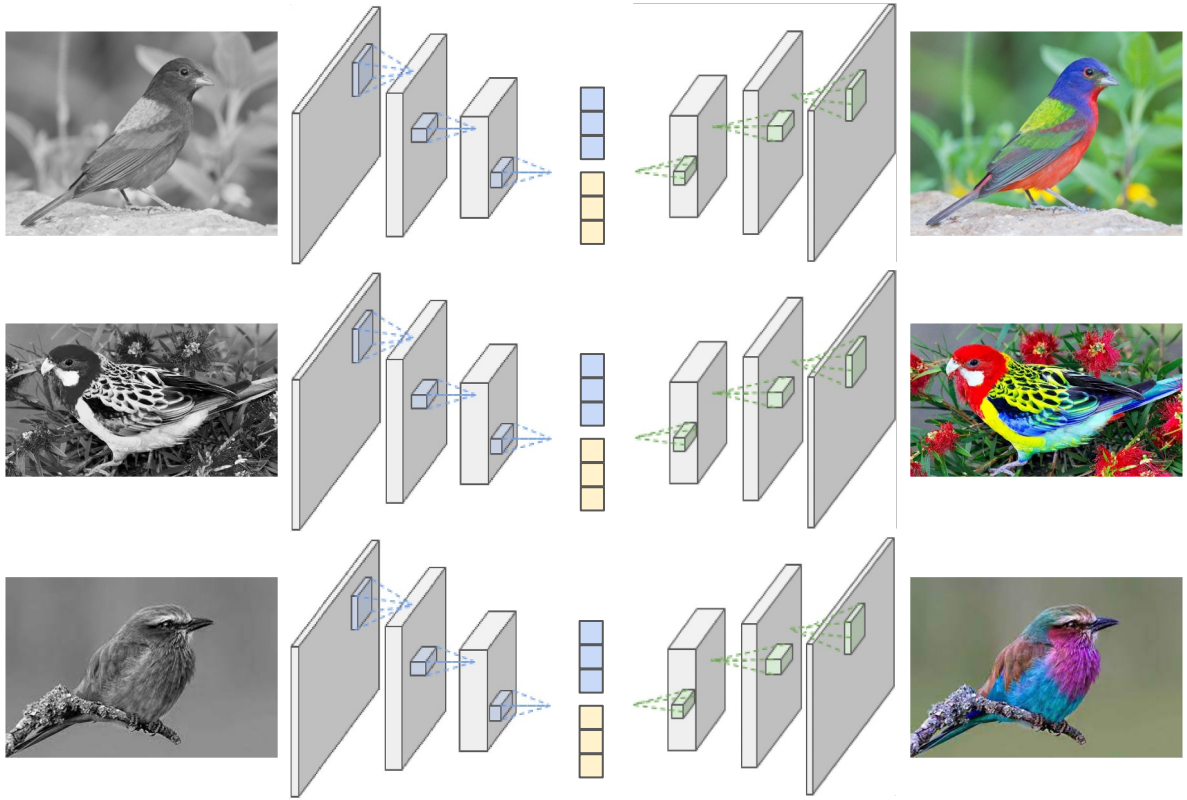
- We learn a mean and variance for the latent manifold for every input, then constrain those to be Gaussian

Variational Autoencoder

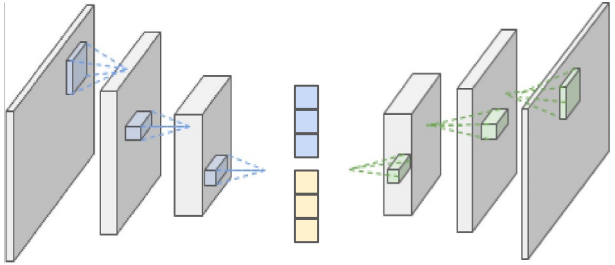


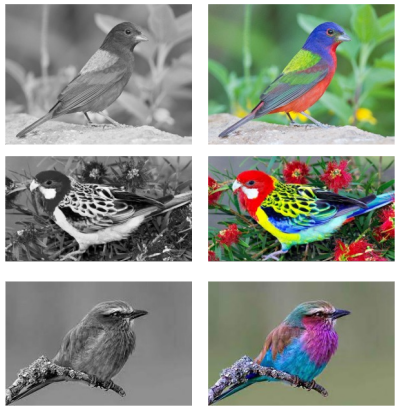
[The Reparameterization Trick in Variational Autoencoders]

Autoencoder Decoder For Downstream Tasks



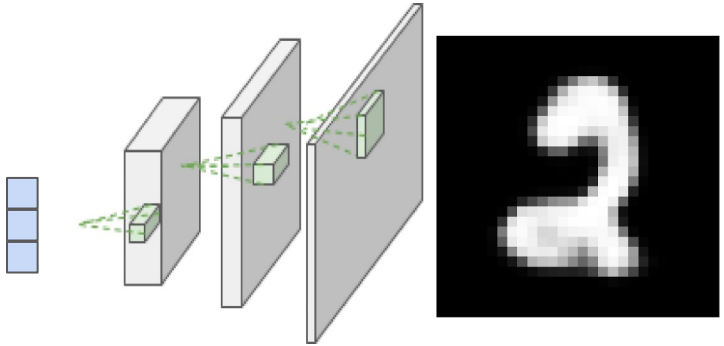
Variational Autoencoder Mode Collapse





- Minimizing reconstruction loss means defaulting to *averages* for high variance tasks like “coloring a bird”
- Average minimizes the loss but never gets the right color!
- How can we penalize taking these “safe bet” local minima?

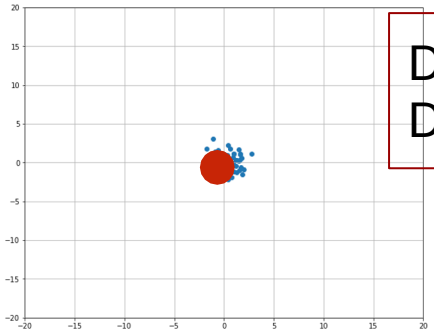
Generative Adversarial Networks



Rand **z**

Decoder
$f_d(\mathbf{z};\theta_d)=\mathbf{x}'$

Upsampled **x**'



Discriminator D(**x'**) -> 0
Discriminator D(**x**) -> 1

Rand *N(0, 1)*

Generator f(**z**) -> **x'**

Generator Network

Gen Images

Train Images

Discriminator Network D

Discriminator Network D

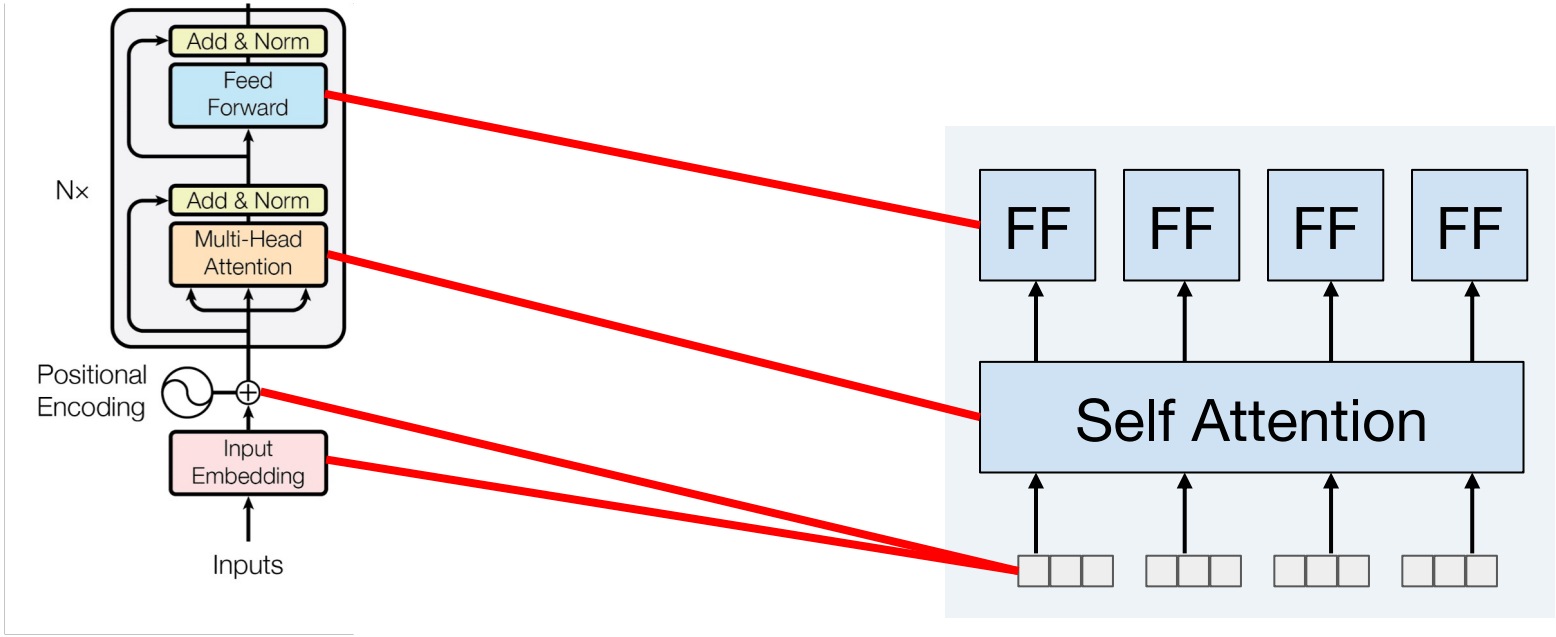
Real/Fake Decision

Real/Fake Decision

Transformers

- Transformer $T : S_a \longrightarrow S_b$ is a function from…
  - A set a in S_a to a set b in S_b
- Simple transformer consists of a single layer encoder E and a single layer decoder D with single-headed attention

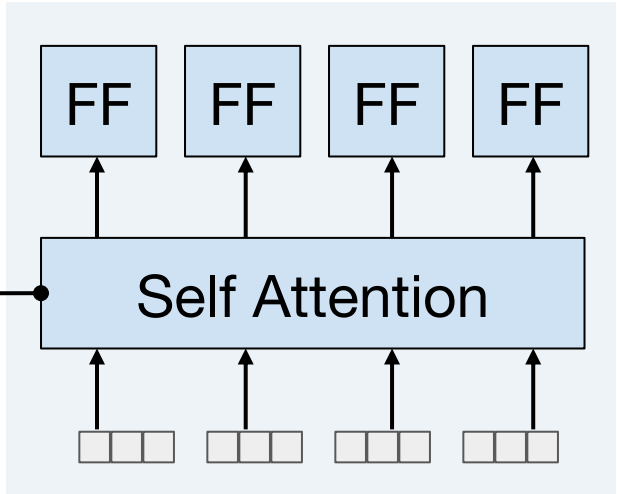
Transformers: Self-attention



Attention Is All You Need

Scaled dot-product attention

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

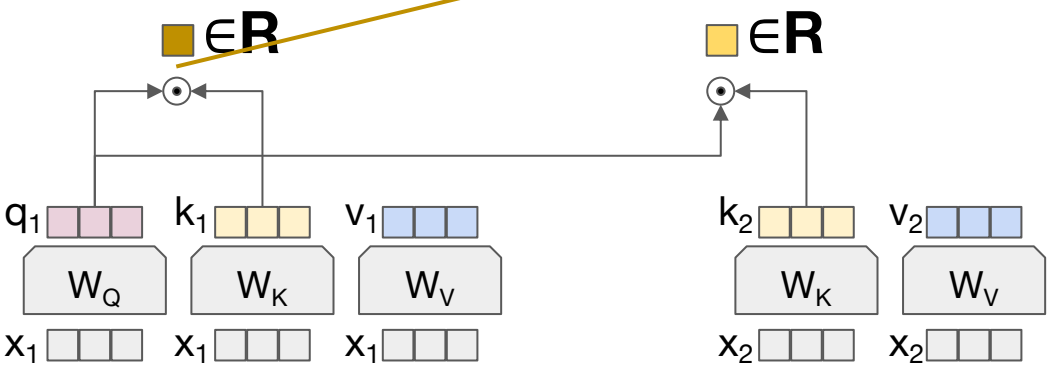


Our input for *self-attention* is the matrix **X** which is the embeddings of the input sequence

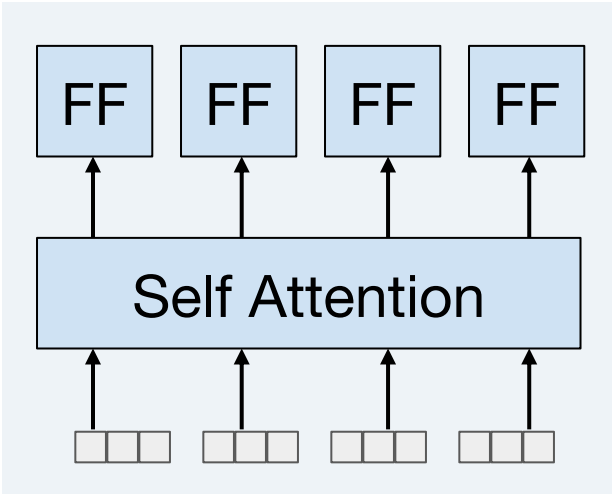
So we call Attention(**X**, **X**, **X**)

Self-Attention

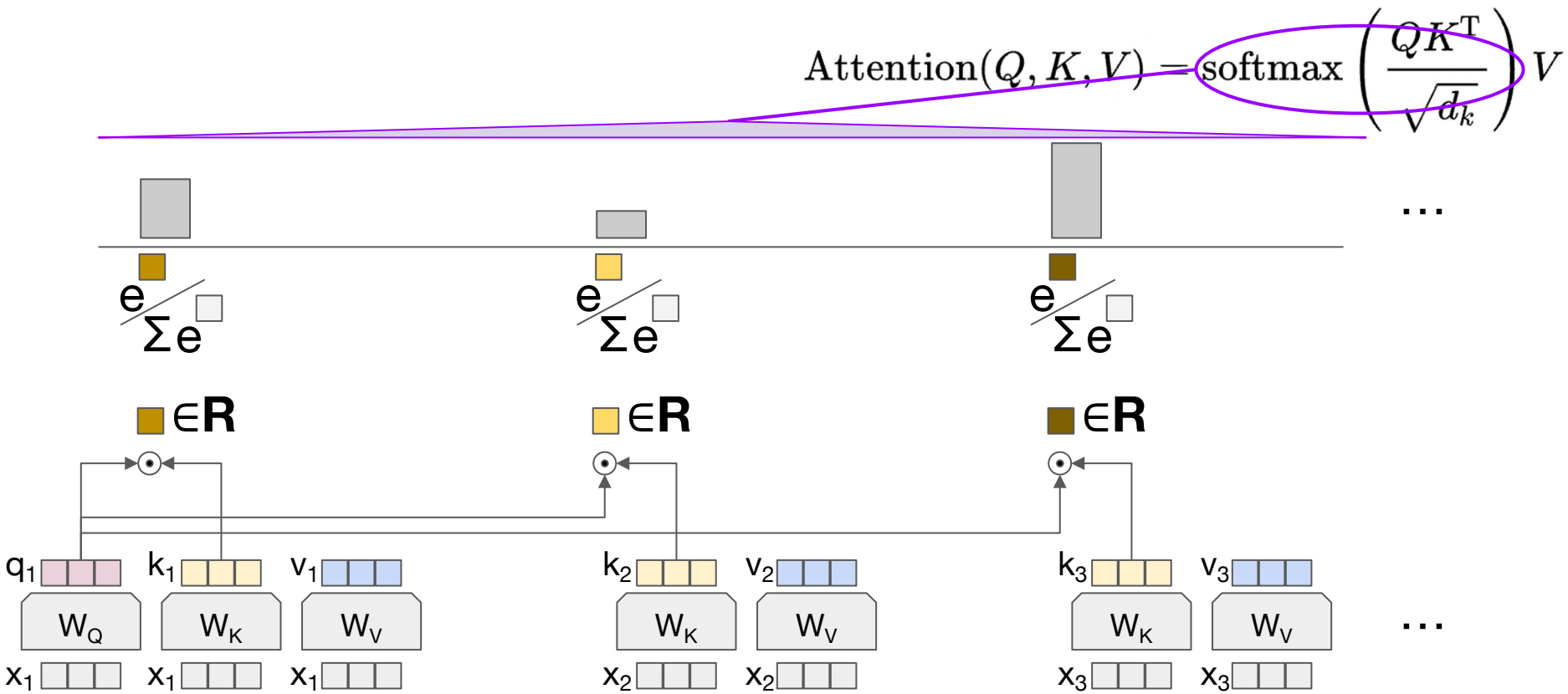
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



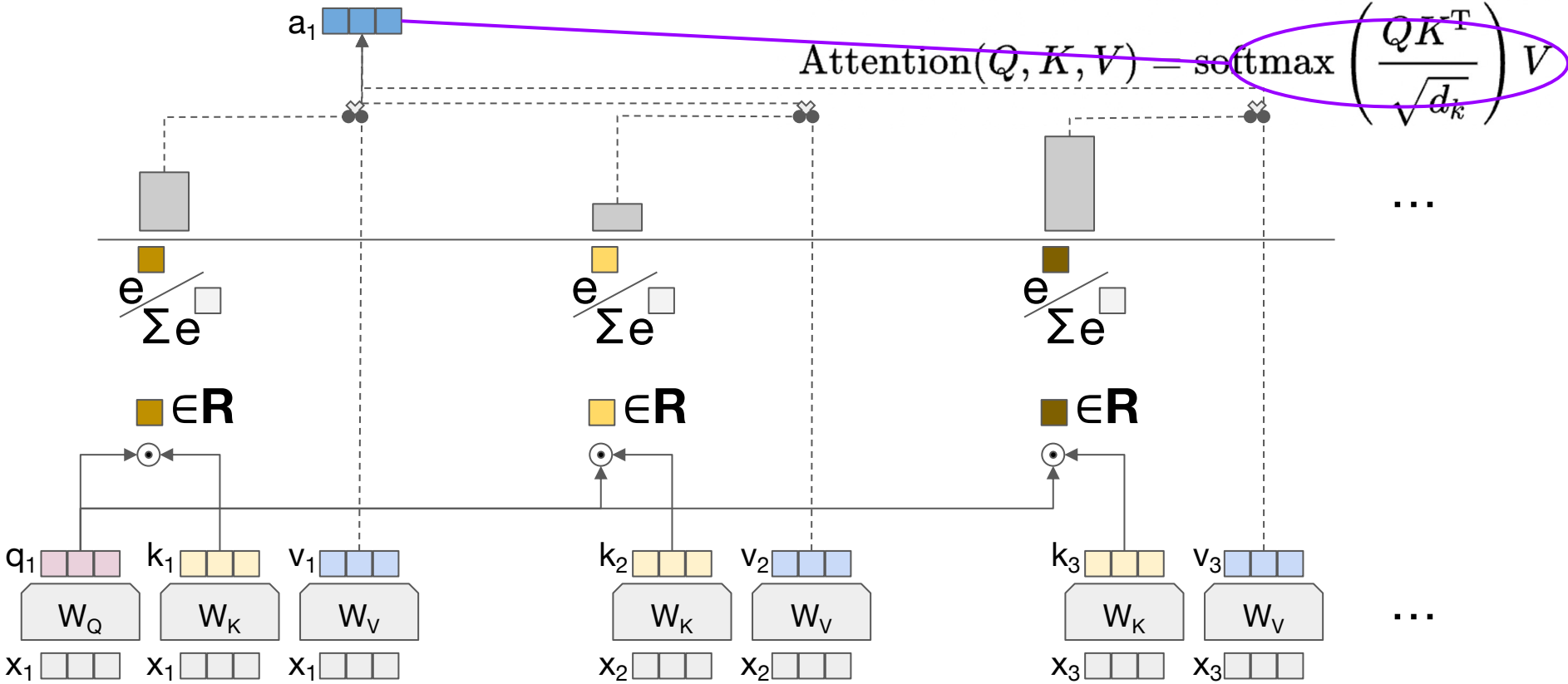
x



Self-Attention

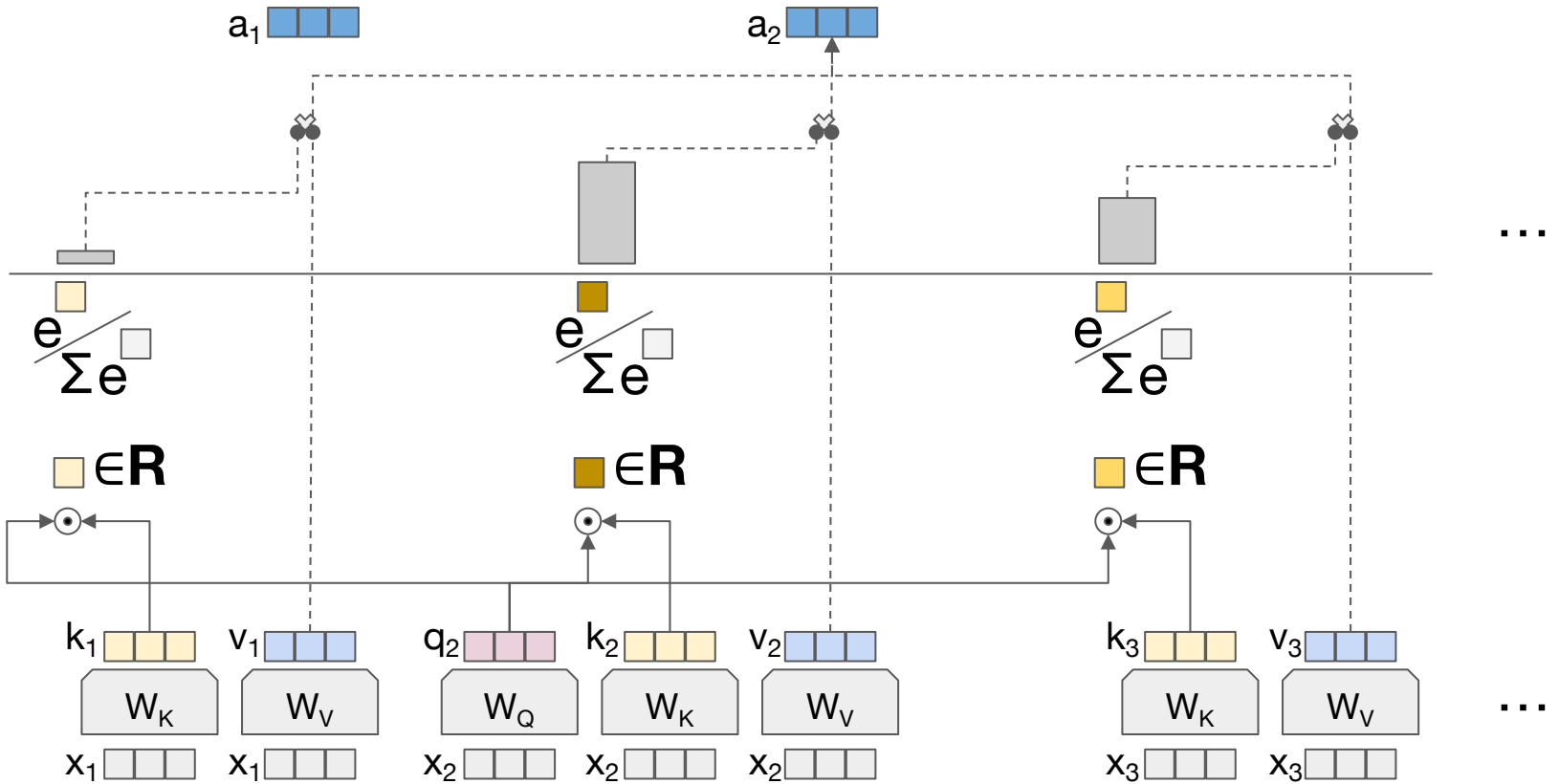


Self-Attention



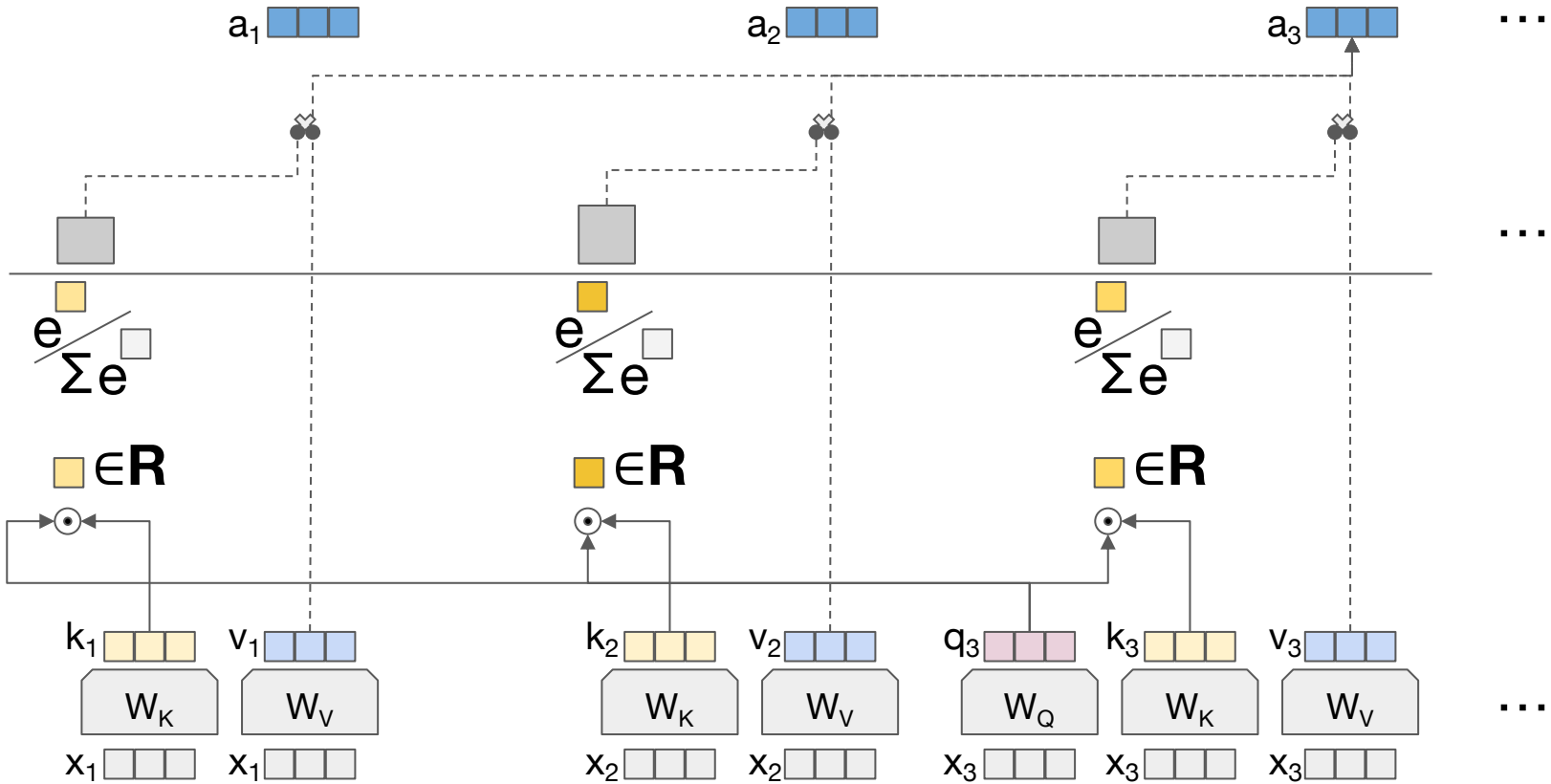
Self-Attention

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

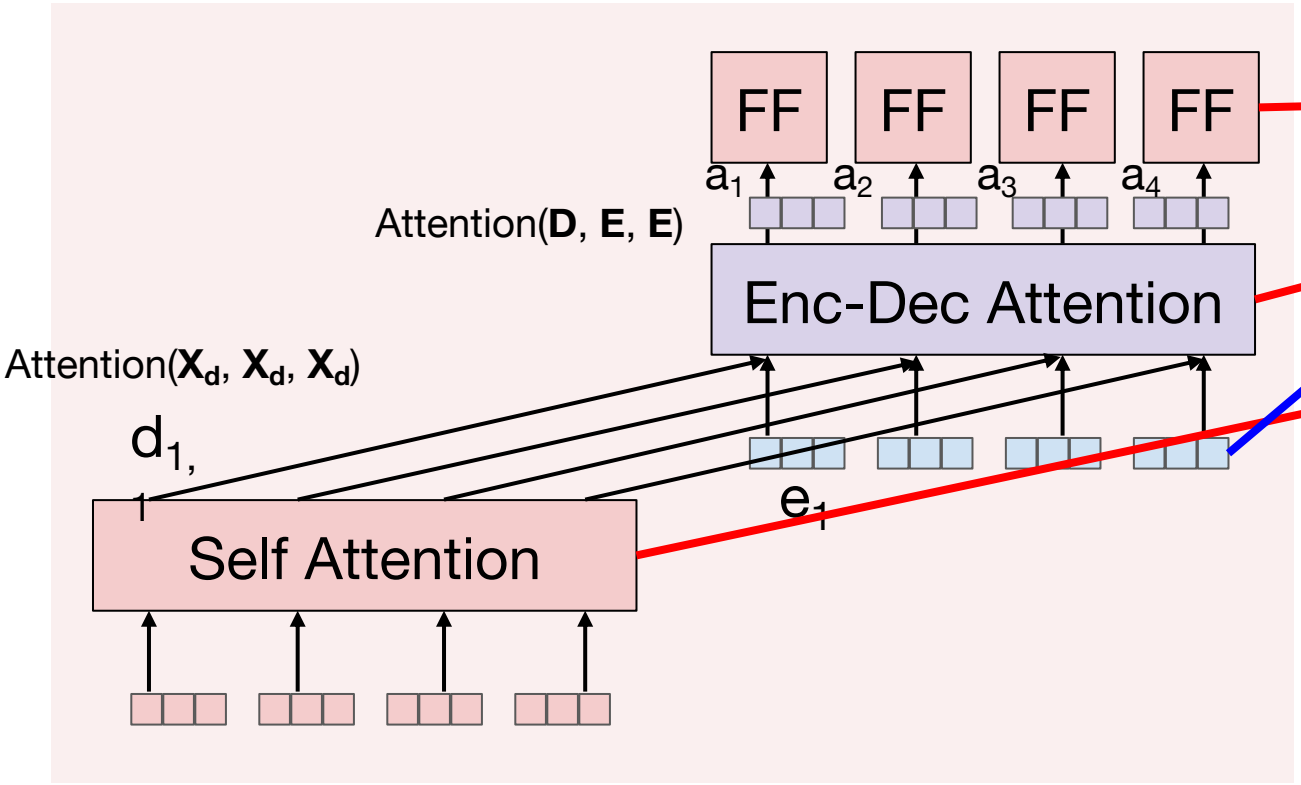


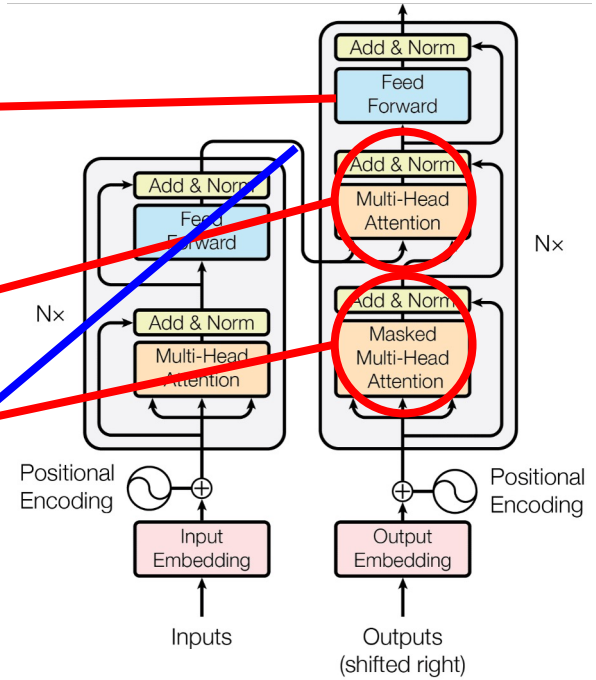
Self-Attention

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



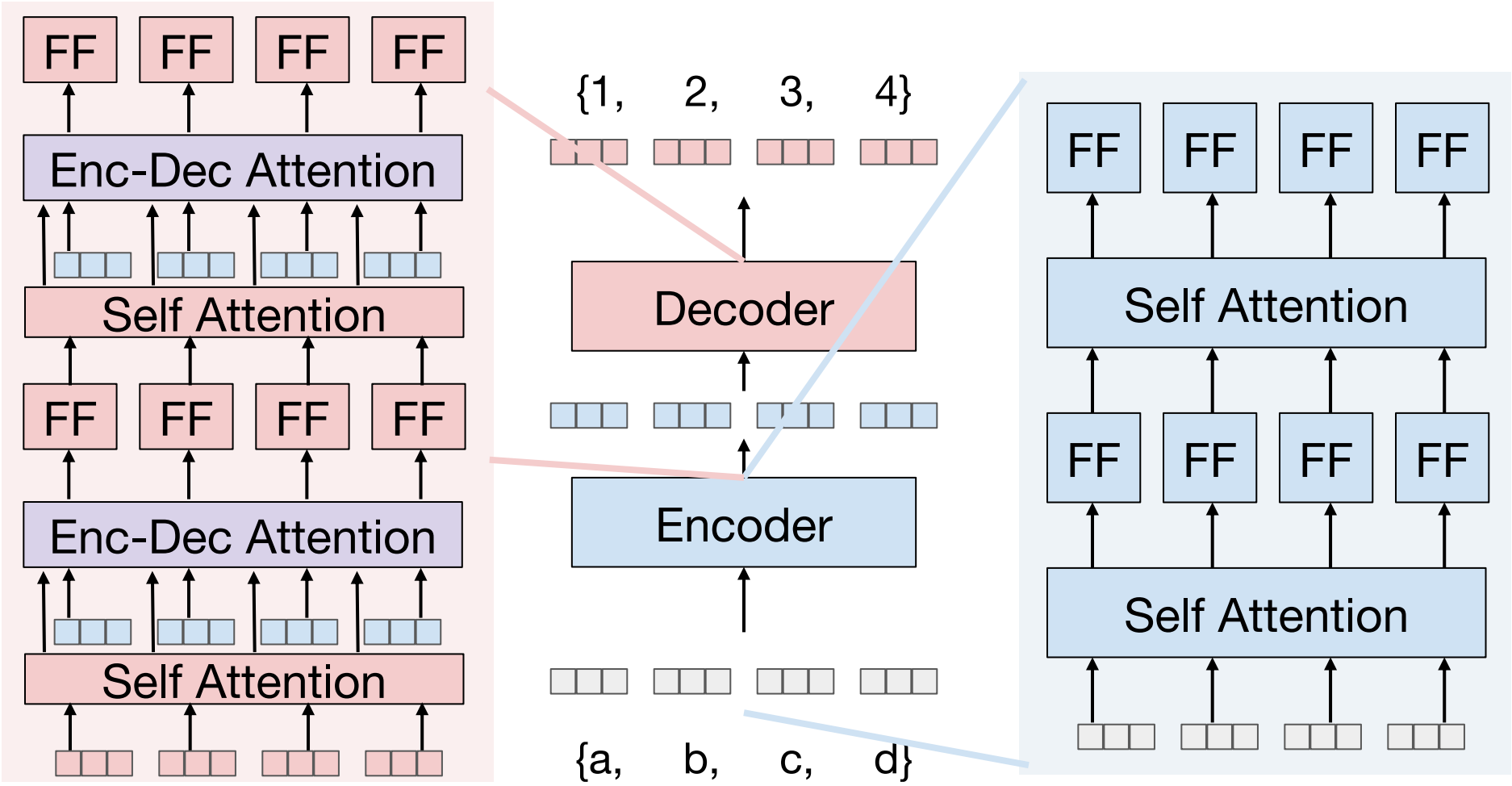
Encoder-Decoder Attention



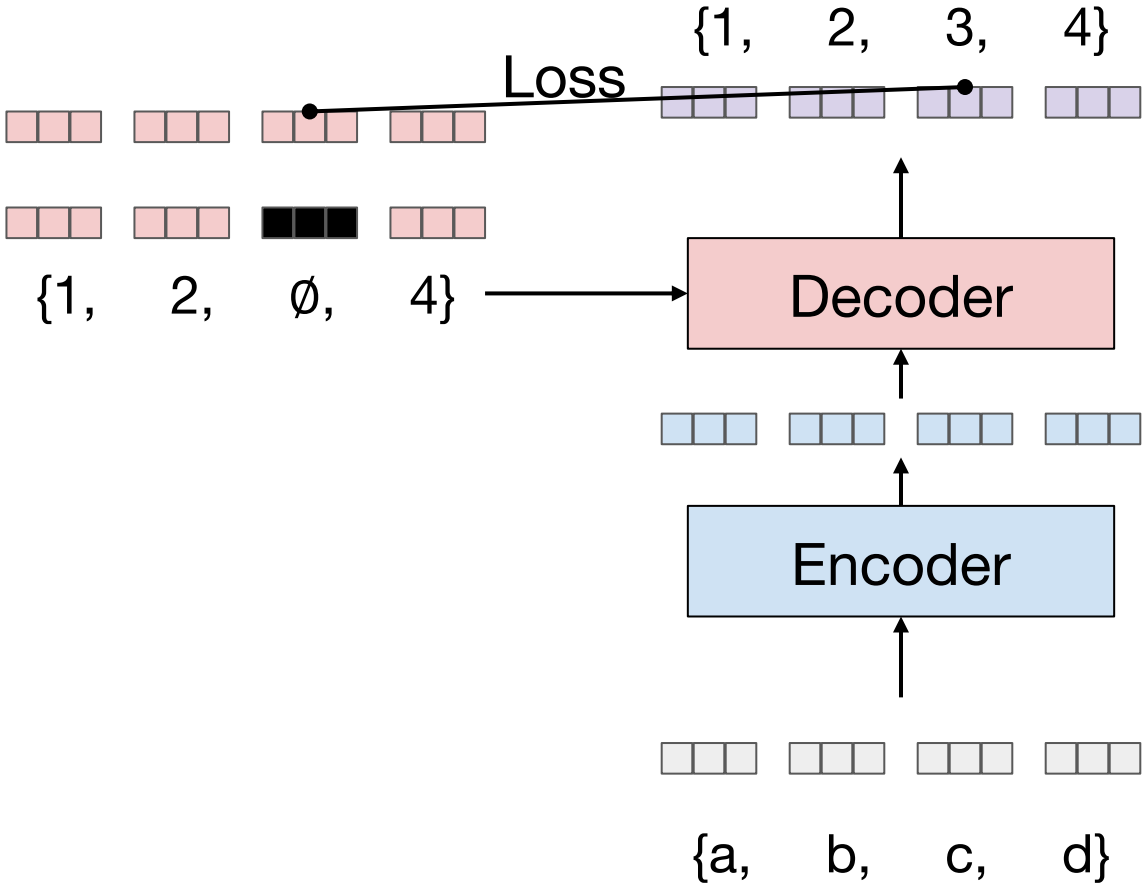


[Attention Is All You Need]

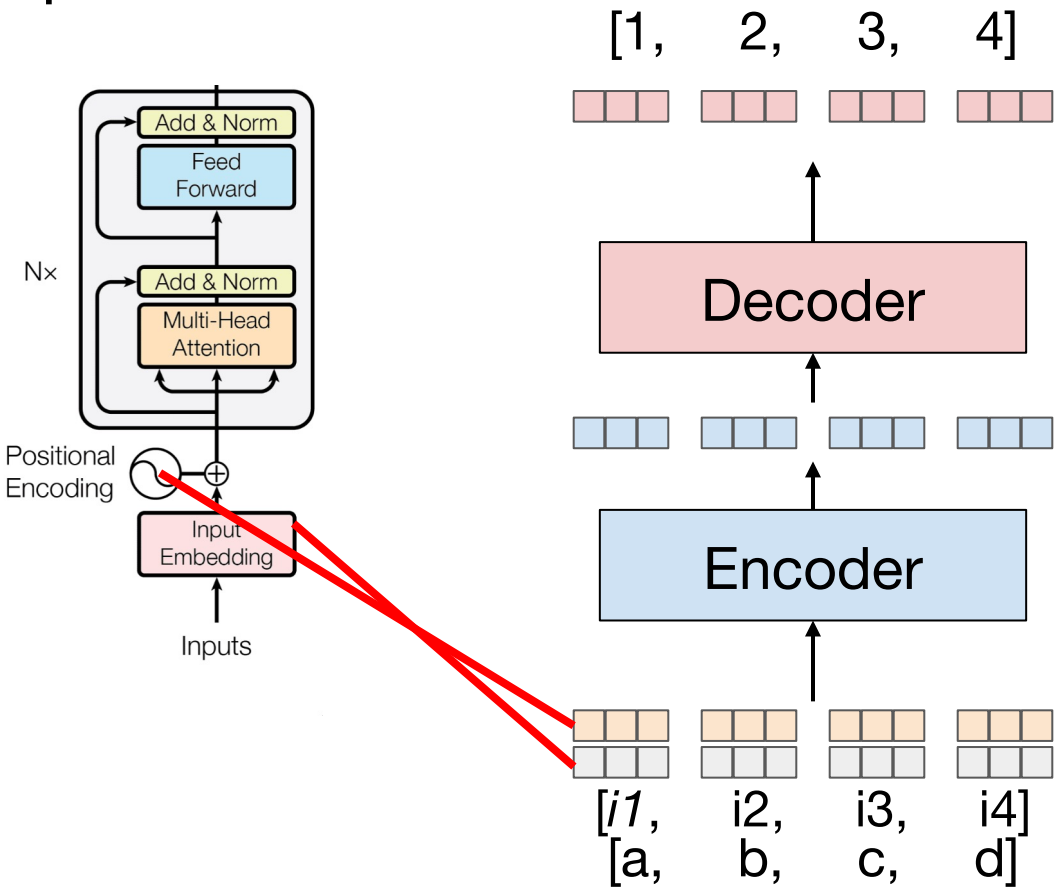
Deeper Encoders / Decoders



Supervision: Masking



Sequences



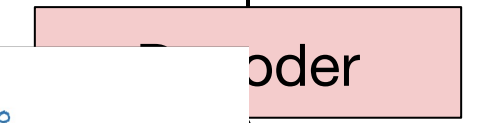
- What if we want our input/output to be *sequences*?
- E.g., sentences!
- Prev:
  - {a, c, d} -> {4, 3, 1},
- Want:
  - [a, c, d] -> [1, 3, 4]

Sequences: What *is* a Positional Encoding?

$$P(k, 2i) = \sin\left(\frac{k}{n^{2i/d}}\right)$$

$$P(k, 2i+1) = \cos\left(\frac{k}{n^{2i/d}}\right)$$

[1, 2, 3, 4]



Sequence	Index of token, k	Positional Encoding Matrix with d=4, n=100			
		i=0	i=0	i=1	i=1
I	0	P_{00}=sin(0) = 0	P_{01}=cos(0) = 1	P_{02}=sin(0) = 0	P_{03}=cos(0) = 1
am	1	P_{10}=sin(1/1) = 0.84	P_{11}=cos(1/1) = 0.54	P_{12}=sin(1/10) = 0.10	P_{13}=cos(1/10) = 1.0
a	2	P_{20}=sin(2/1) = 0.91	P_{21}=cos(2/1) = -0.42	P_{22}=sin(2/10) = 0.20	P_{23}=cos(2/10) = 0.98
Robot	3	P_{30}=sin(3/1) = 0.14	P_{31}=cos(3/1) = -0.99	P_{32}=sin(3/10) = 0.30	P_{33}=cos(3/10) = 0.96

Positional Encoding Matrix for the sequence 'I am a robot'

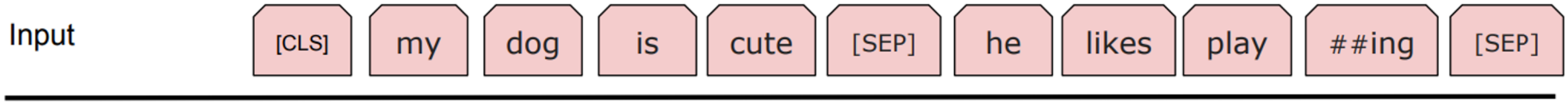
cos(p(0), p(1)) = 0.77		
cos(p(0), p(2)) = 0.28	cos(p(1), p(2)) = 0.77	
cos(p(0), p(3)) = -0.01	cos(p(1), p(3)) = 0.28	cos(p(2), p(3)) = 0.77

i3, i4]

- Creating a consistent "indicator vector" for each position
- "Nearness" between adjacent positions preserved by encoding functions.

https://machinelearningmastery.com/a-gentle-introduction-to-positional-encoding-in-transformer-models-part-1/

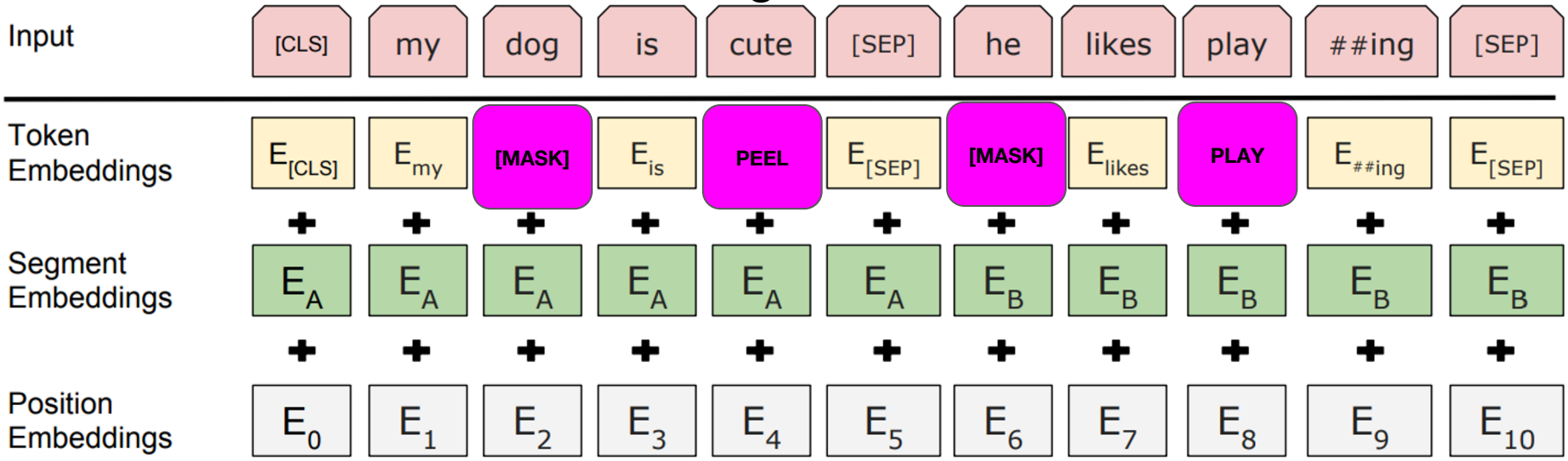
Bidirectional Encoder Representations from Transformers



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Masked Language Modeling Pretraining

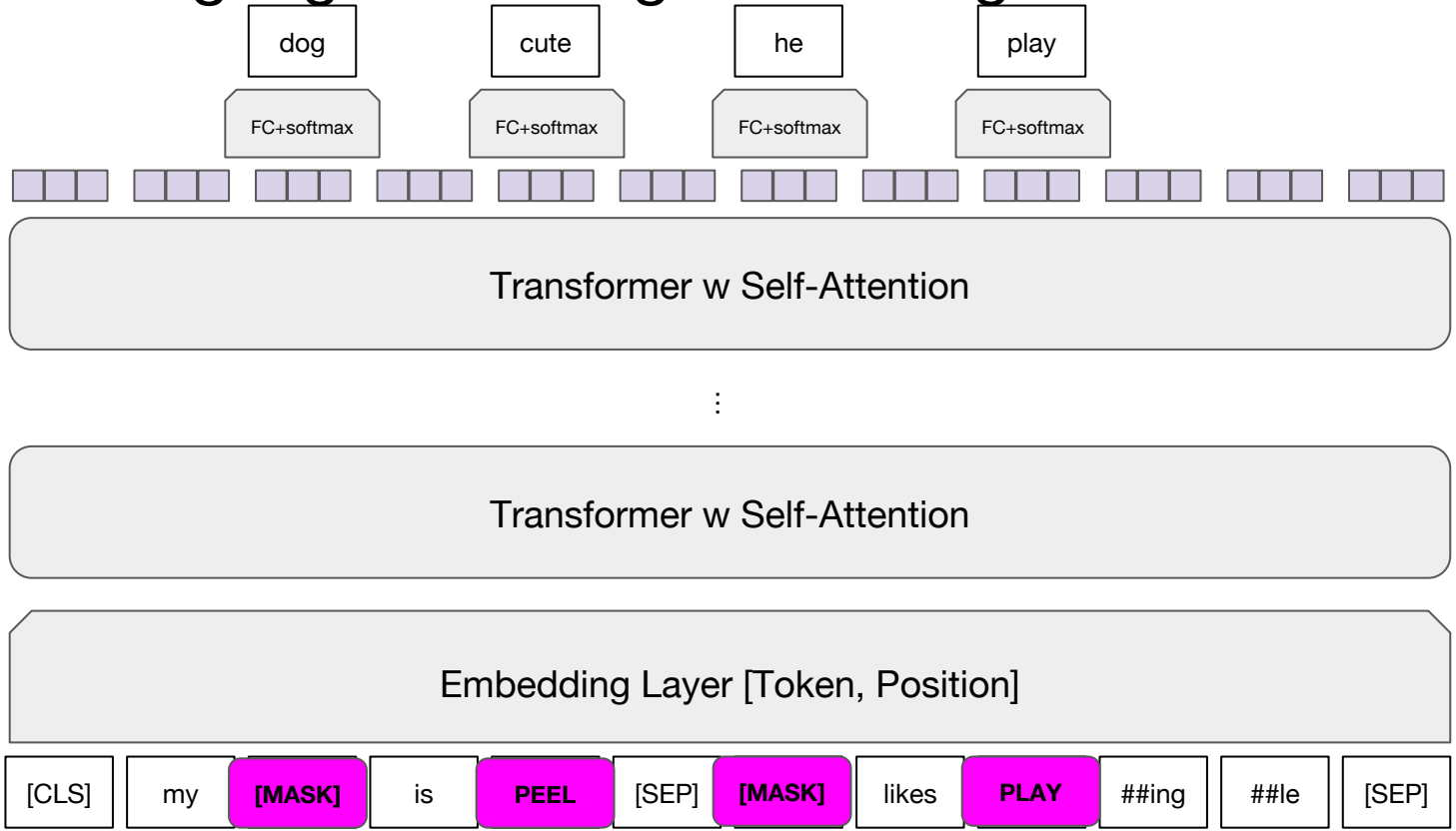
15% of words selected for training



80% of those [MASK] 10% random token 10% original token

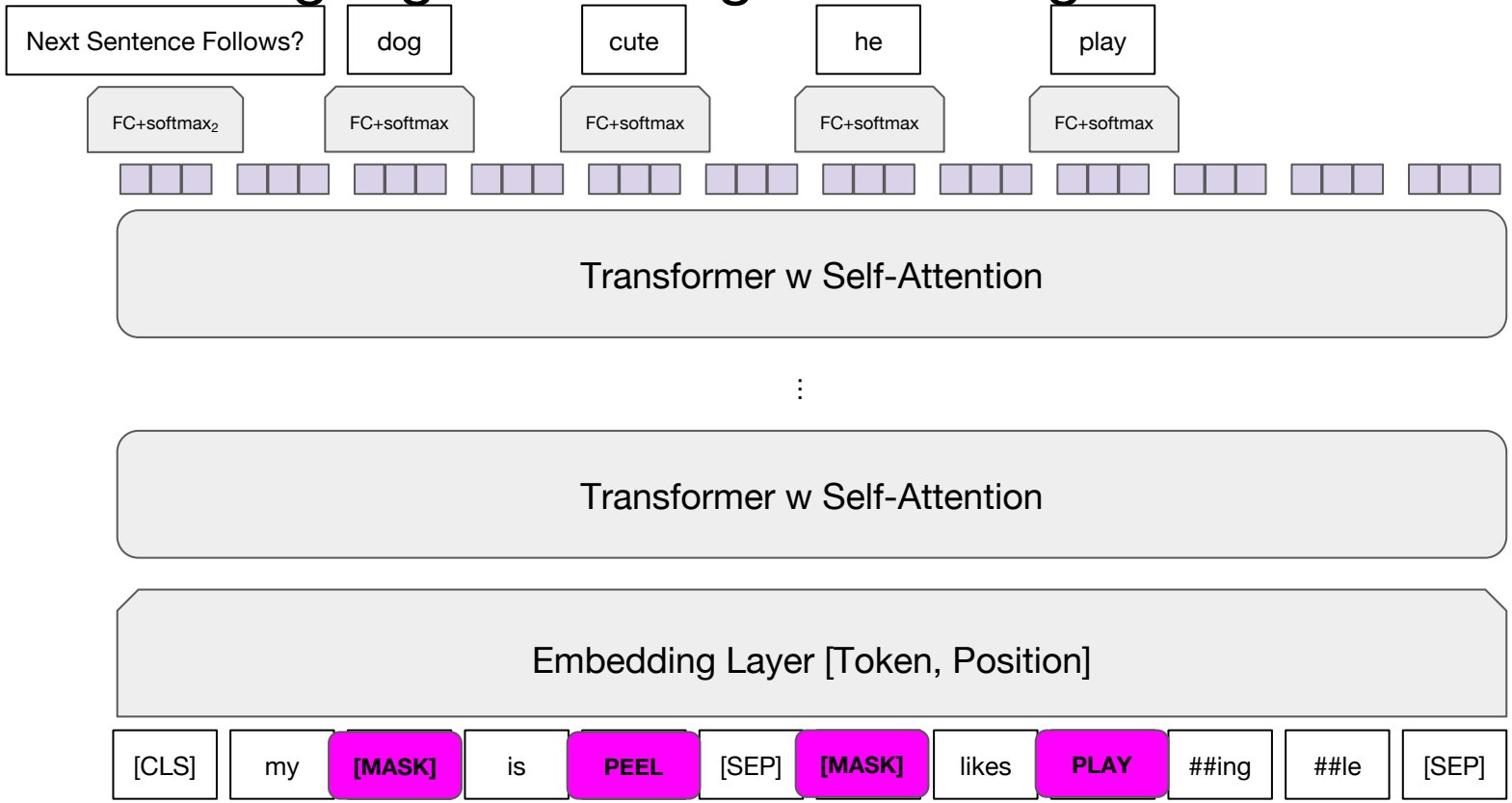
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Masked Language Modeling Pretraining



[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding]

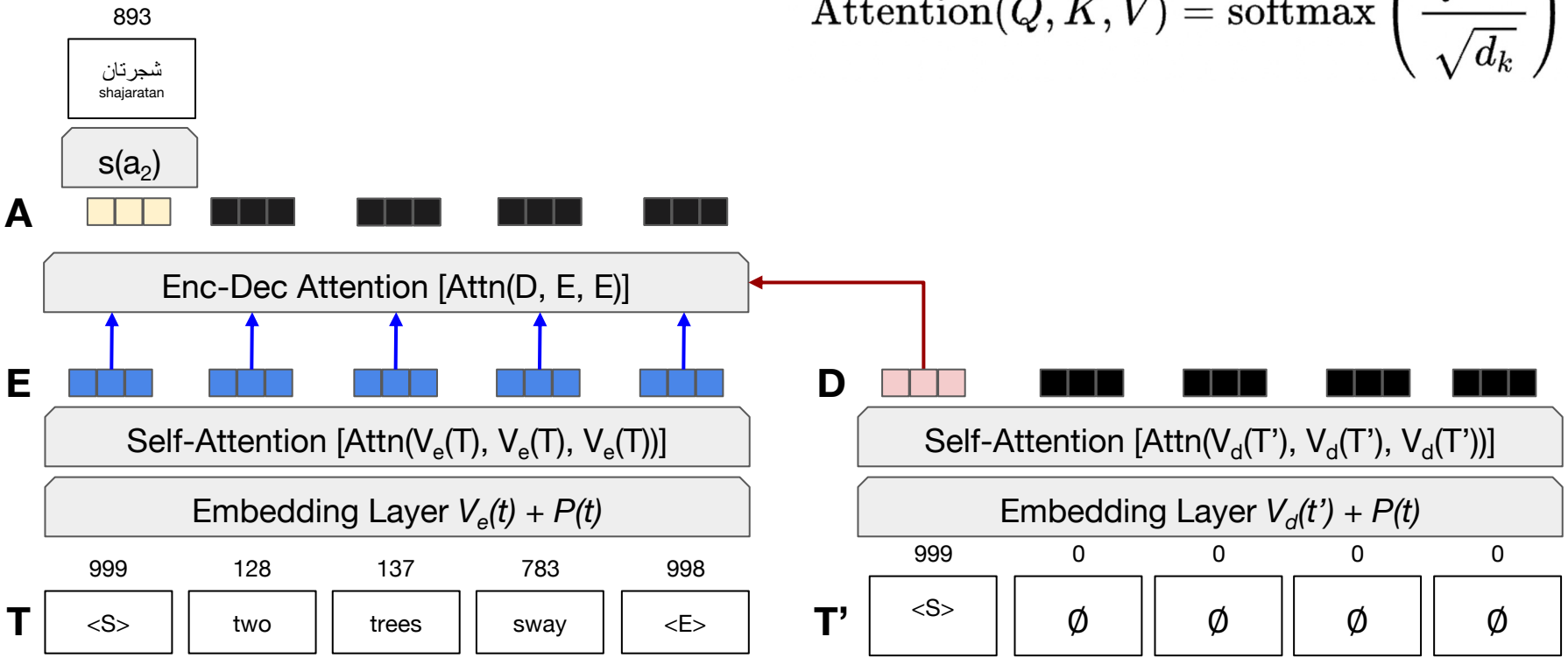
Masked Language Modeling Pretraining



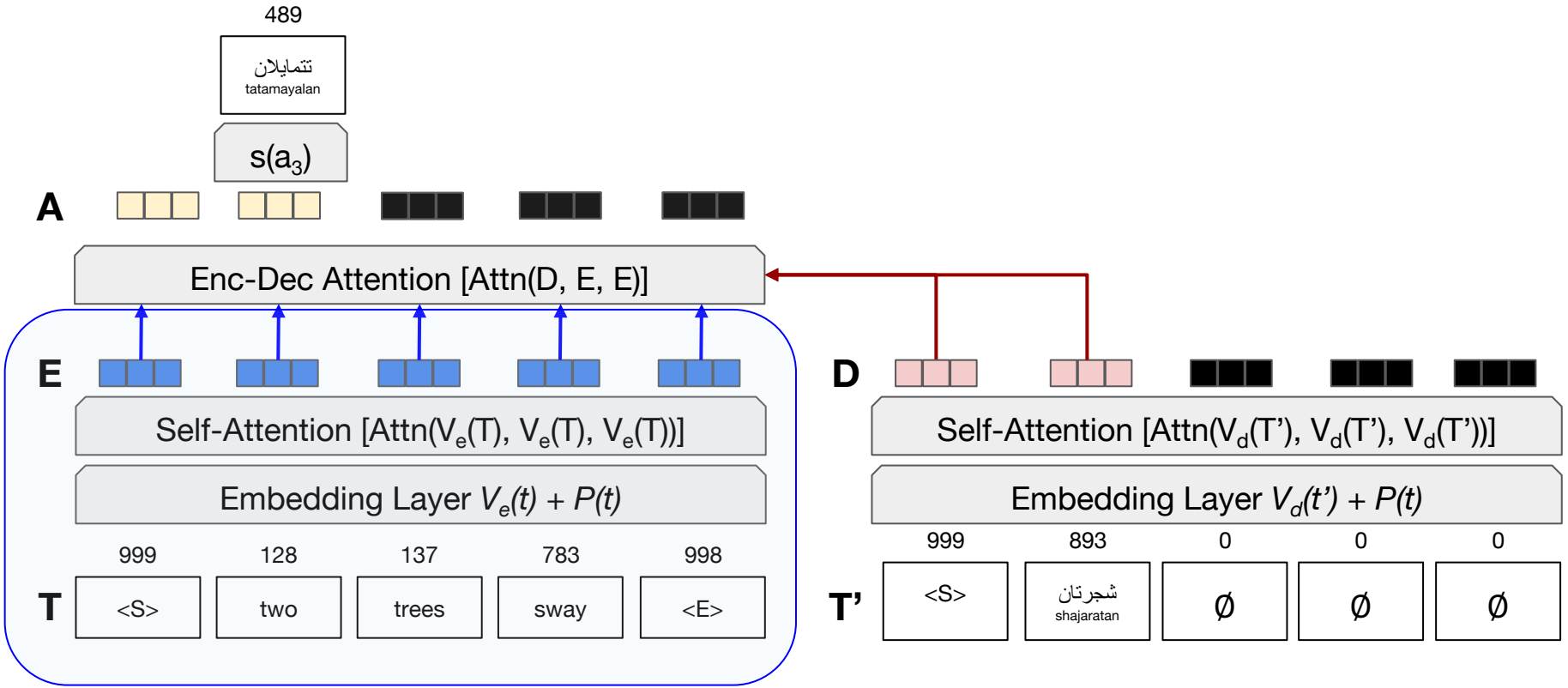
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Sequence-to-Sequence Transformer

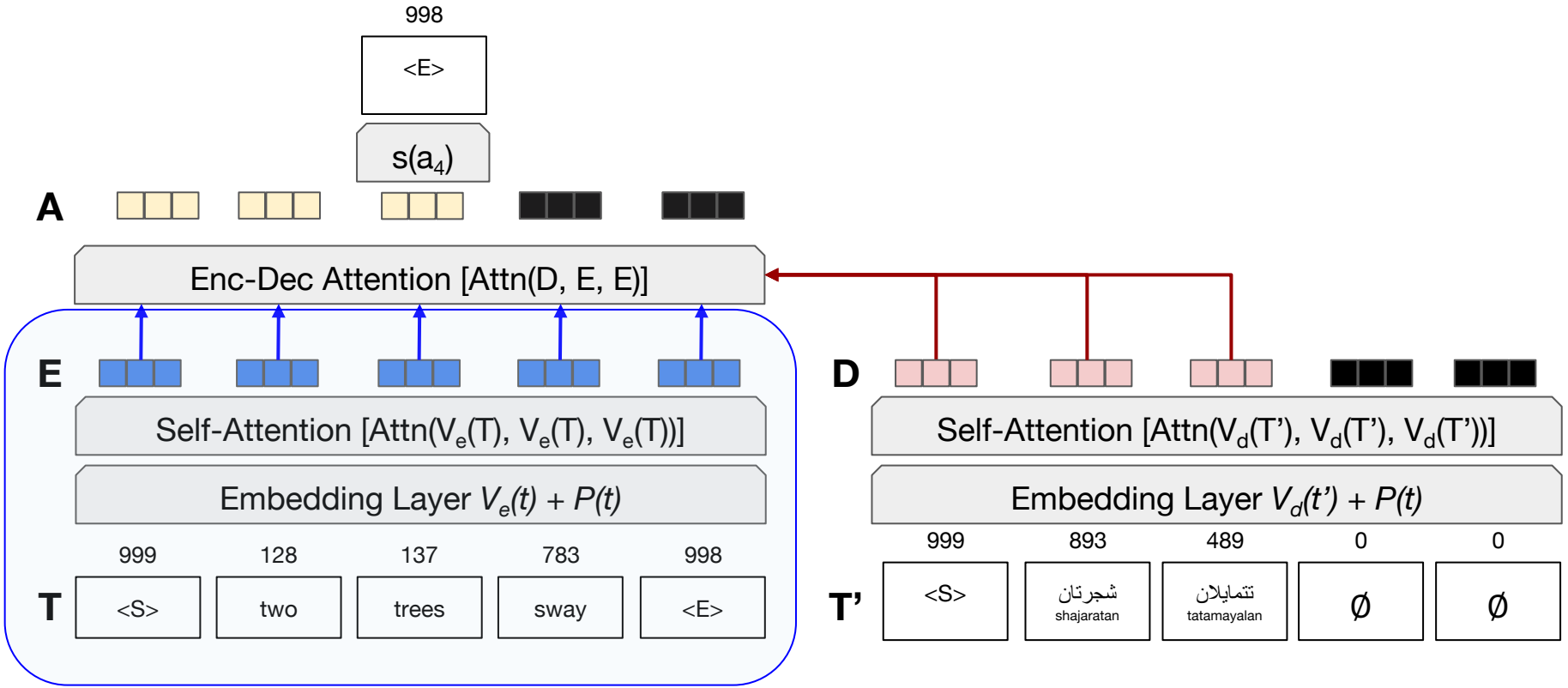
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V$$



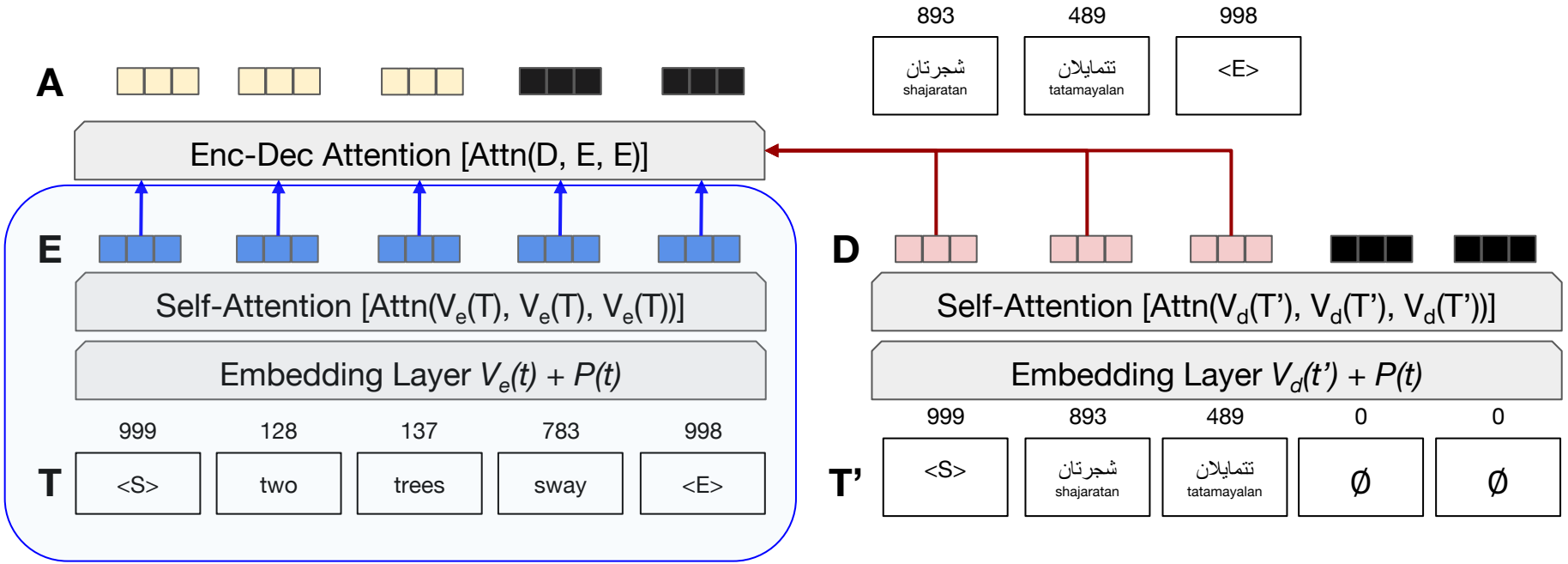
Sequence-to-Sequence Transformer



Sequence-to-Sequence Transformer



Sequence-to-Sequence Transformer



Transformers: The Future of the Present

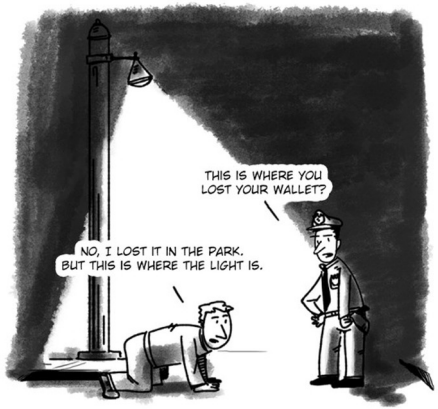
- Transformers *seem* like they might be the universal encoders we've been waiting for in multimodal space
- Language tokens, image regions, image patches, audio snippets, joint positions, transformers don't care!

Overview of Today's Plan

- ~~Course organization and deliverables~~
- ~~Lecture 7 Recap~~
  - Any questions before we move on?
- Multimodal Deep Learning

Streetlight Effect

A policeman sees a drunk man searching for something under a streetlight and asks what the drunk has lost. He says he lost his keys and they both look under the streetlight together. After a few minutes the policeman asks if he is sure he lost them here, and the drunk replies, no, and that he lost them in the park. The policeman asks why he is searching here, and the drunk replies, "this is where the light is".



https://en.wikipedia.org/wiki/Streetlight_effect

Images and Text for Retrieval

Text Query

"A tropical bird perches in the jungle."

Candidate Images






Images and Text for Retrieval

Image Query

Candidate Captions

“A rabbit sits in the palm of a hand.”

“Men are talking on a basketball court.”

“A tropical bird perches in the jungle.”

“Children play soccer in a field.”

“A white fox is looking at the camera.”

“Sports equipment is staged for a photo.”

Formulating the Retrieval Problem

Images	$\mathcal{I}$
Captions	$\mathcal{L}$
Scoring Function	$F : \mathcal{I} \times \mathcal{L} \to \mathbb{R}$

Formulating the Retrieval Problem

$$F : \mathcal{I} \times \mathcal{L} \rightarrow \mathbb{R}$$

F(# red pixels, # word “red”) = 1 if (a > 0, b > 0) else 0

F(RGB bins, color word counts) =
sum of #(color word, bin)/(#color word + #bin) in training data

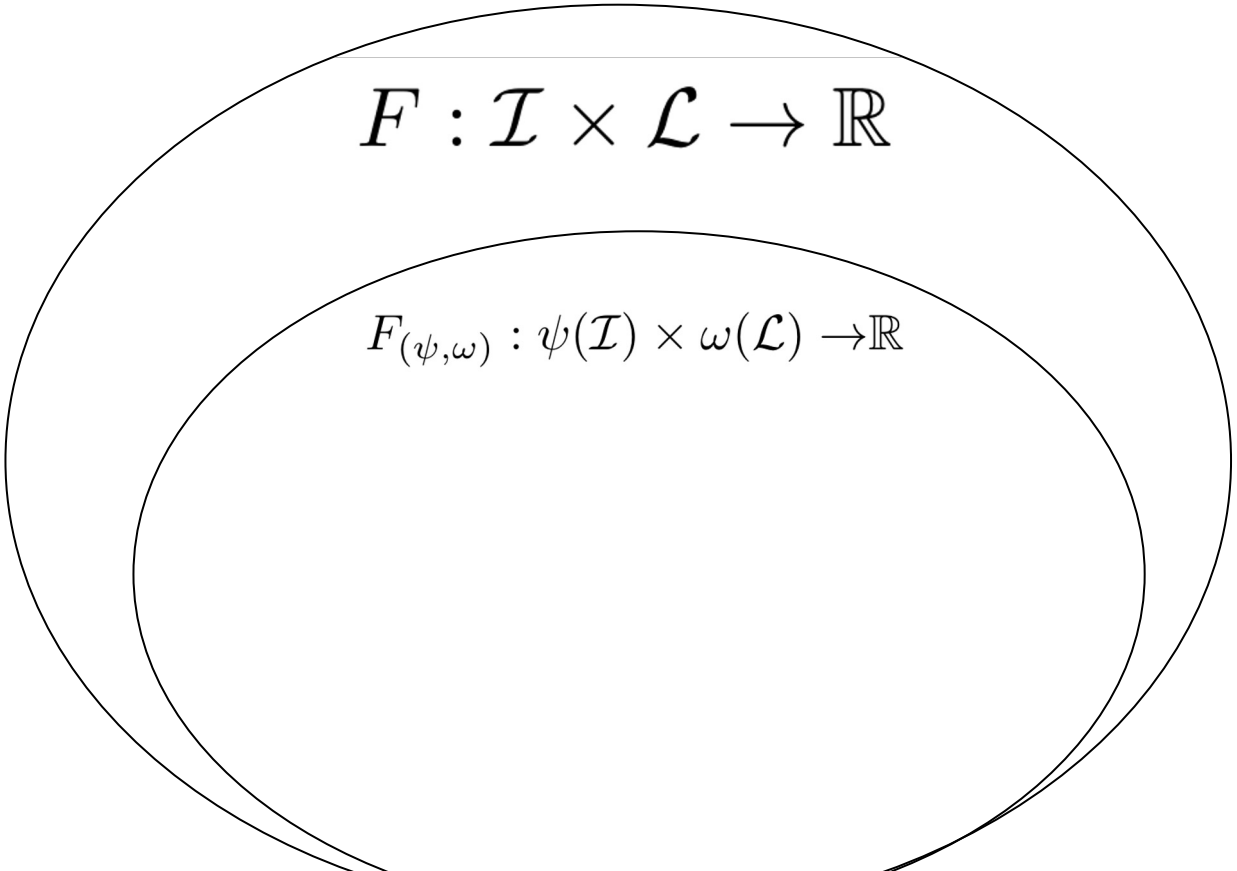
F(detected objects, word counts) =
sum of #(object, word)/(#object + #word) in training data

F(image pixels, token sequence) =
NN trained with contrastive matching loss

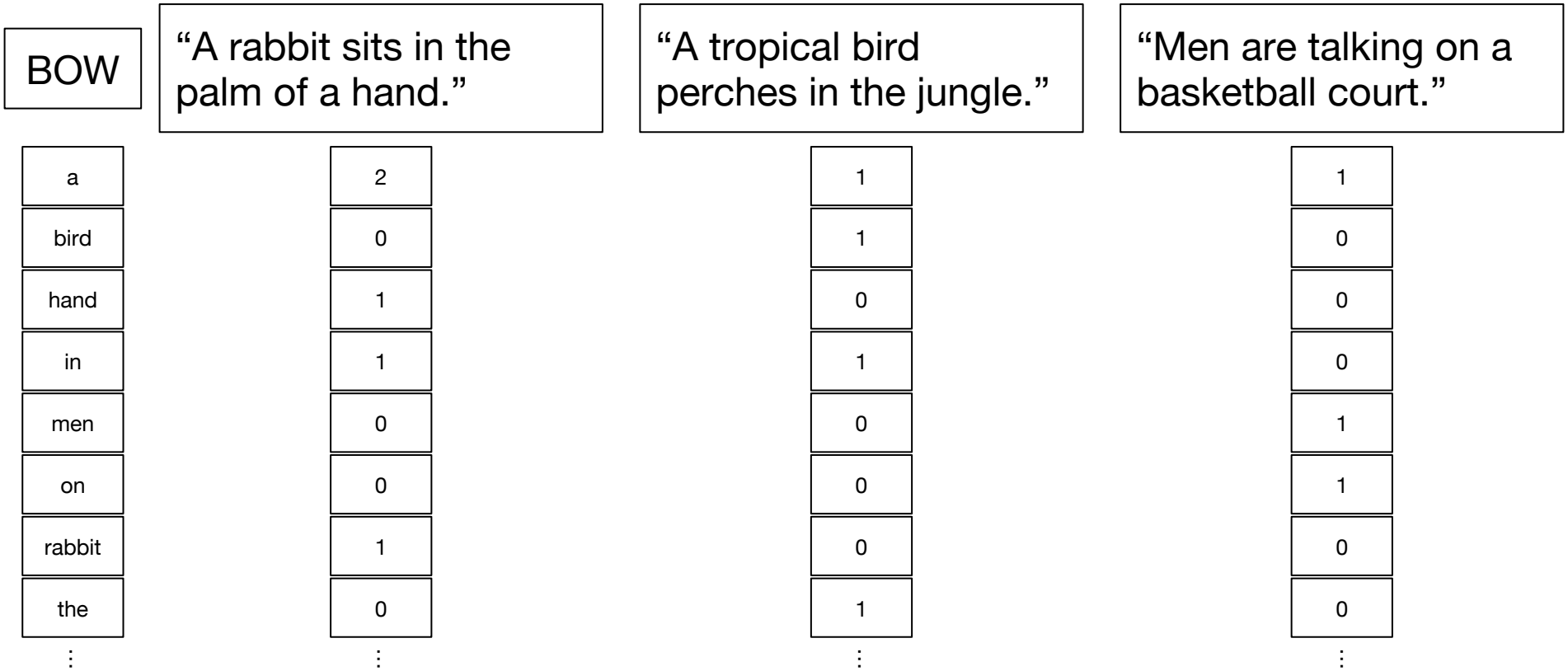
Formulating the Retrieval Problem

Image Features	$\psi : \mathcal{I} \rightarrow \mathbb{R}^n$
Caption Features	$\omega : \mathcal{L} \rightarrow \mathbb{R}^m$
Scoring Function	$F_{(\psi,\omega)} : \psi(\mathcal{I}) \times \omega(\mathcal{L}) \rightarrow \mathbb{R}$

Formulating the Retrieval Problem

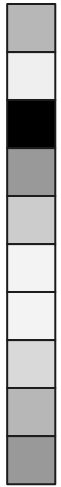


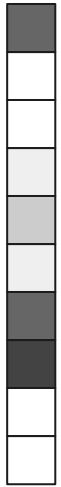
Feature Extraction for Language



Feature Extraction for Vision

RGB
kNN bins



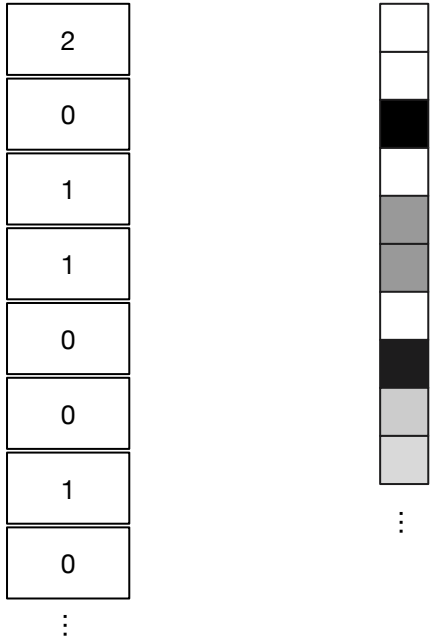


A Simple Retrieval Solution

$$\psi : \mathcal{I} \rightarrow \mathbb{R}^n$$

$$\omega : \mathcal{L} \rightarrow \mathbb{R}^m$$

$$F_{(\psi,\omega)} : \psi(\mathcal{I}) \times \omega(\mathcal{L}) \rightarrow \mathbb{R}$$



Contextual Information

You shall know a word by the company it keeps (Firth, J. R. 1957:11)

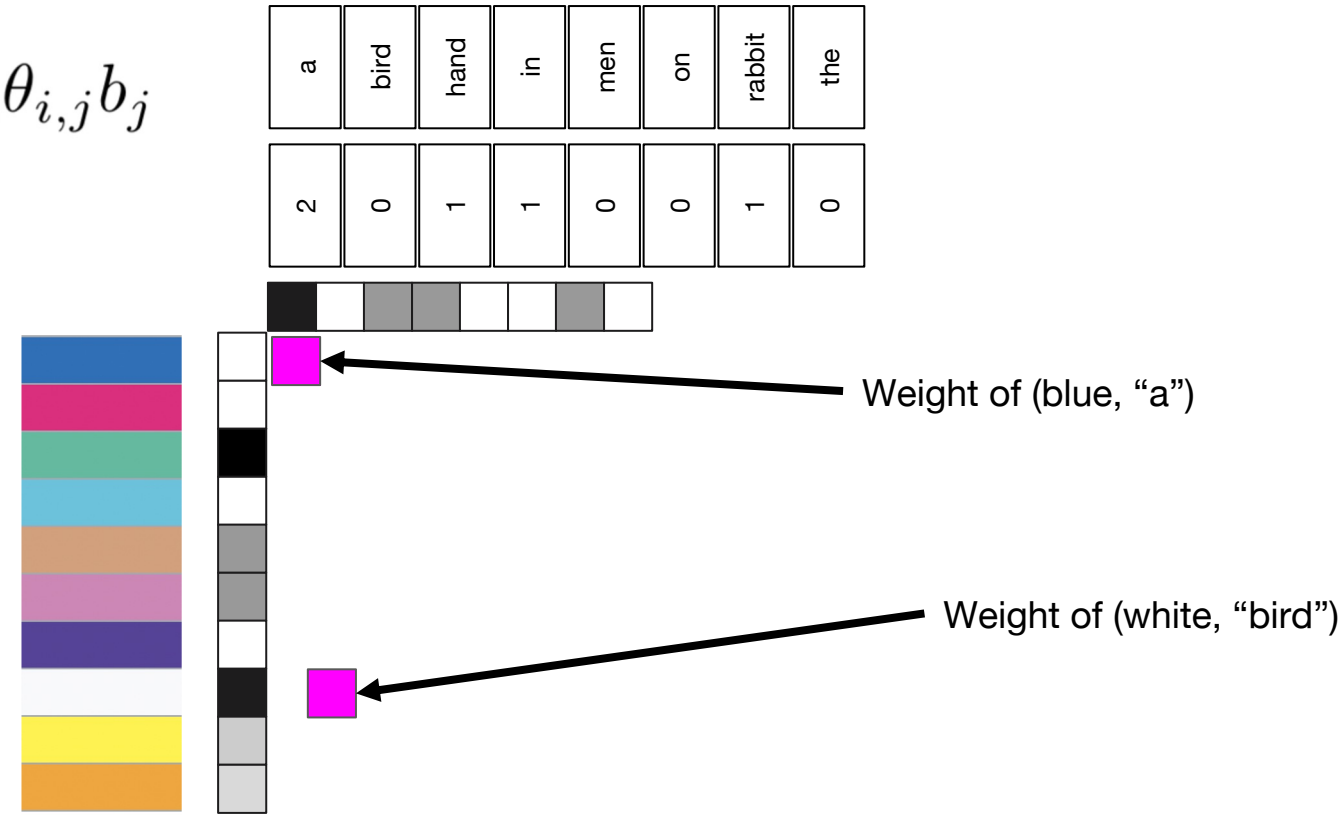
You shall know the features of modality A by the company they keep in the features of modality B. And vice versa.

Formulating the Retrieval Problem as a Linear Model

$$F_{(\psi,\omega)} : \psi(\mathcal{I}) \times \omega(\mathcal{L}) \rightarrow \mathbb{R}$$

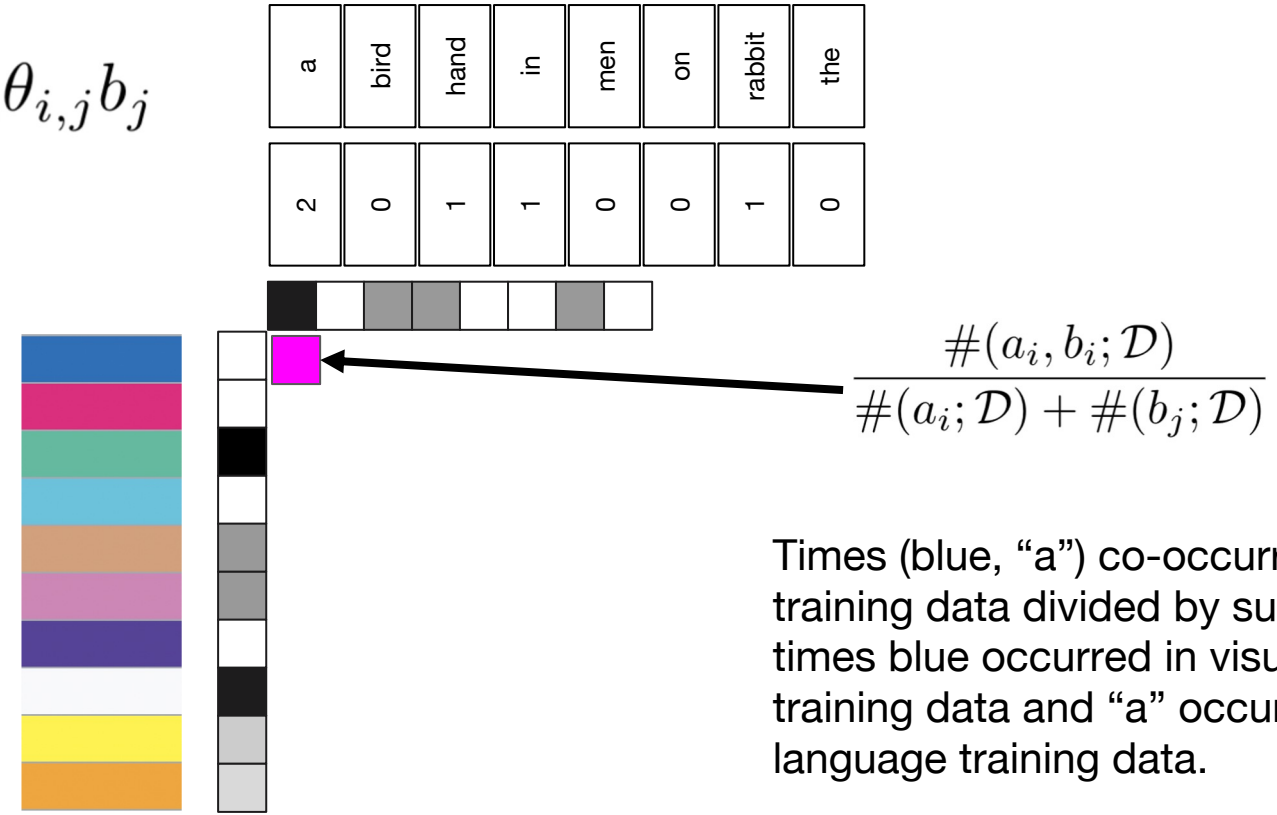
Formulating the Retrieval Problem as a Linear Model

$$\sum_{i=1,j=1}^{n,m} a_i \theta_{i,j} b_j$$



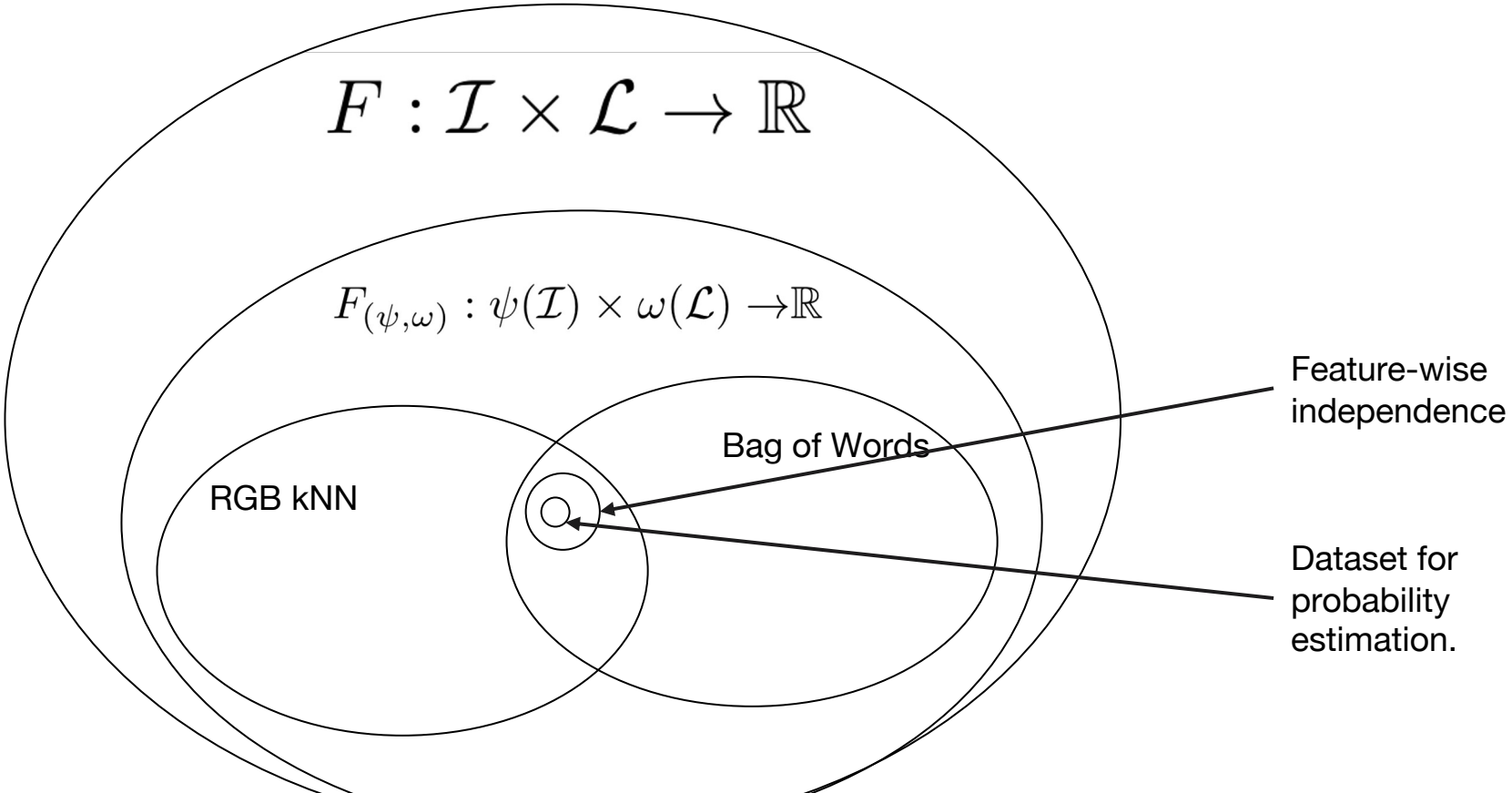
Formulating the Retrieval Problem as a Linear Model

$$\sum_{i=1,j=1}^{n,m} a_i \theta_{i,j} b_j$$



Times (blue, “a”) co-occurred in training data divided by sum of times blue occurred in visual training data and “a” occurred in language training data.

Formulating the Retrieval Problem



Tokenization Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

“Perching parakeet in a wire frame cage.”

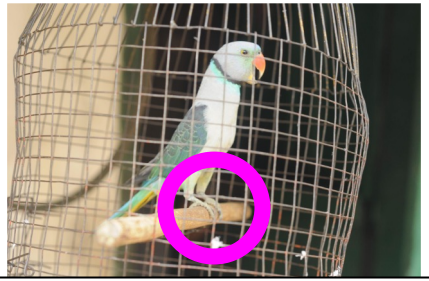
“A snow owl lands on a wooden perch.”

Stemming

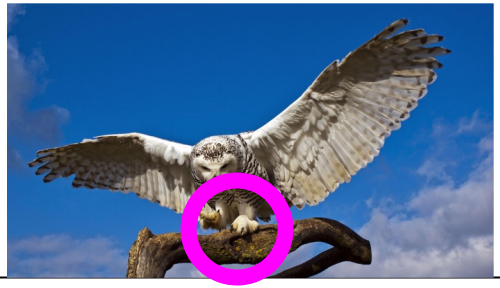
a	tropic	bird	perch	in	the	jungl	.
perch	parakeet	in	a	wire	frame	cage	.
a	snow	owl	land	on	wood	perch	.

Tokenization Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”



“Perching parakeet in a wire frame cage.”



“A snow owl lands on a wooden perch.”

Stemming	a	tropic	bird	**perch**	in	the	jungl	.
	perch	parakeet	in	a	wire	frame	cage	.
	a	snow	owl	land	on	wood	**perch**	.

Tokenization Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

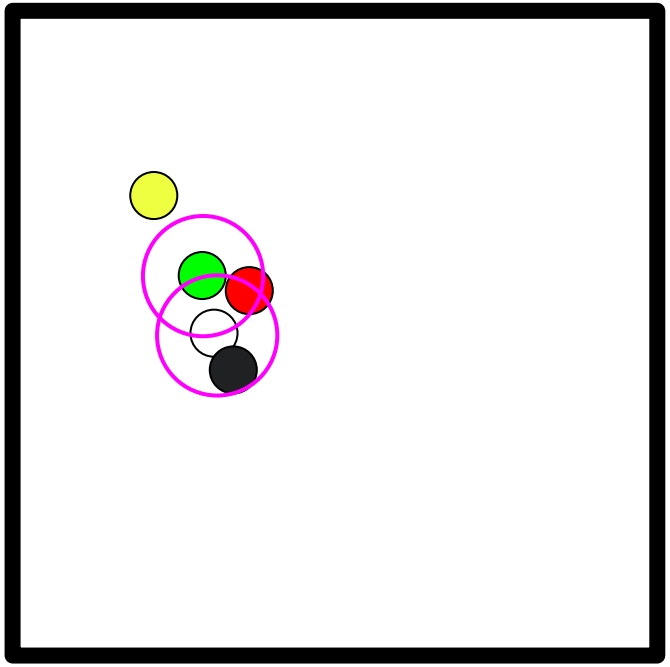
a	tropic	bird	perch	in	the	jungl	.
perch	parakeet	in	a	wire	frame	cage	.
a	snow	owl	land	on	wood	perch	.

Pretrained Language Token Embeddings

- Cosine similarity of “bird”, “owl”, “parakeet” helps share information across training data
- Taking a guess: what are the nearest neighbors of “green” in word2vec embedding space?
  - Blue, white, red, yellow, black, grey, purple, pink, light, gray
- What can we learn for highly polysemous words like “play”?
  - “play guitar”, “play piano”, “play basketball”, “play tag”
- Text-based embeddings help us share information ***only to the extent that words used in a similar context share a visual representation***
  - which is true for, say, birds or trees, but not colors

Pretrained Language Token Embeddings

$\omega(\mathcal{L})$



Training

Inference

“A white bunny rabbit held in a green field.”

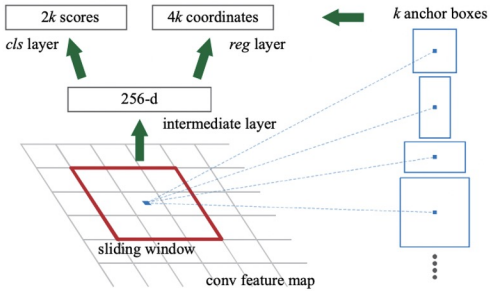
“A black rabbit watches a red sunset.”

Feature Extraction for Vision

Object Classification

AlexNet, VGG, ResNet, …

Object Detection



Faster R-CNN, YOLO, …

[Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, 2015.]

Feature Extraction for Vision

Object Classification

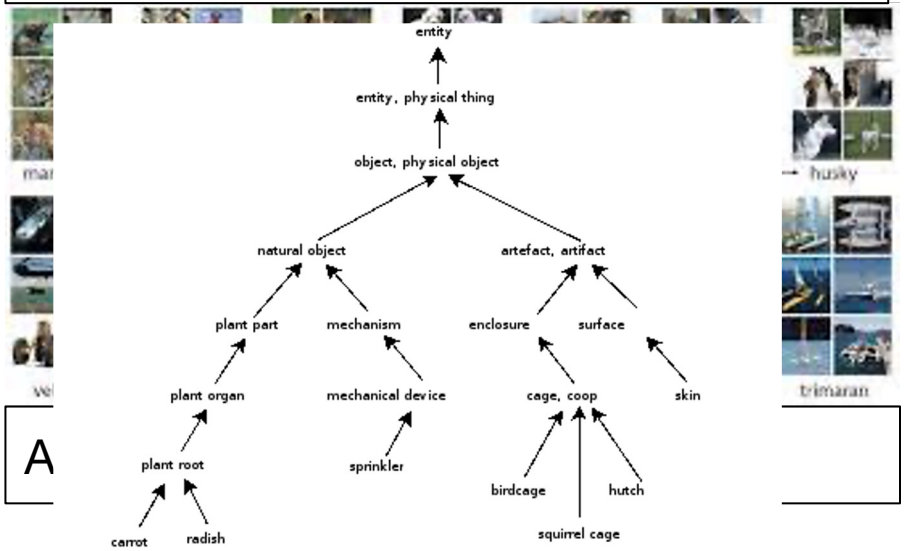
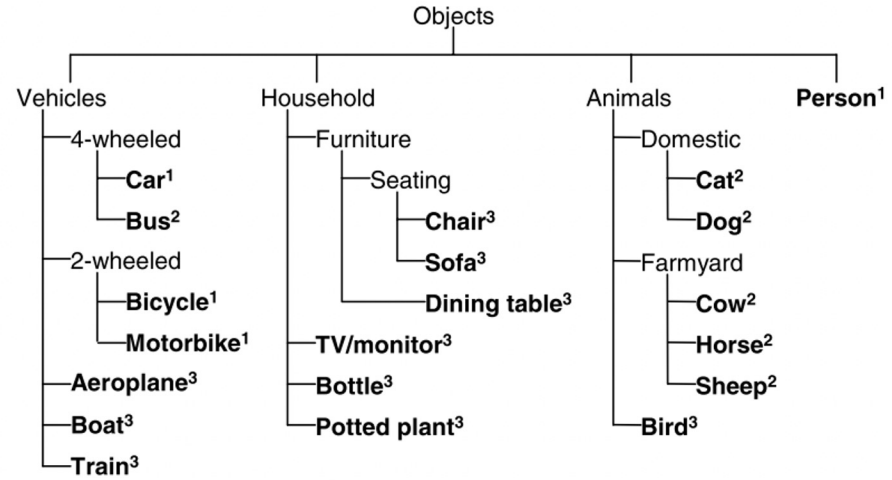


Figure 1. "is a" relation example

Object Detection



[The PASCAL Visual Object Classes (VOC) Challenge, 2010.]

Image Feats. Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

Object Classification

toucan

cockatoo

great grey owl

https://deeplearning.cms.waikato.ac.nz/user-guide/class-maps/IMAGENET/

Image Feats. Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

“Perching parakeet in a wire frame cage.”

“A snow owl lands on a wooden perch.”

Object Detection

bird	potted plant
bird	boat
bird	aeroplane

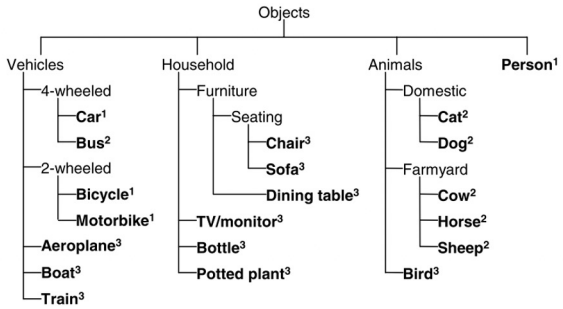
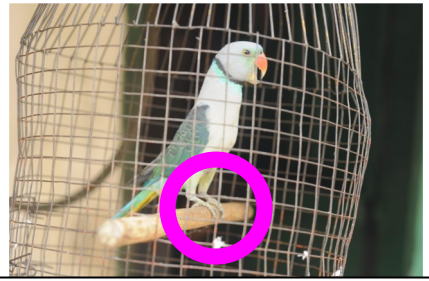
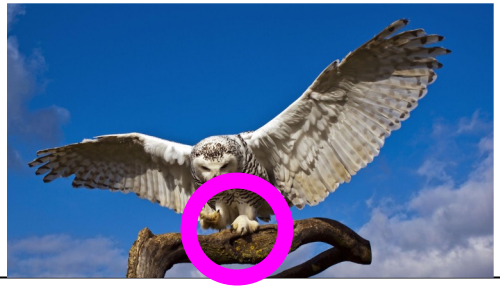


Image Feats. Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”



“Perching parakeet in a wire frame cage.”



“A snow owl lands on a wooden perch.”

bird	potted plant
bird	boat
bird	aeroplane

a	tropic	bird	**perch**	in	the	jungl	.
perch	parakeet	in	a	wire	frame	cage	.
a	snow	owl	land	on	wood	**perch**	.

Image Feats. Considering Language and Vision Retrieval

“A tropical bird perches in the jungle.”

bird	potted plant
bird	boat
bird	aeroplane

“Perching parakeet in a wire frame cage.”

a	tropic	bird	perch	in	the	jungl	
perch	parakeet	in	a	wire	ame	cage	.
a	snow	owl	land	on	wood	rch	.

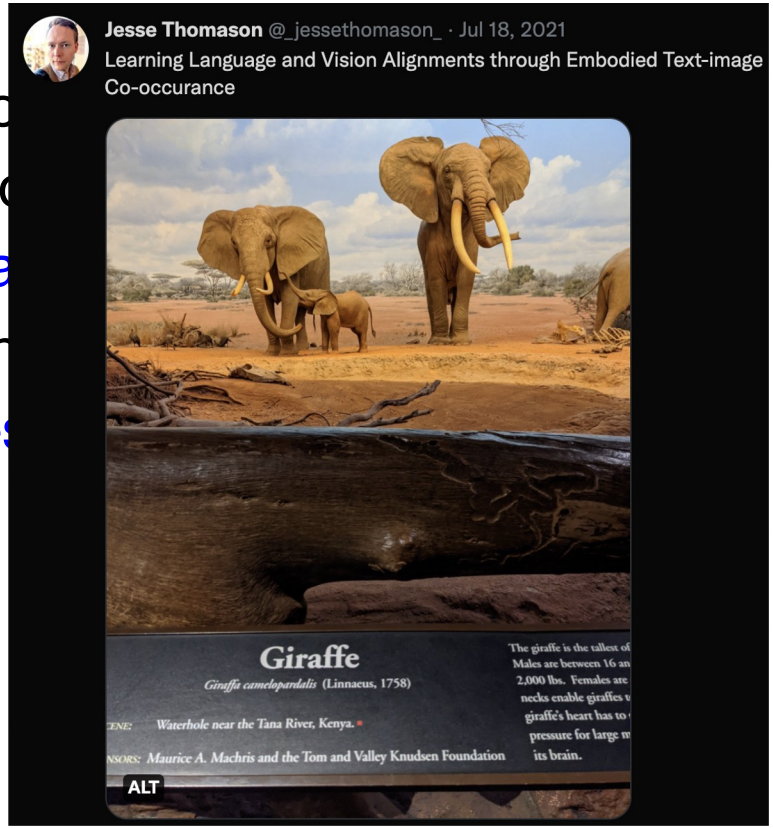
“A sn owl lands n a w oden perch.”

Mitigating Issues With Image Feat. Extraction

- CNNs trained on object classification
  - Penultimate layer as a feature vector
- CNNs trained to do object detection
  - Local penultimate feature reps of each bounding box
- Patch-based alternatives for transformers
  - Learn transformations of small image patches as “embeddings” for a transformer.
- Learning without labels: directly align caption and images

What We Glossed Over

- There are *many* non-neural ways to
  and vision features that span decad
  - POS, fast point feature histogra
- How can we build PMI into loss fun
  - Hinge losses; contrastive losse
- Co-occurrence isn't ground truth



Conditional Discriminative Models for Retrieval (ImSearch)

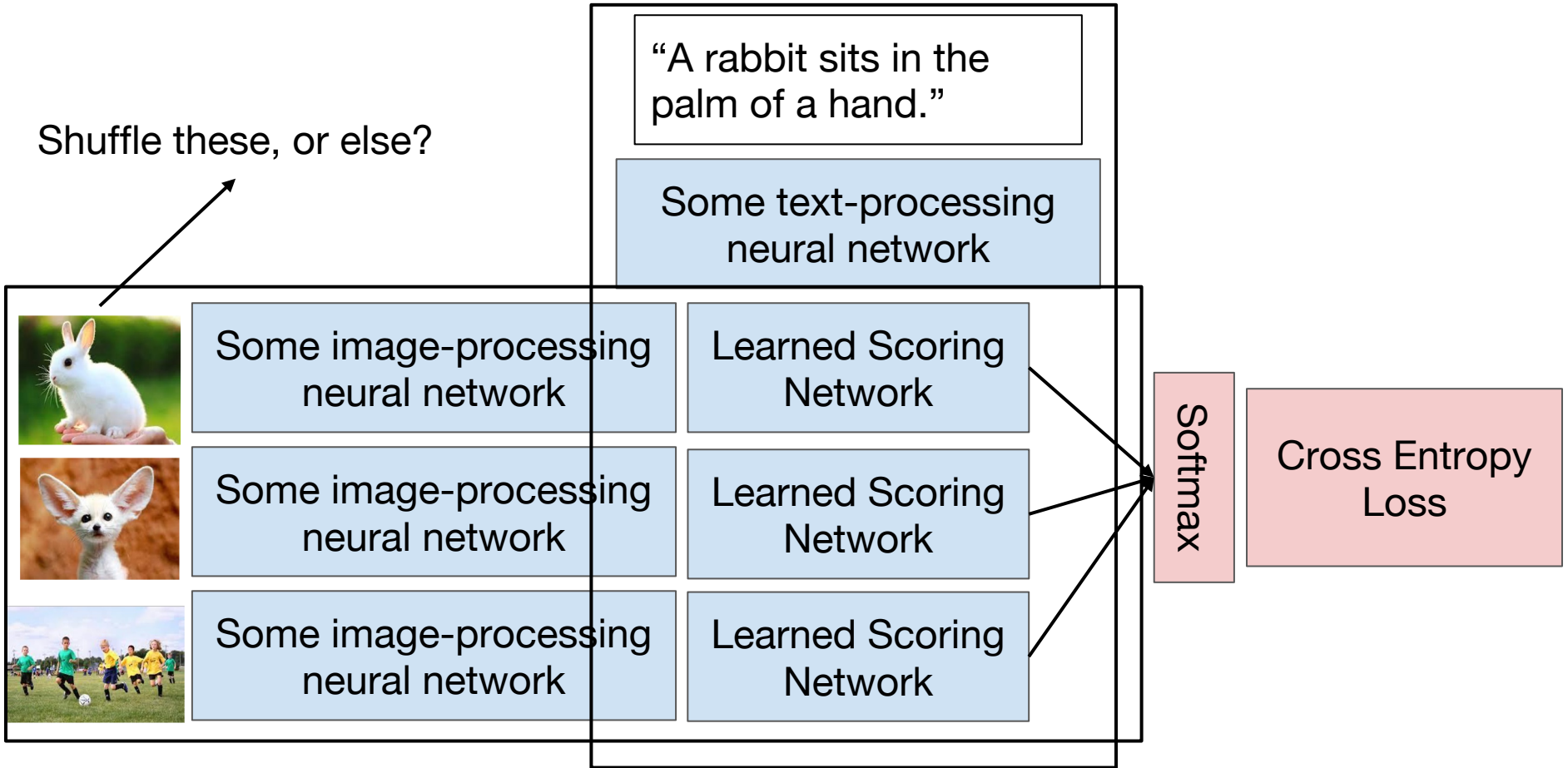
Training Data	
	“A rabbit sits in the palm of a hand.”
	“A white fox is looking at the camera.”
	“Children play soccer in a field.”
	“Men are talking on a basketball court.”

$$F : \mathcal{I} \times \mathcal{L} \rightarrow \mathbb{R}$$

$$P(X|Y)$$

- Maximize the probability of the image X given the caption Y in the training data.
- Example: for each training caption, batch it with true image and distractor images, then use cross entropy.

Conditional Discriminative Models for Retrieval (ImSearch)



Conditional Discriminative Models for Retrieval (ImSearch)

Training Data	
	“A rabbit sits in the palm of a hand.”
	“A white fox is looking at the camera.”
	“Children play soccer in a field.”
	“Men are talking on a basketball court.”

$$F : \mathcal{I} \times \mathcal{L} \rightarrow \mathbb{R}$$

$$P(X|Y) \qquad P(Y|X)$$

- Are these expressions symmetric?
- P(X|Y) learn which words indicate which visual features.
- P(Y|X) learn which visual features indicate which words.
- Iff symmetric,
  $P(a|b)=P(b|a)$ for all a,b

Conditional Discriminative Models for Retrieval

Training Data	
dog	“terrier”
dog	“husky”
guitar	“play”
frisbee	“play”

$$F : \mathcal{I} \times \mathcal{L} \rightarrow \mathbb{R}$$

$P(X|Y)$

“play”
guitar
frisbee

$P(Y|X)$

guitar
“play”

$$P(X|Y)P(Y) = P(Y|X)P(X)$$

P(guitar|”play”)=0.5 …*P(“play”)=0.25

P(“play”|guitar)=1 …*P(guitar)=.25

Estimating Joint Distributions

Image Features	$\psi : \mathcal{I} \rightarrow \mathbb{R}^n$
Caption Features	$\omega : \mathcal{L} \rightarrow \mathbb{R}^m$
Scoring Function	$F_{(\psi,\omega)} : \psi(\mathcal{I}) \times \omega(\mathcal{L}) \rightarrow \mathbb{R}$

Estimating Joint Distributions

Scoring Function $F_{(\psi,\omega)} : \psi(\mathcal{I}) \times \omega(\mathcal{L}) \rightarrow \mathbb{R}$

- Define F as cosine similarity between image and caption embeddings
- Then we can learn:
  - (minimally) a projection network that maps image embeddings and caption embeddings to the same space
  - (bonus) an image feature extractor (psi)
  - (bonus) a language feature extractor (omega)

"A rabbit sits in the palm of a hand."

"Children play soccer in a field."

"A white fox is looking at the camera."

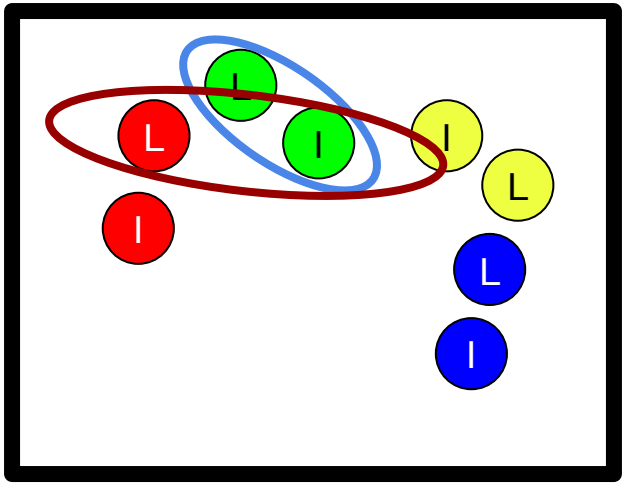
"Men are talking on a basketball court."

Image Embedder $\psi(\mathcal{I})$

Caption Embedder $\omega(\mathcal{L})$

Pull matching imag embeddings together.

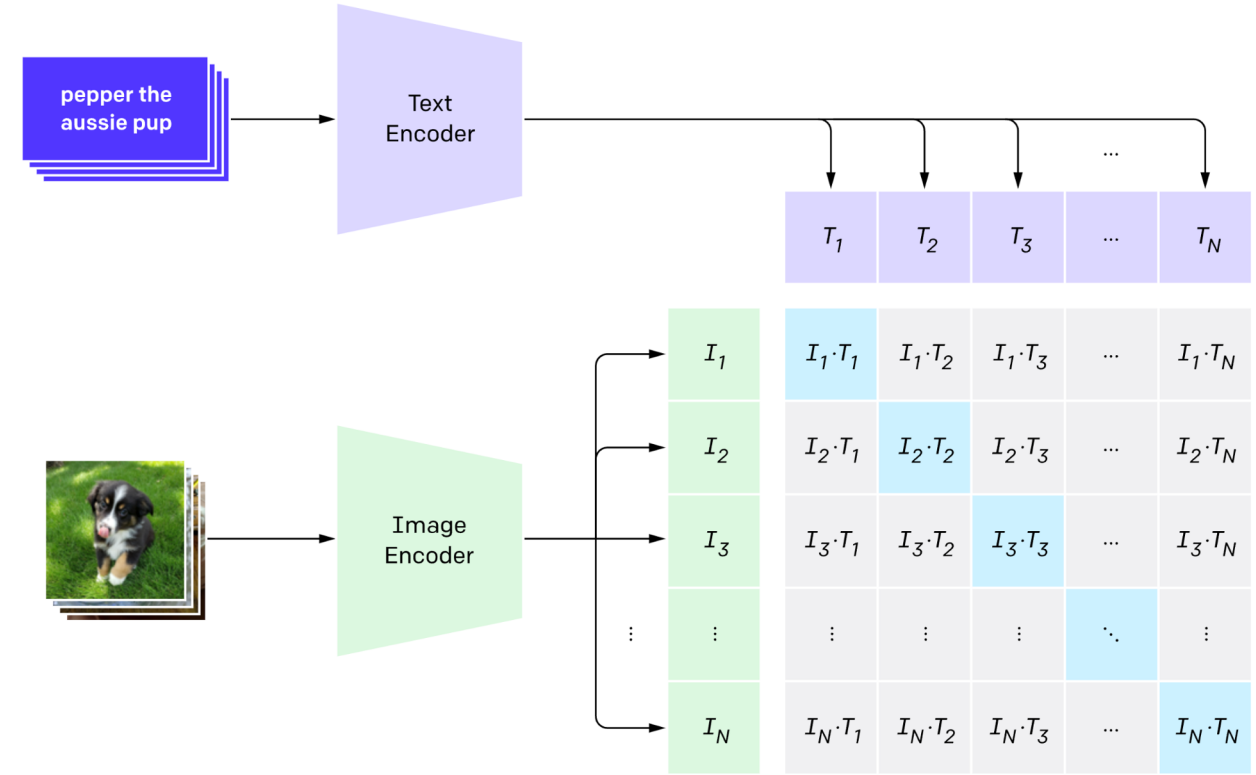
Learned Projection



Push distractor on embeddings apart.

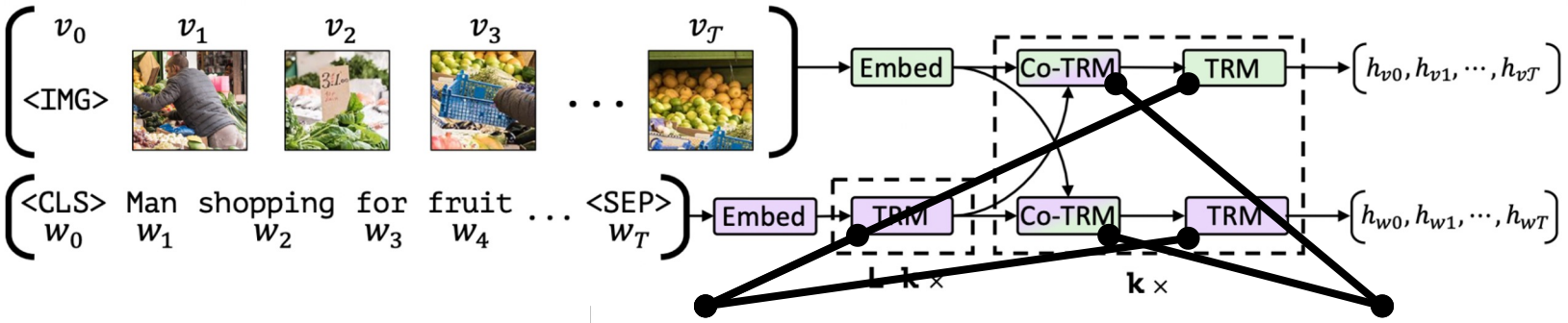
Learned Projection

Contrastive Language–Image Pre-training (CLIP)

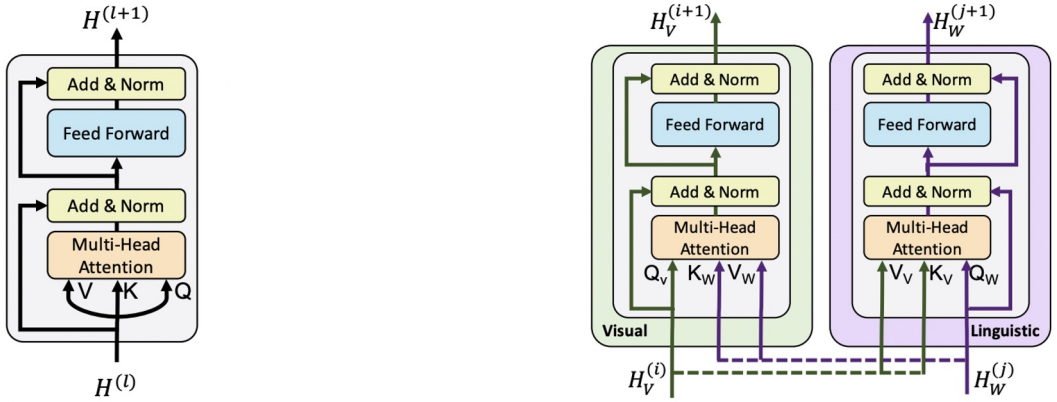


https://openai.com/blog/clip/

Language and Vision: Regions

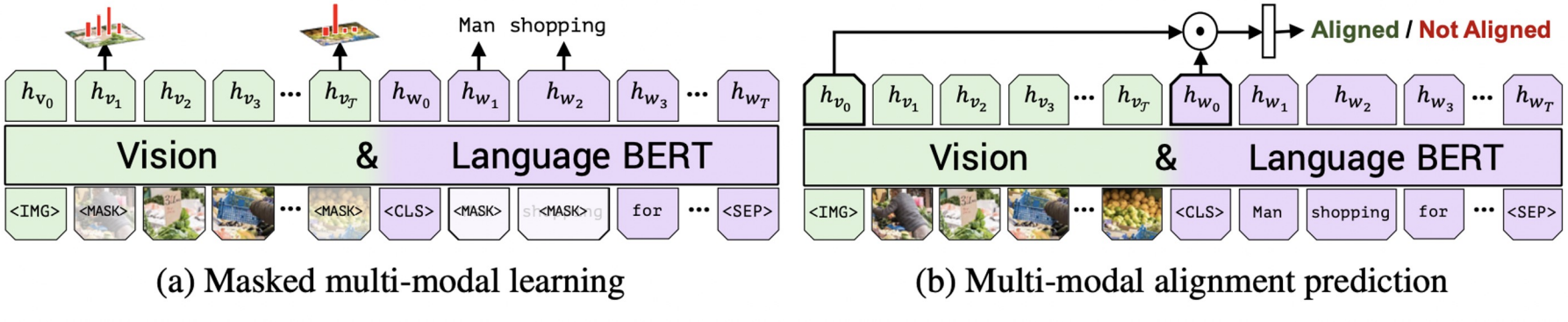


- Word attentions H_W as keys/values for image queries
- Region attentions H_V as keys/values for language queries



ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

Language and Vision: Regions

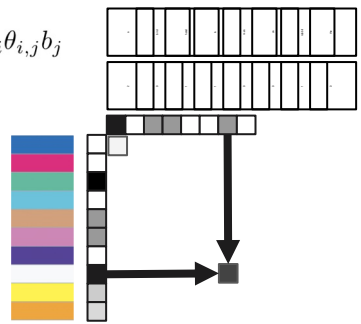


(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

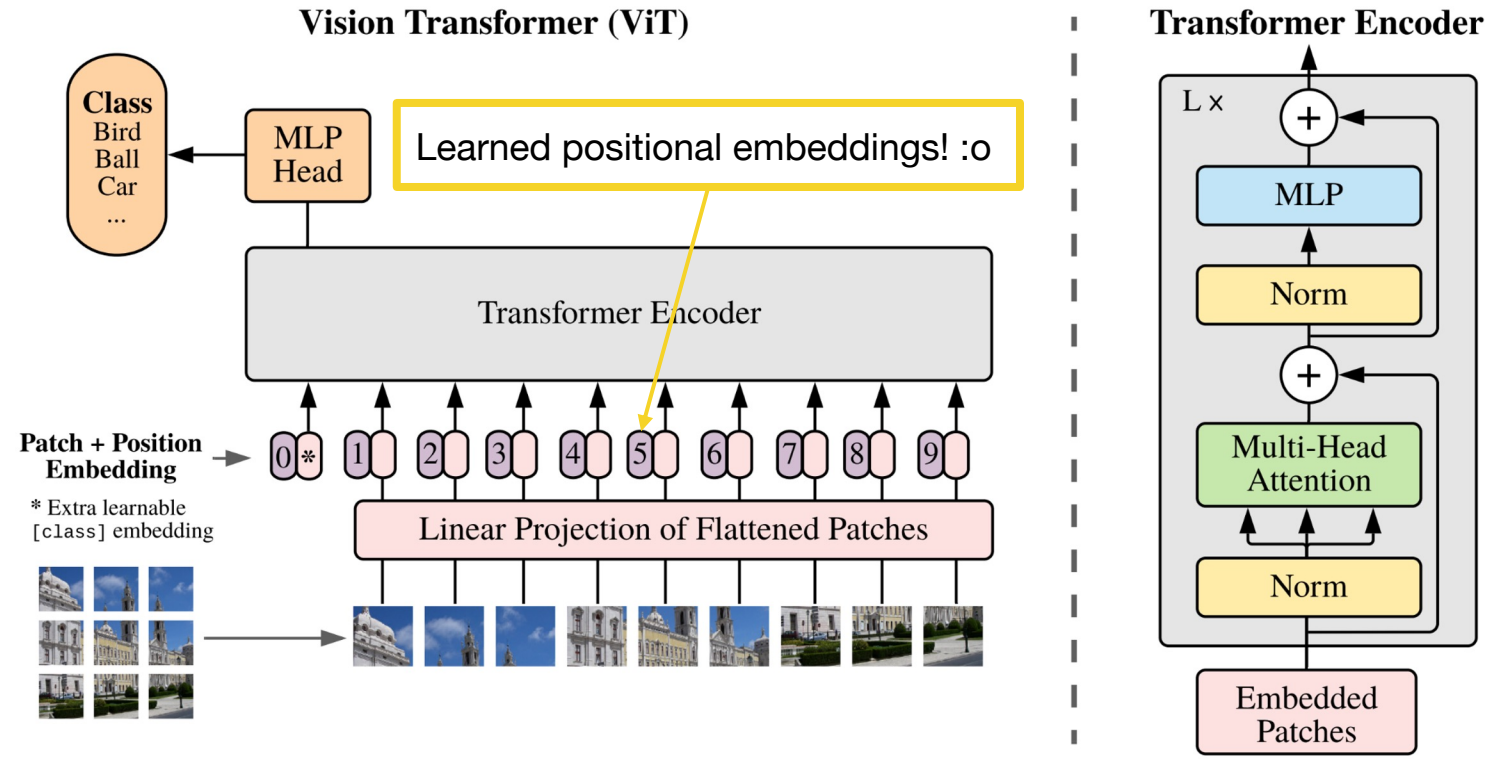
- Training outputs to have high cosine similarity is the same core idea as our lil linear model

$$\sum_{i=1,j=1}^{n,m} a_i \theta_{i,j} b_j$$



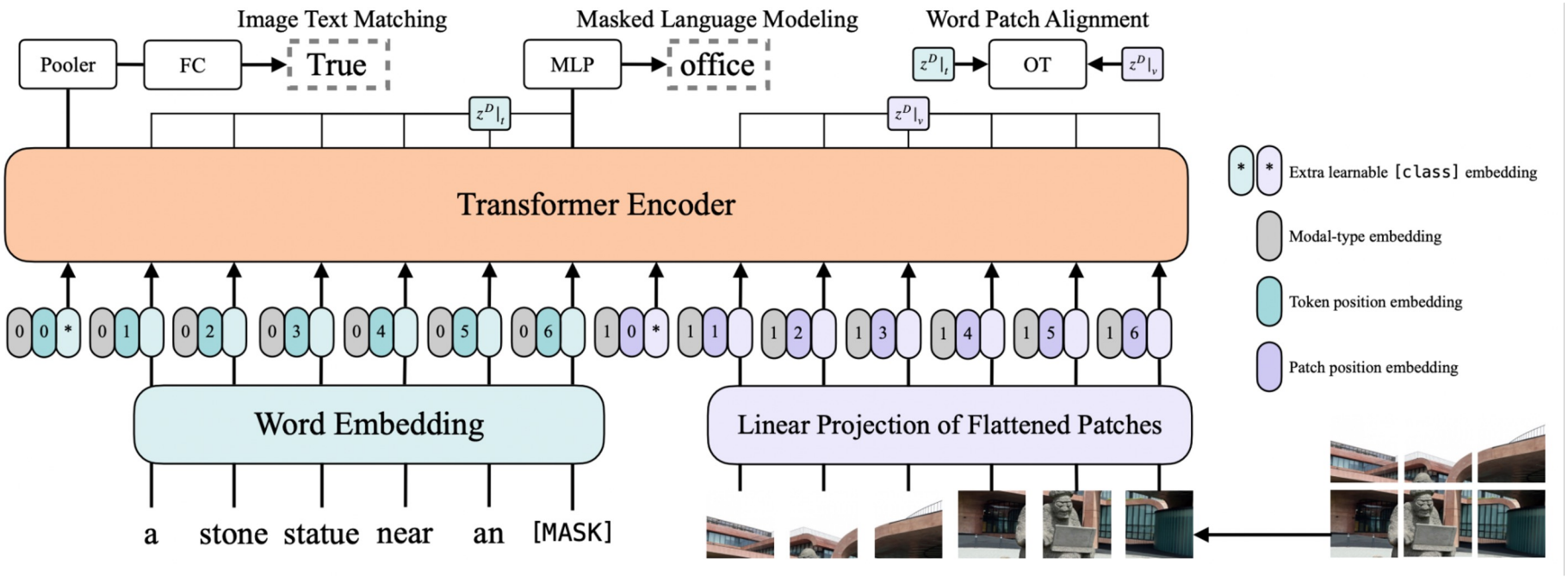
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

Transformers for Vision: Patches



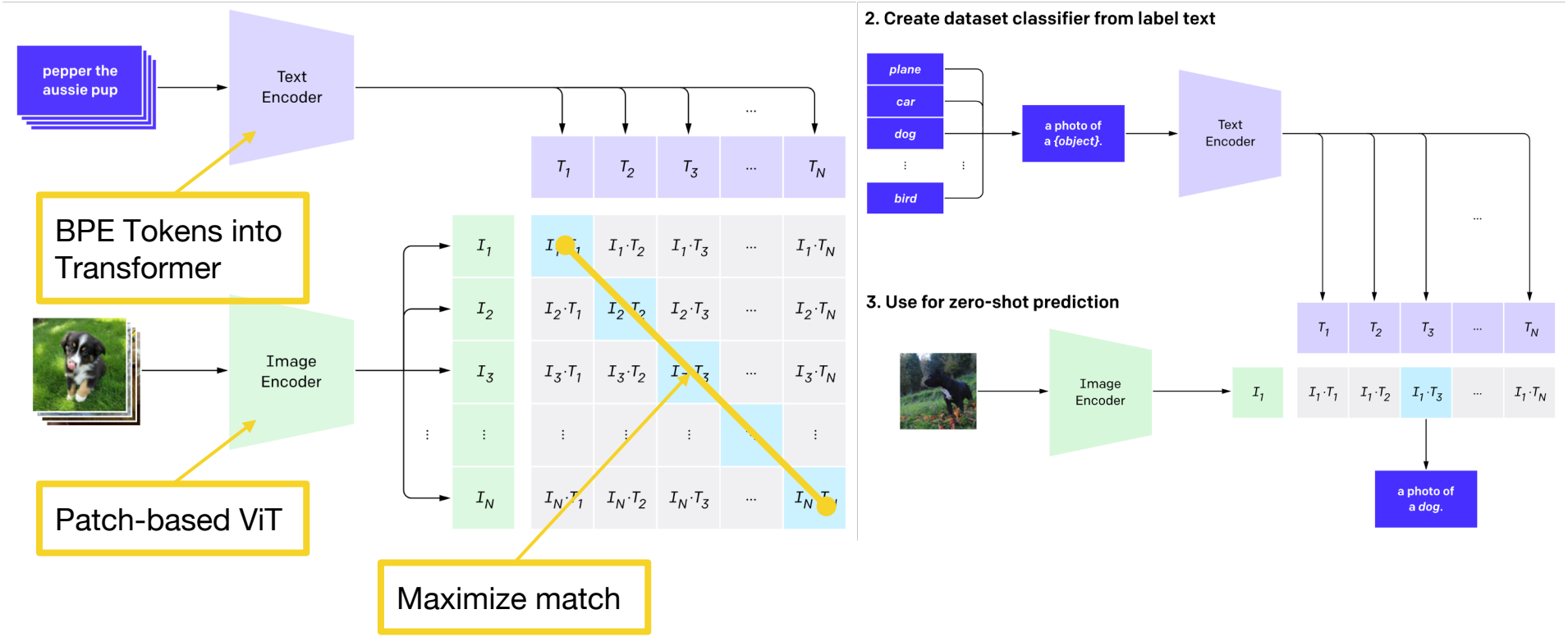
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Language and Vision: Patches



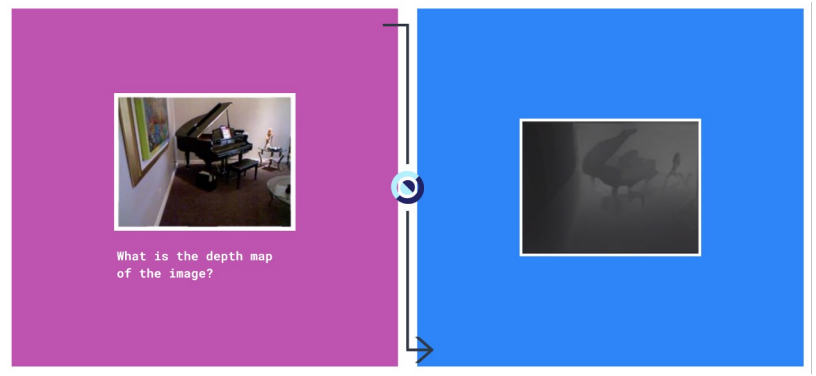
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision

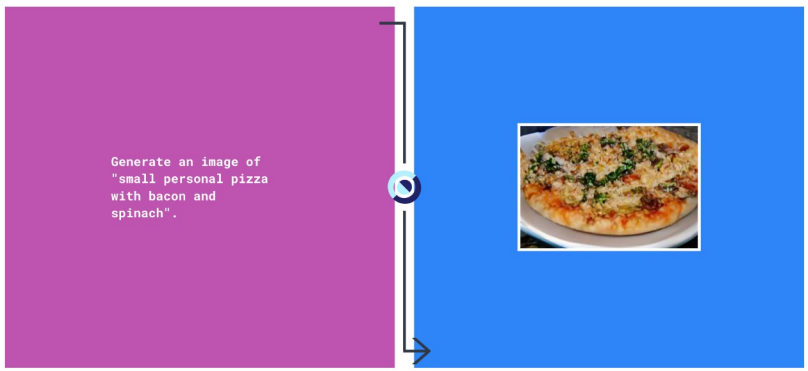
Contrastive Language–Image Pre-training (CLIP)

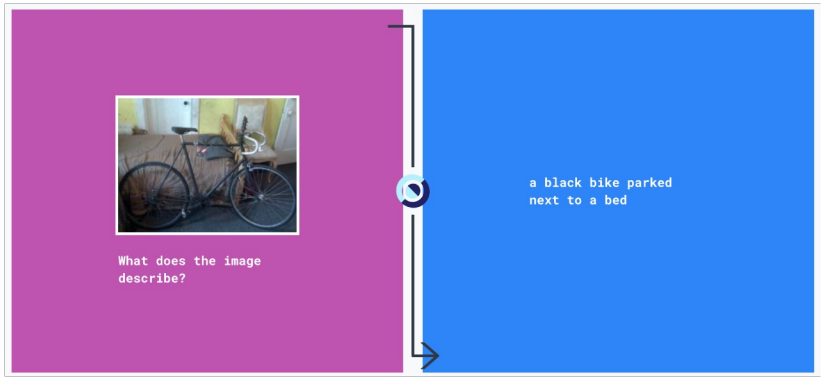


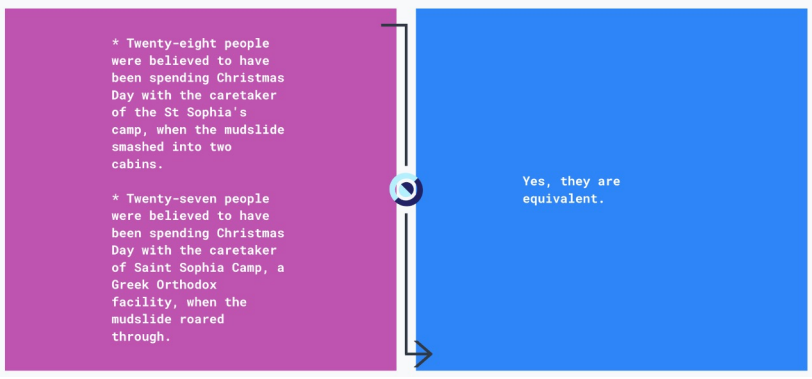
https://openai.com/blog/clip/

Multi-task Multi-modal Model: Unified IO









https://unified-io.allenai.org/

Multi-task Multi-modal Model: GPT-4

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA

- Algorithm details?
- No one will ever know.
- Since CLIP, OpenAI has committed to making deeply scientifically uninteresting AI artifacts.

https://openai.com/research/gpt-4

CSCI 566: Deep Learning and Its Applications

Jesse Thomason

Lecture 8: Multimodal Deep Learning