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

ENSF 619.02 | Fall 2021

By: Shirin Yamani (yamani.shirin@ucalgary.ca), Mahsa Malek (mahsa.malek@ucalgary.ca)

Instructor: Dr. Roberto Medeiros de Souza

Course: Advanced topic in Machine Learning for Image Analysis.

Content

Transformer

1

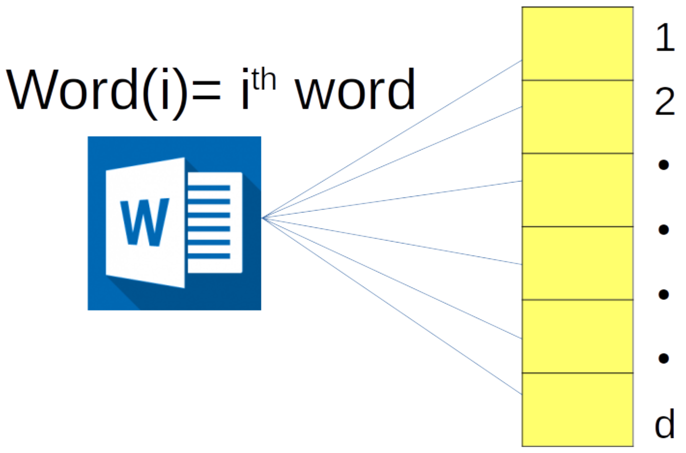
Notational Meaning of Transformer

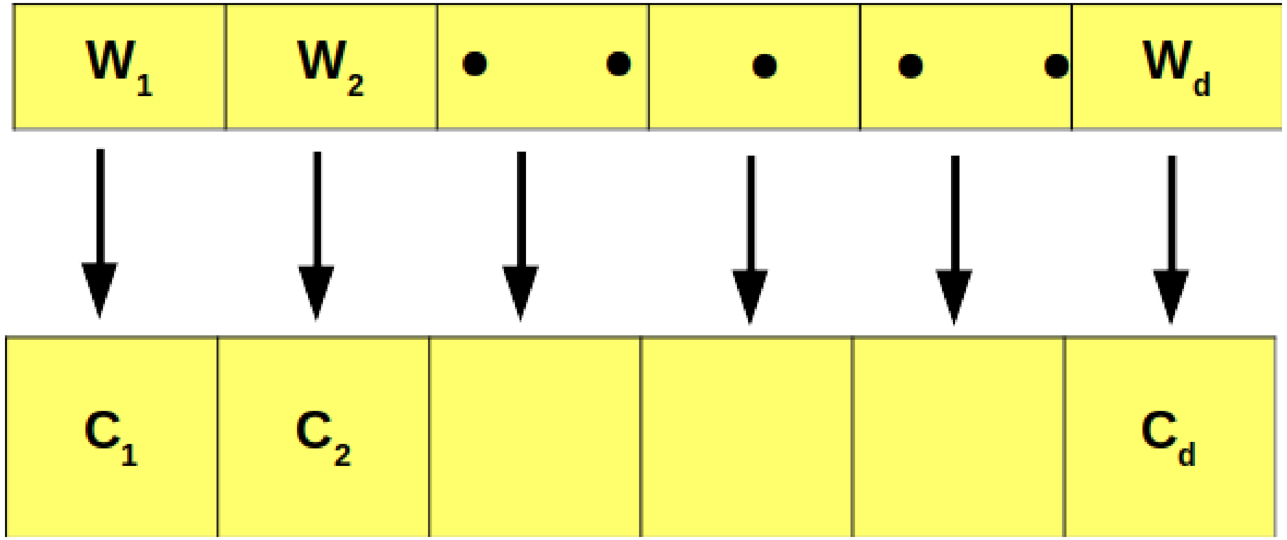
2

Vision Transformer

3

Introduction to Transformer

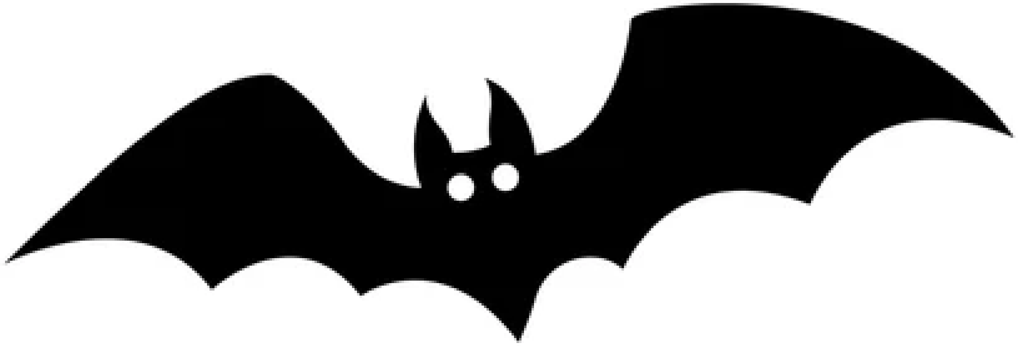




Word mapping

Transformer

Bat ?!



Accounting for the Word Context

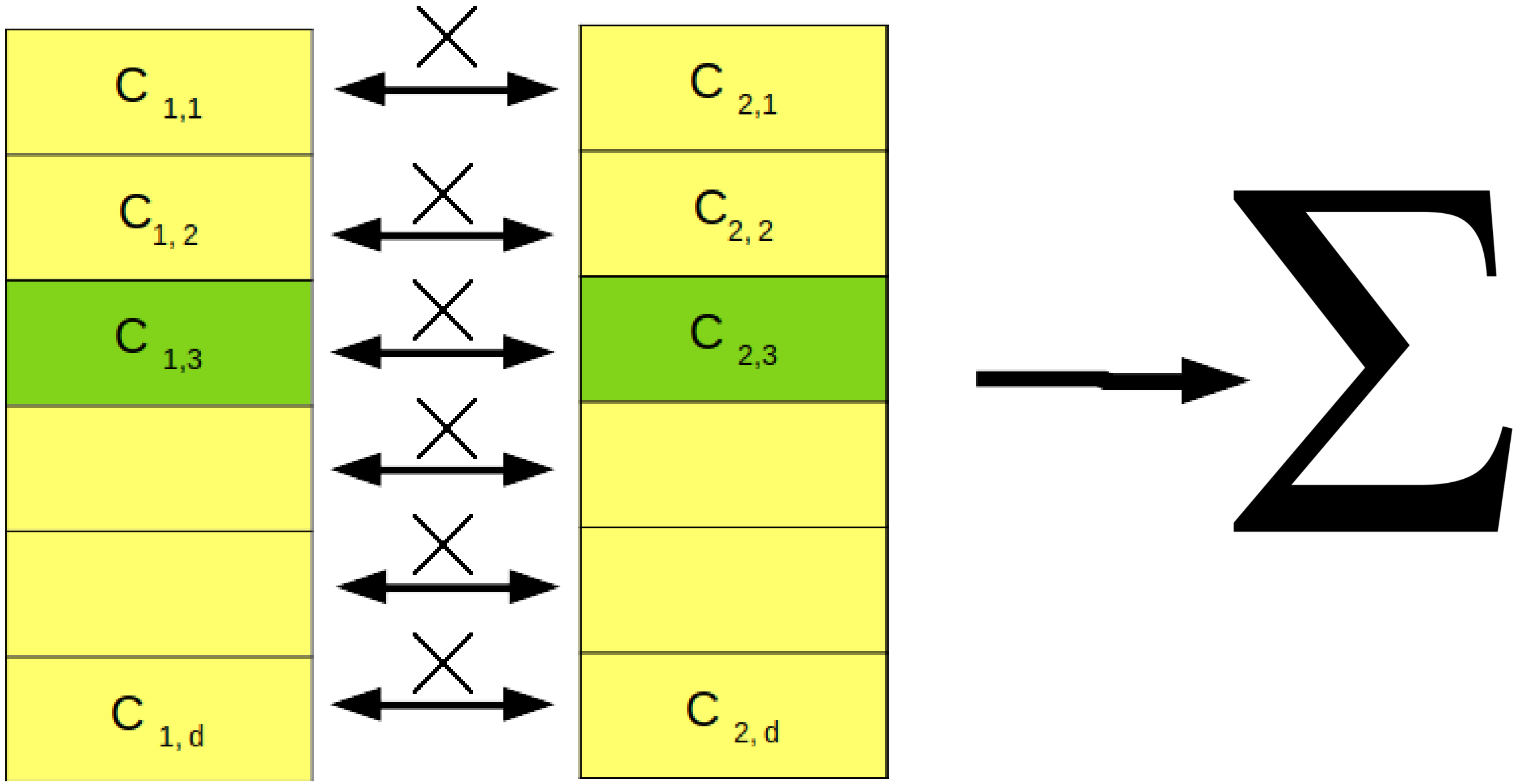
- Mapping each word to a single vector is restrictive!
- Words often have different meanings!

Transformer

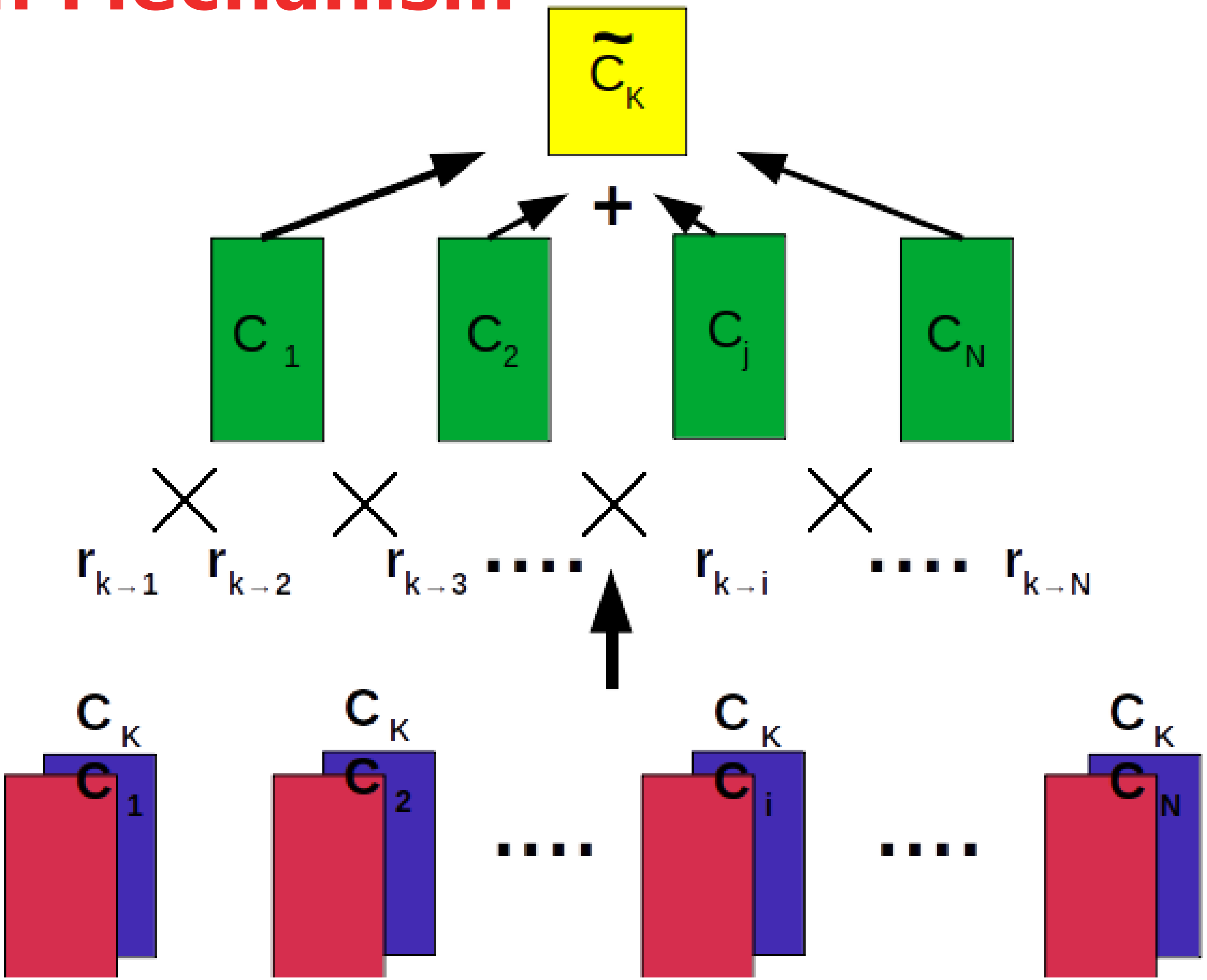
Inner Product between two vectors!

Goal!
Quantify how similar kth word is to the sequence!

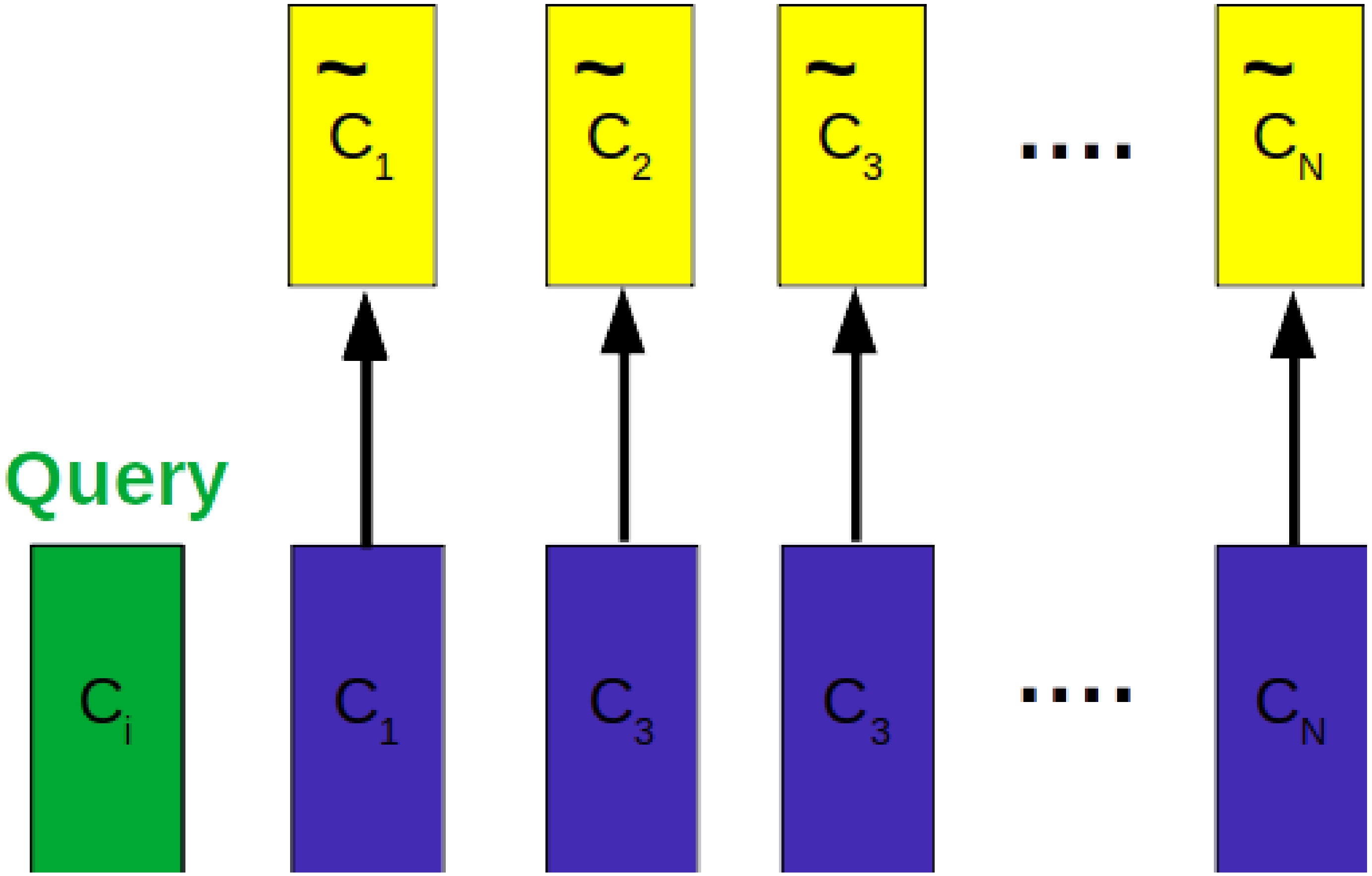
- **If inner product is positive, then words are similar**
- **If inner product is negative, then words are dissimilar**



Attention Mechanism



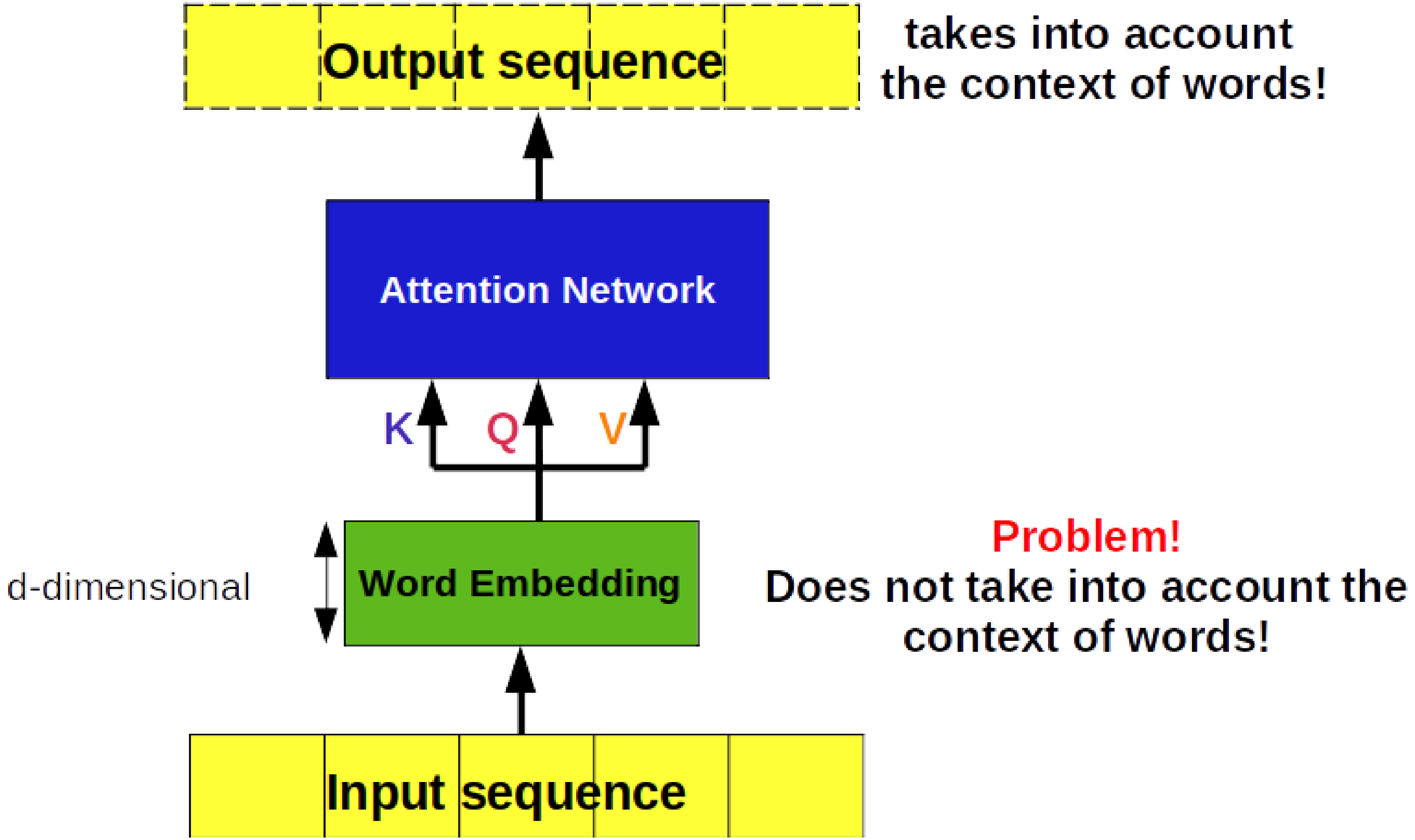
Transformer



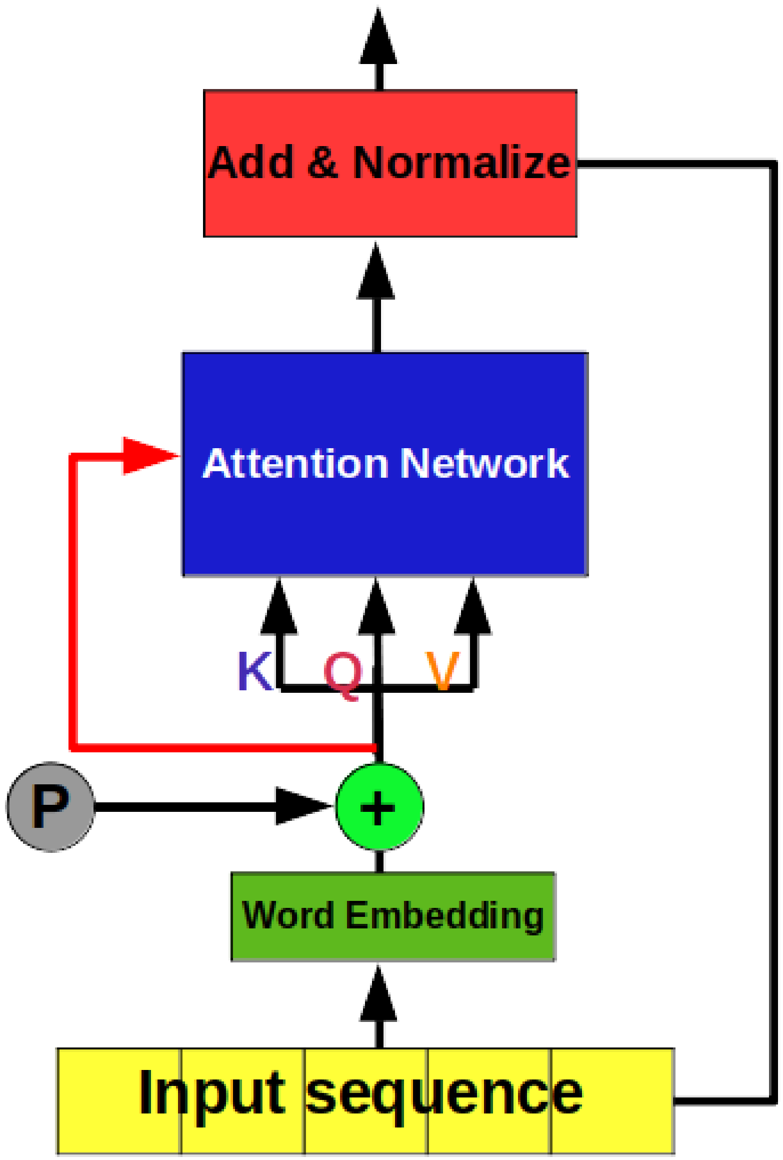
Attention

- Original sequence of words
- Map to corresponding code(Problem: original code independent of surrounding words!)
- Updated codes take into account the context of surrounding words!

Notional meaning of Transformer



Transformer Encoder



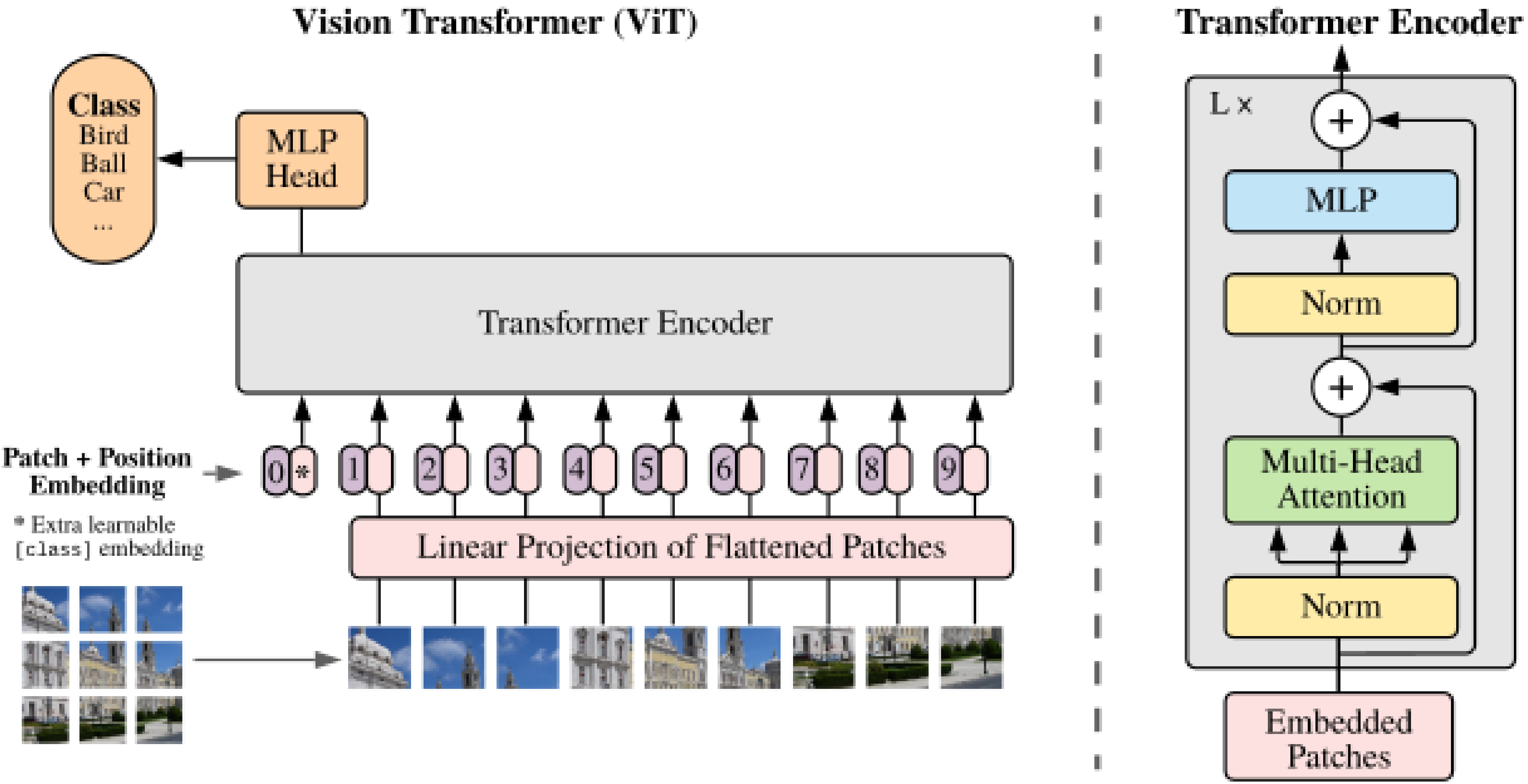
Positional Embedding

Accounts for the Order

1. Vaswani et al., 2017

Vision Transformer (ViT)

UNIVERSITY OF CALGARY



Vision Transformer (ViT) Method:
1) ViT Components

- Close to original Transformer (Vaswani et al., 2017).
- Transformer receives as input 2D images.
- Reshape the image into 2 sequence of flattened 2D patches with same resolution of the original image.
- ViT uses constant latent vector size D through all of its layers.
- Position embeddings are added to the patch embeddings to retain positional information.

Vision Transformer (ViT)
Method:
2) Fine-tuning and higher resolution

- Typically, pre-train ViT on large datasets, and fine-tune to (smaller) downstream tasks.
  - Remove the pre-trained prediction head and attach a zero-initialized $D \times K$ feedforward layer, where K is the number of downstream classes.
  - Fine-tune at a higher resolution than pre-training -> keep the patch size the same, which results in a larger effective sequence length.

2. Dosovitskiy, Alexey, et al. 2021

Vision Transformer (ViT)
Experiment:

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Details of Vision Transformer model variants

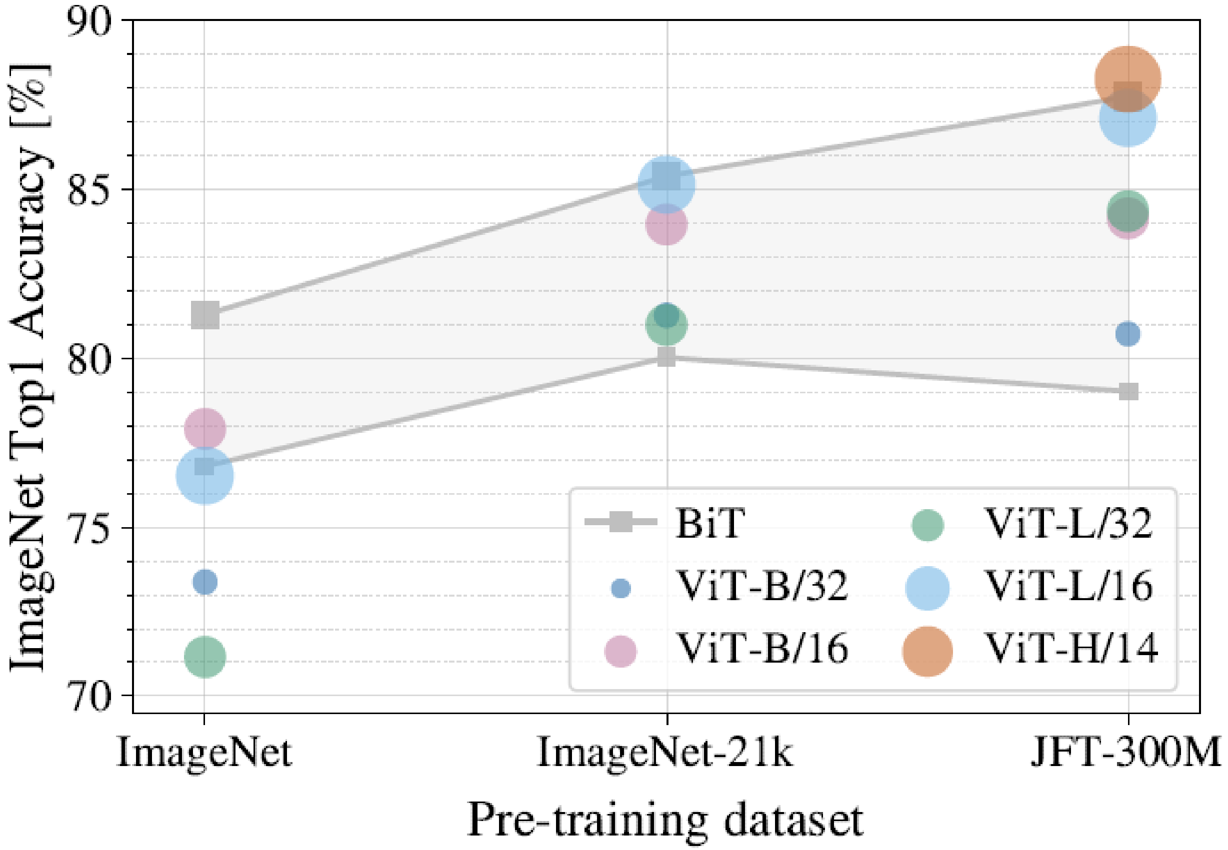
2. Dosovitskiy, Alexey, et al. 2021

Vision Transformer (ViT)
Experiment:

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

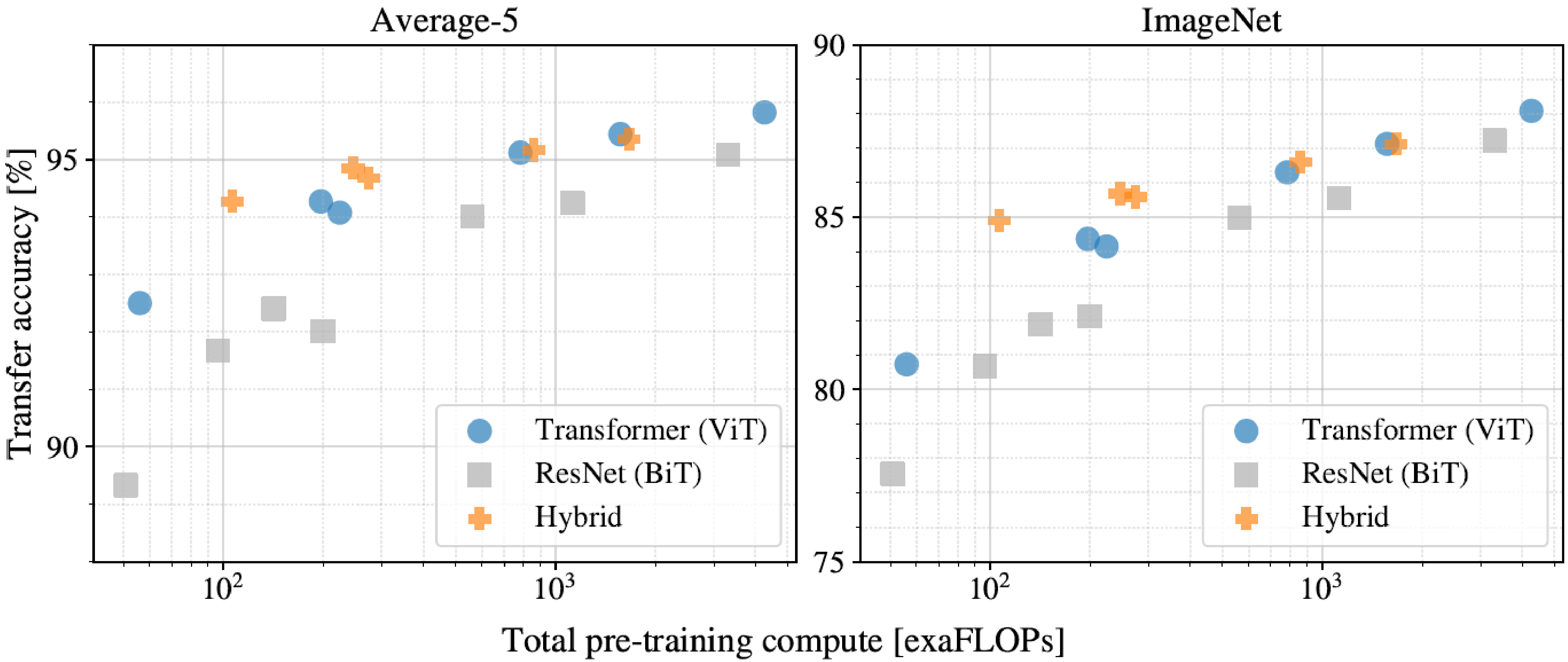
Comparison with state of the art on popular image classification benchmarks

Vision Transformer (ViT)
Experiment:



Transfer to ImageNet

Vision Transformer (ViT)
Experiment:



Performance vs pre-training compute for different architectures

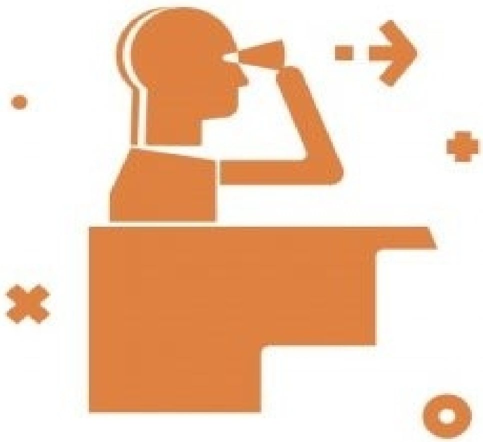
Vision Transformer (ViT) Conclusion

- ViT interprets an image as a sequence of patches and process it by a standard Transformer encoder as used in NLP.

- cheap to pre-train.

Vision Transformer (ViT)
Future work

- Apply ViT to other computer vision tasks.
- Exploring self-supervised pre-training methods.
- Scaling of ViT



2. Dosovitskiy, Alexey, et al. 2021

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017.
2. Dosovitskiy, Alexey, et al. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.” *ArXiv:2010.11929 [Cs]*, June 2021.