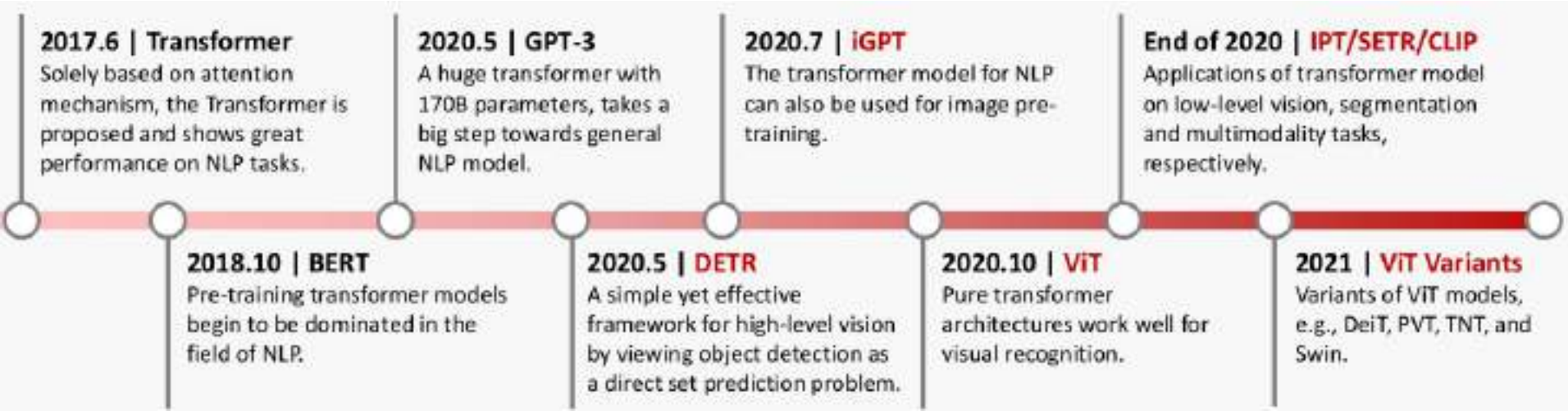
Computer Vision for Embedded Systems

Yung-Hsiang Lu
Purdue University
yunglu@purdue.edu

PURDUE UNIVERSITY

Transformers

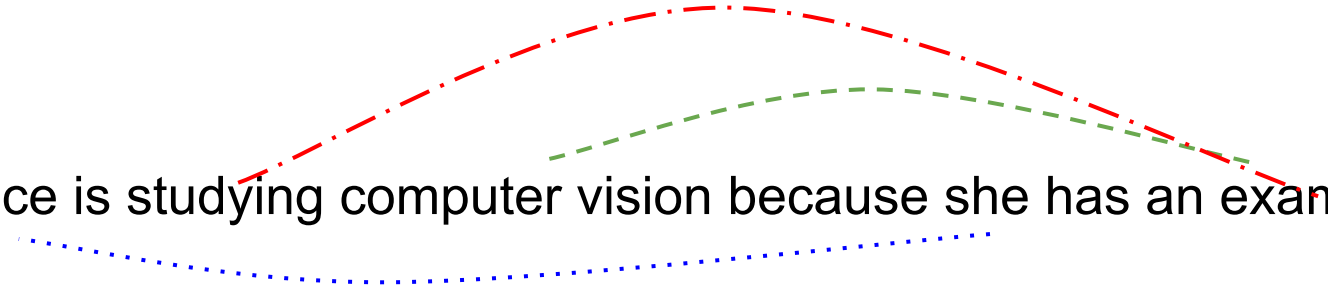


BERT: Bidirectional Encoder Representations from Transformers
GPT: Generative Pre-trained Transformer
DETR: End-to-End Object Detection with Transformers
iGPT: Image Generative Pre-trained Transformer
ViT: Vision Transformer

K. Han et al., "A Survey on Vision Transformer," in IEEE Transactions on Pattern Analysis and Machine Intelligence 2022 , doi: 10.1109/TPAMI.2022.3152247

Problems of CNN (Convolutional Neural Networks)

- CNN is a feed-forward structure, no feedback



Alice is studying computer vision because she has an exam tomorrow.

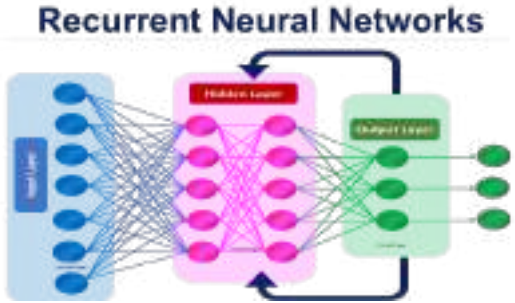
Alice is studying computer vision now because she has an exam tomorrow.

Bob put the book on the table. It is needed for tomorrow's exam.

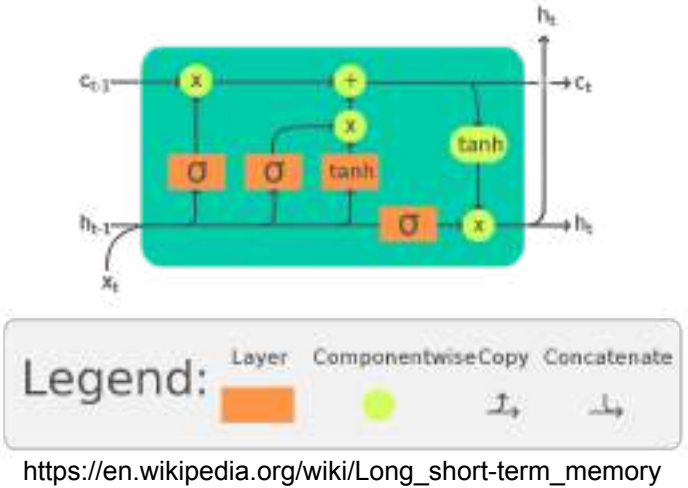
- Feedback (e.g., RNN and LSTM) is used for processing sequences of data (video and natural languages)

Problems of RNN (Recurrent Neural Networks)

- LSTM (long-short term memory) was popular for processing sequences of data but LSTM has several problems
  - difficult to capture relationships far away. Recurrent neural networks have difficulty handling long sentences.
  - vanishing gradient
  - sequential processing



https://medium.datadriveninvestor.com/recurrent-neural-network-58484977c445

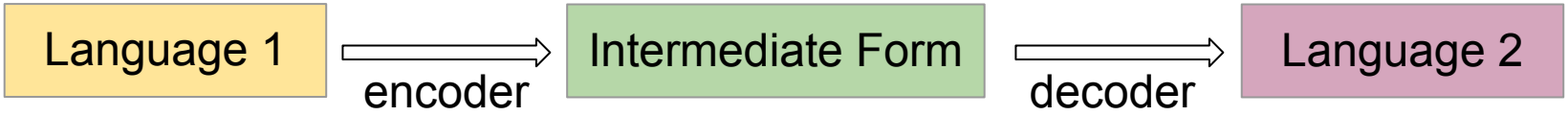


https://en.wikipedia.org/wiki/Long_short-term_memory

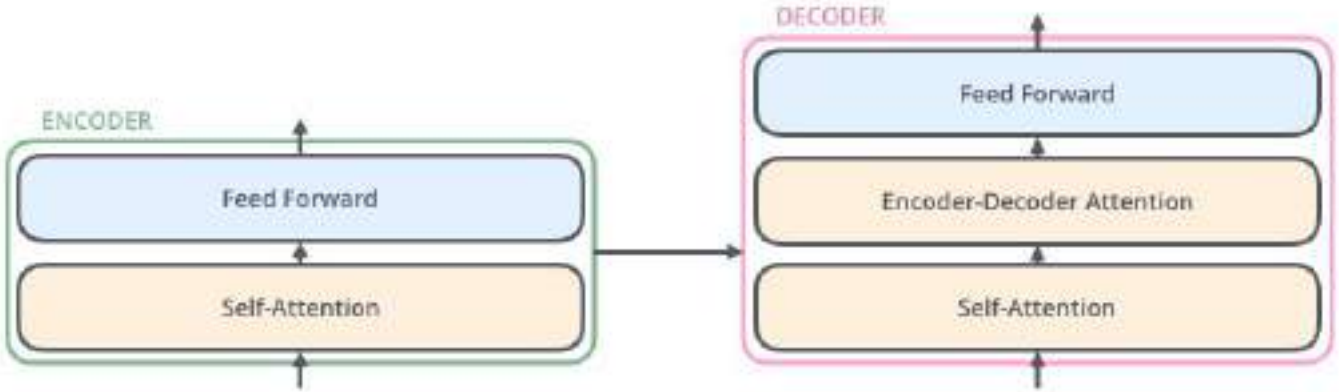
Encoder + Decoder + Intermediate Form

Widely used languages: English, Mandarin, Hindi, Spanish, French, Arabic

N languages: translation A ⇒ B: N^2 possibilities (actually N (N-1))

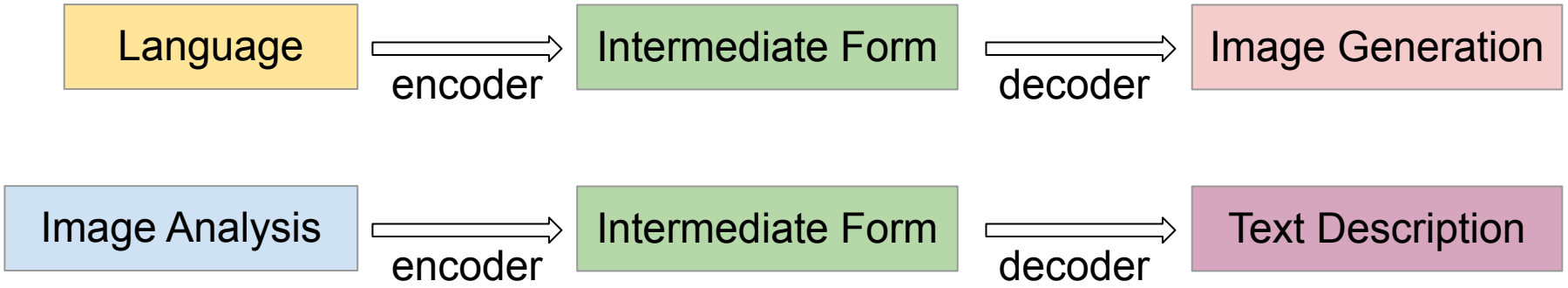


A ⇒ IF: N possibilities, IF ⇒ B: N possibilities, total 2N possibilities



Encoder + Decoder = Variety of Tasks

The encoder / decoder structure can solve many tasks

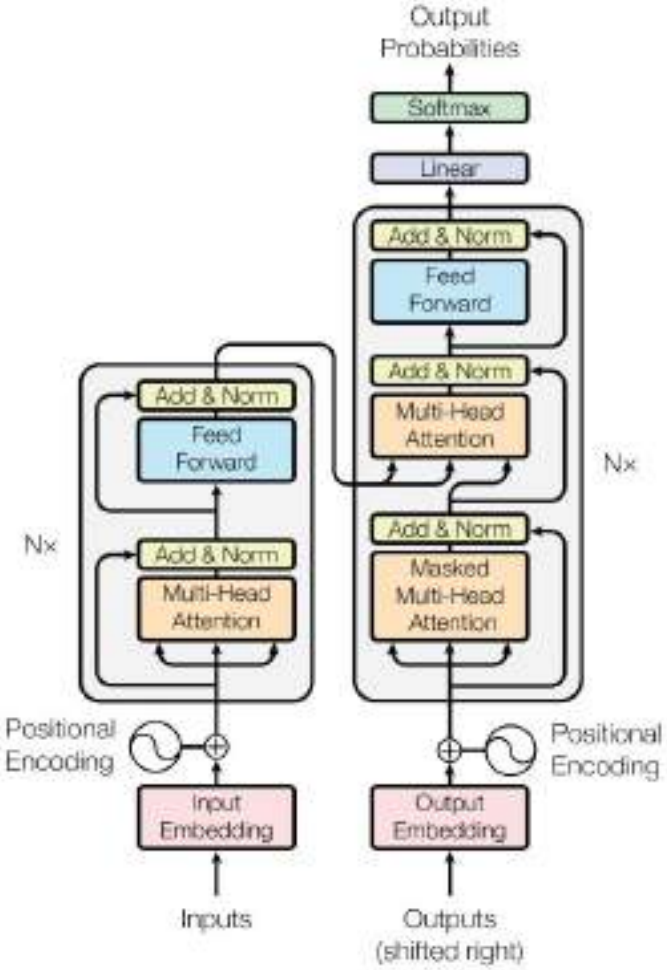


Transformers for Language Processing

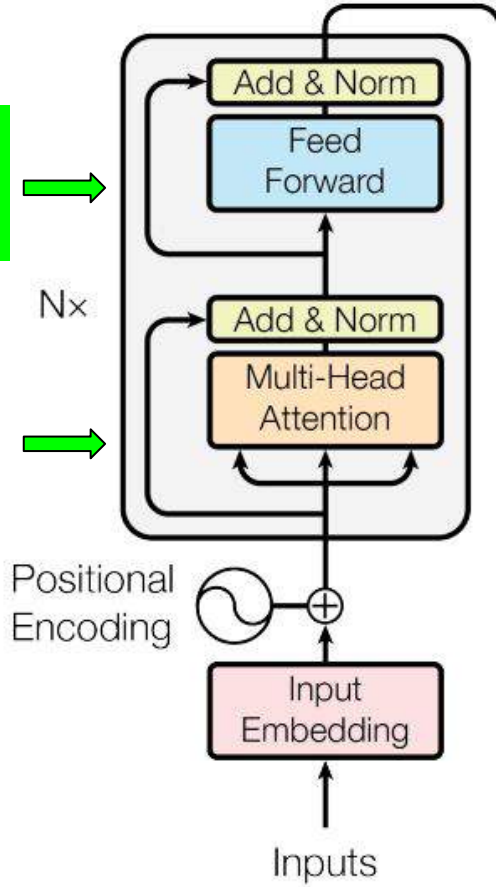
- Positions are important information
  - Alice is happy because Bob is not coming.
  - Bob is happy because Alice is not coming.
- To keep position information, a sequence of data must be processed sequentially. Is that true? Is it possible to process the words in parallel?

Attention Is All You Need, 2017 Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia,

Acknowledgments: Math of Intelligence, Rasa, and Halfling Wizard on Youtube
PyTorch Implementation: http://nlp.seas.harvard.edu/annotated-transformer/


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

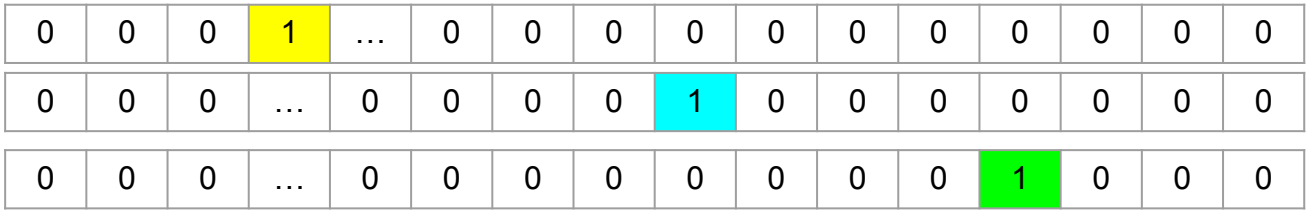
Residual connection
i.e., skip layers



Input Embedding

represent each word by a vector (e.g., I am learning computer vision.)

- option 1: long vectors, with a specific position for each word
  - I
  - am
  - learning



⇒ very long vectors, inefficient, unscalable

- option 2: one number for each word
  - I: 2549; am: 367; learning: 16579

⇒ implicitly create ordering

Input Embedding

- Each word is represented by a 512-element vector
  - I
  - am
  - learning

0.5	0.07	-0.8	0.14	…	-0.67	0.09	-0.01	0.74
-0.67	0.08	0.01	…	-0.4	0.02	0.1	0.23	0.05
0.2	-0.3	0.2	-0.5	0.07	0.6	-0.8	0.09	-0.5

- The values are obtained by training. Synonyms and related words are closer in the high-dimensional space. Antonyms are farther away from each other (should have opposite directions). The relative directions of two vectors can be determined by the inner product.

Position Encoding

I am not surprised that Cathy has been to Chicago.

I am surprised that Cathy has not been to Chicago.

- Recurrent neural networks process the input sequential to keep positions
- Transformers encode position information and allow parallel processing
- Option 1: use the position indexes (I am learning computer vision)
  - I: 0
  - am: 1
  - learning: 2

2	2	2	…	2	2	2	2	2
0.2	-0.3	0.2	-0.5	+0.07	0.6	-0.8	0.09	-0.5

⇒ destroy the words' meanings; later positions have larger values (why?)

Position Encoding

- Option 2: use the position indexes over length $\frac{\text{position}}{\text{length}}$

  I am learning computer vision (5 words)

  - I: 0
  - am: 0.2
  - learning: 0.4

  ⇒ sensitive to the sentence's length.

  I am learning computer vision **now**
  I am learning computer vision **right now**
  I am learning computer vision **at this moment**

Position Encoding (proposed in this paper)

$$\sin\left(\frac{\text{position}}{10000^{\frac{2i}{d}}}\right) \qquad \cos\left(\frac{\text{position}}{10000^{\frac{2i}{d}}}\right)$$

d model size (512 in this paper)

i index

- When i increases, 10000^{2i} increases, the frequency decreases
- The values are always between -1 and 1 ⇒ positions will not dominate the embeddings

Query, Key, Value

Q = computer vision for embedded systems

	K	**V**
Course	Topics	Website
A	machine learning, artificial intelligence, back propagation, neural networks	WA
B	Python, PyTorch, TensorFlow	WB
C	computer vision, quantization, privacy, data bias, transformer	WC
D	parallel programming, GPU, Cuda	WD
E	natural language processing, transformer, translation	WE

$$\mathbf{QK^T \rightarrow P}$$

Query, Key, Value

Q = computer vision for embedded systems

	K	**V**
Course	Topics	Website
A	machine learning, artificial intelligence, back propagation, neural networks	WA
B	Python, PyTorch, TensorFlow	WB
C	computer vision, quantization, privacy, data bias, transformer	**WC**
D	parallel programming, GPU, Cuda	WD
E	natural language processing, transformer, translation	WE

$$QK^T \rightarrow P$$

Matrices for Query, Key, Value

input embedding + positions = X (a vector for each word)

XM_Q = Q (query)

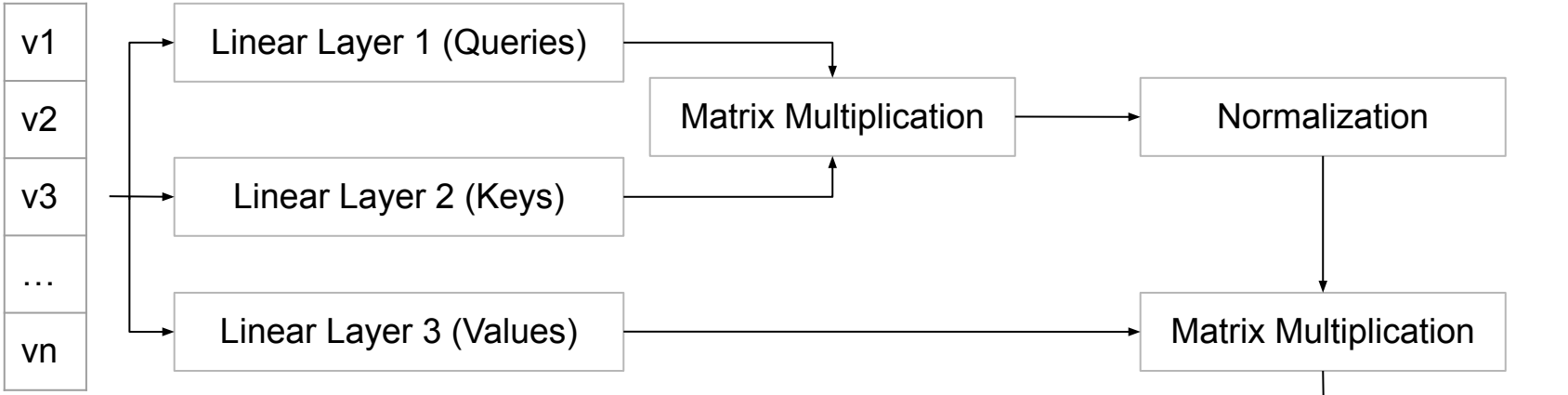
XM_K = K (key)

XM_V = V (value)

M_Q, M_K, M_V are matrices whose values can be tuned (i.e., trained).

They are as linear layers in neural networks (no activation functions).

Self Attention

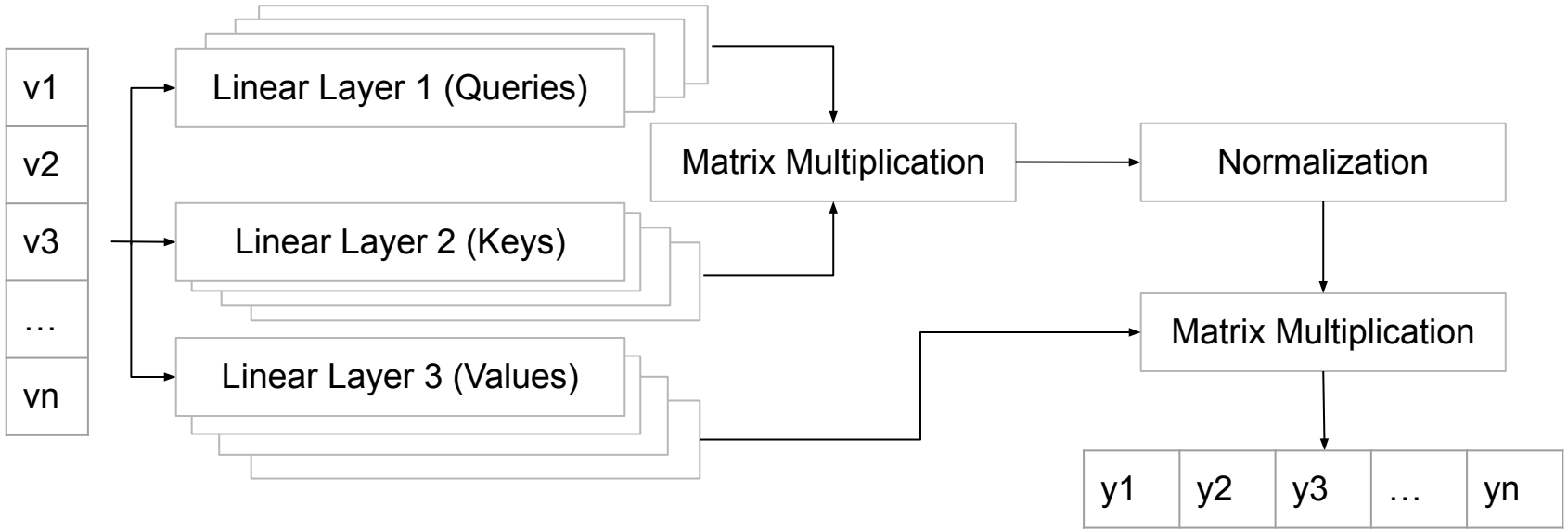


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

https://www.youtube.com/watch?v=tIvKXrEDMhk

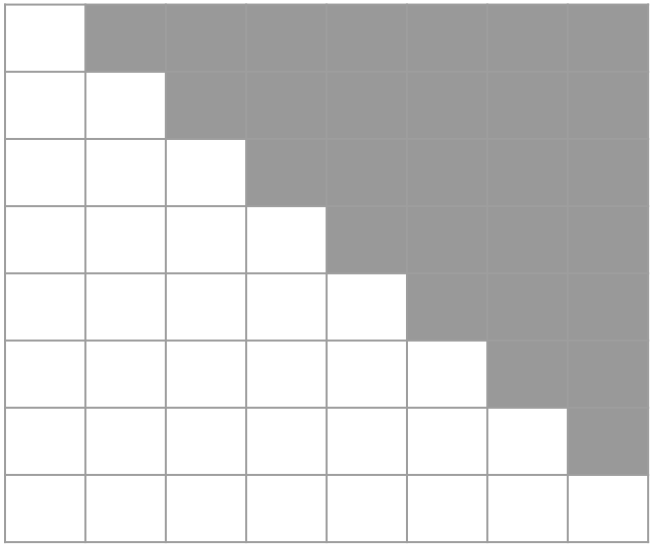
Multi-Head Attention

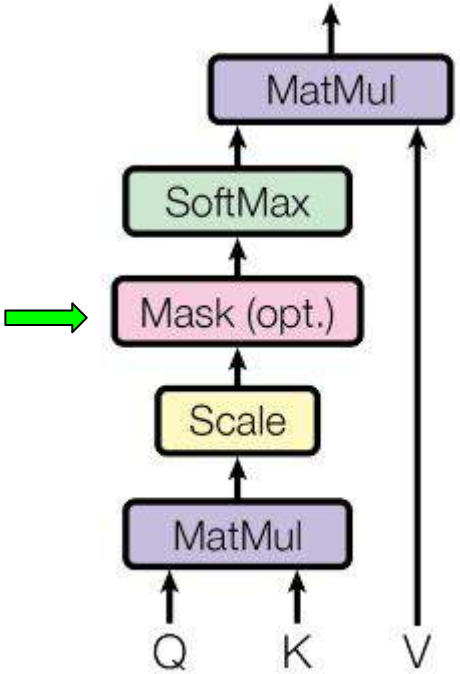
I am reading the book of Computer Vision written by the professor.

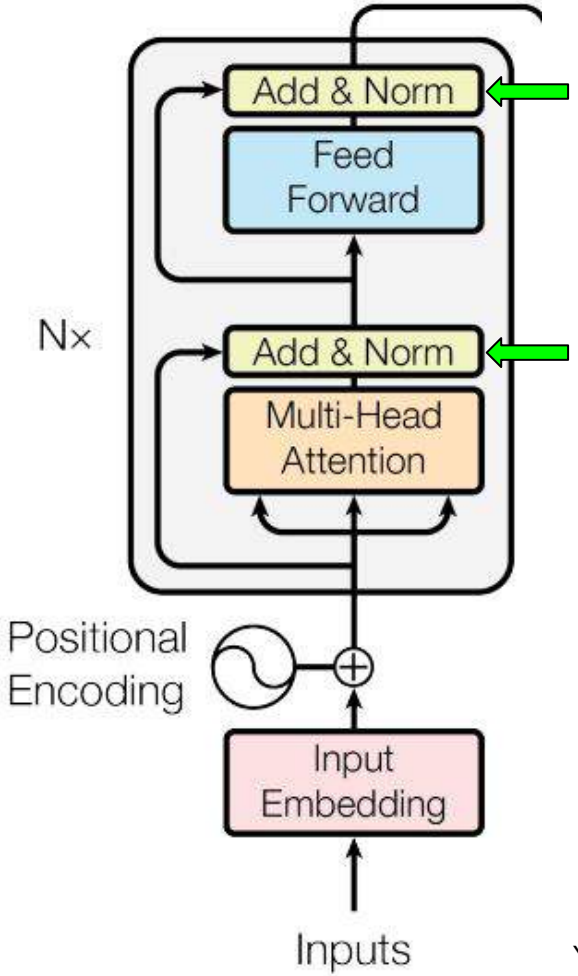


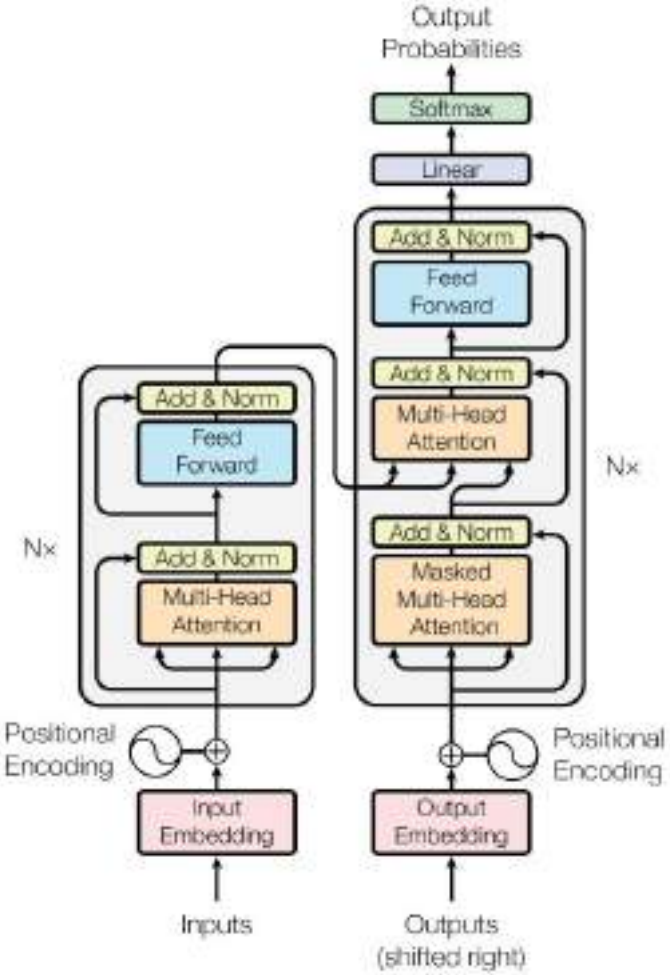
Masked Attention

The input sequence cannot see future inputs






Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

BLEU Score

BiLingual Evaluation Understudy

BLEU : a Method for Automatic Evaluation of Machine Translation, Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, Annual Meeting on Association for Computational Linguistics 2002

Quantify the Quality of Machine Translation (MT)

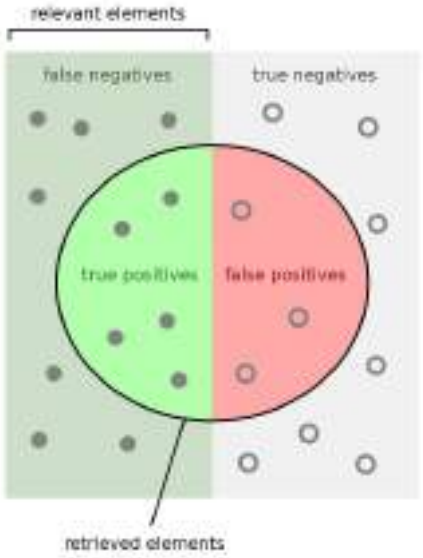
The same concept may be expressed differently, e.g.,

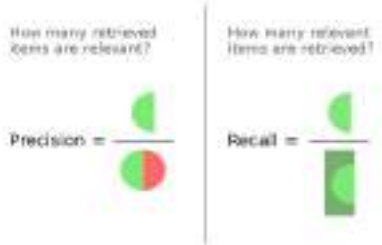
- I am studying computer vision.
- I am learning computer vision.
- Computer vision is the subject I am studying now.
- Today I spend time learning computer vision.

Solution: "match" machine translations with human written translations

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

percentage of the MT words appear in the references?





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

How is BLEU score calculated?

Precision: percentage of words MT also appear in the references?

Reference: The cat is on the mat.
MT1: The cat is mat the on.
Precision MT1 = 1 (every word in MT appears in the reference)
MT2: the the the the the the the
Precision MT2 = 2/7

n-gram: consider n words together as a unit

The cat is on the mat: "The cat", "cat is", "is on", "on the", "the mat"

"The cat", "cat is", "is mat", "mat the", "the on".

https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1

GLUE

General Language Understanding Evaluation

GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding
Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman 2019

Natural Language Understanding (NLU) Tasks

Nine NLU tasks divided into three categories: (1) Single-Sentence Tasks, (2) Similarity and Paraphrase Tasks, (3) Inference Tasks.

Single-Sentence Tasks:

1. Corpus of Linguistic Acceptability: decide sequences of words are grammatical English sentences
2. Stanford Sentiment Treebank: decide the sentiments of sentences from movie reviews as positive or negative

Natural Language Understanding (NLU) Tasks

Similarity and Paraphrase Tasks

1. Microsoft Research Paraphrase Corpus: decide whether sentence pairs from online news are semantically equivalent
2. Quora Question Pairs: decide whether pairs of questions are semantically equivalent
3. Semantic Textual Similarity Benchmark: assign scores of similarities of sentence pairs from news

Natural Language Understanding (NLU) Tasks

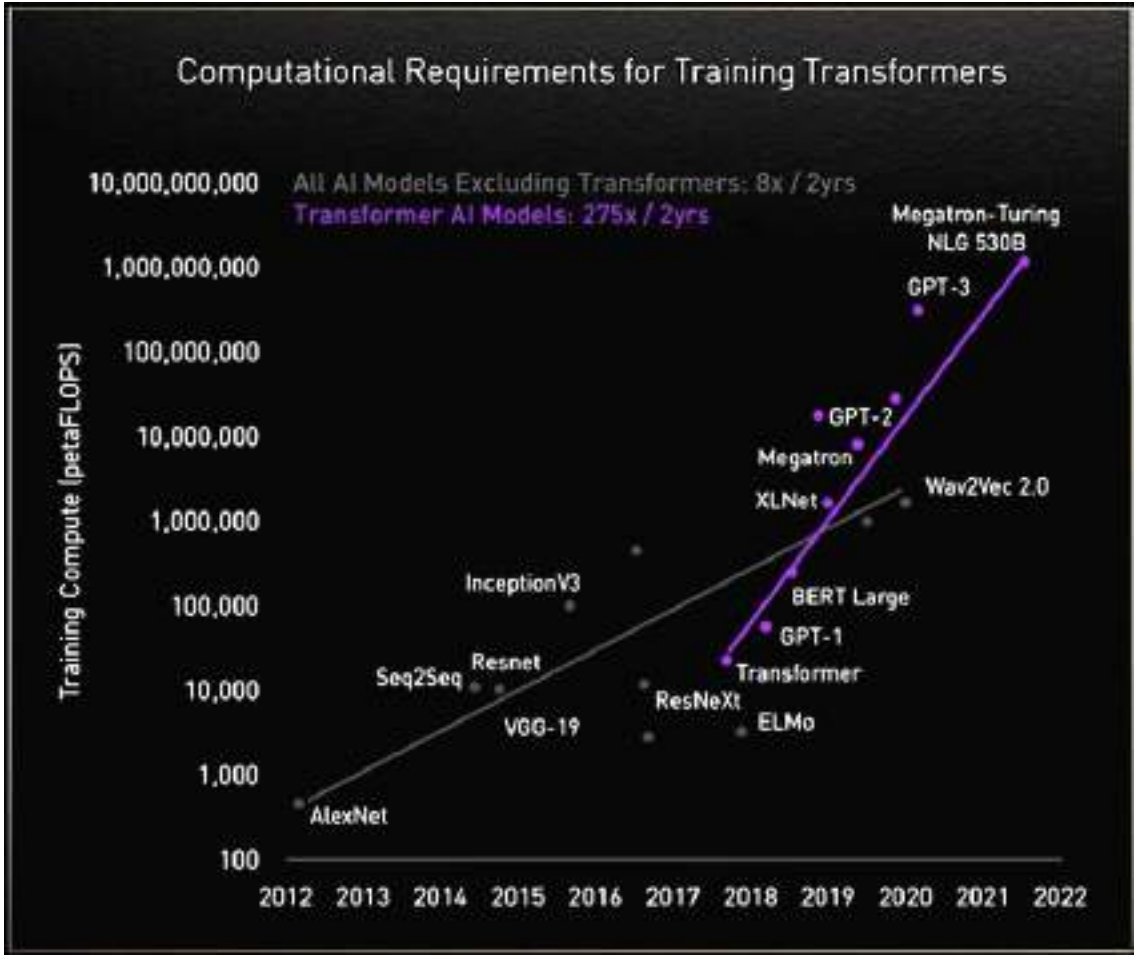
Inference Tasks

1. Multi-Genre Natural Language Inference Corpus: decide whether one sentence entails or contradicts another sentence or neither
2. Stanford Question Answering Dataset: decide whether a sentence in a paragraph (from Wikipedia) answers a question
3. Recognizing Textual Entailment
4. Winograd Schema Challenge: select a noun referred by a pronoun in a sentence

Tags	Sentence 1	Sentence 2	Fwd	Bwd
Lexical Entailment (Lexical Semantics), Downward Monotone (Logic)	The timing of the meeting has not been set, according to a Starbucks spokesperson.	The timing of the meeting has not been considered, according to a Starbucks spokesperson.	N	E
Universal Quantifiers (Logic)	Our deepest sympathies are with all those affected by this accident.	Our deepest sympathies are with a victim who was affected by this accident.	E	N
Quantifiers (Lexical Semantics), Double Negation (Logic)	I have never seen a hummingbird not flying.	I have never seen a hummingbird.	N	E

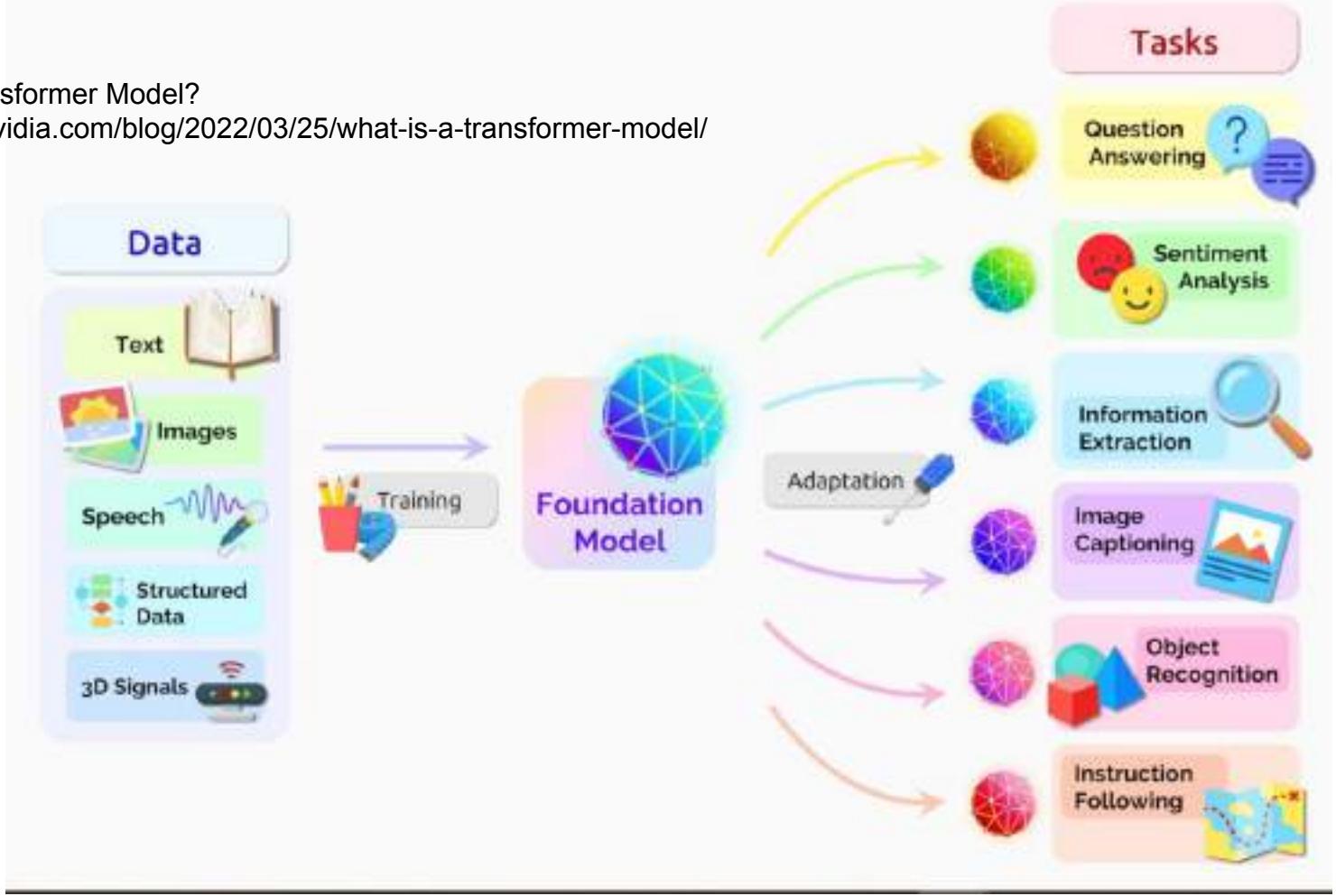
Table 3: Examples from the diagnostic set. *Fwd* (resp. *Bwd*) denotes the label when sentence 1 (resp. sentence 2) is the premise. Labels are *entailment* (E), *neutral* (N), or *contradiction* (C). Examples are tagged with the phenomena they demonstrate, and each phenomenon belongs to one of four broad categories (in parentheses).

Transformers need (much) more computation



https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

What Is a Transformer Model?
https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/



Attention

She poured water from the pitcher to the cup until it was full.

She poured water from the pitcher to the cup until it was empty.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova 2019

BERT

1. Pre-Training: unlabeled data
2. Fine-Tuning: labeled data

unified architecture for multiple language tasks: