TRANSFORMERS FOR VISION

Gertjan Burghouts

Deep Learning SOTA sessions

Credits (Thank You!): Wouter, Arthur, Fieke, Raimon, David, Ombretta, Frank (LwLL student)

What's in it for *me*?

- Transformers are **fun** stuff! enjoy this technical ride
- This new architecture may play a **huge** role in deep learning (DL) for Vision.
- Cross-overs between Language and Vision,
  - relevant for our work on image + text (e.g., internet images, intelligence).
  - fostering collaboration with NLP folks (e.g., TNO Data Science).
- Many ideas that are applicable to other Vision tasks,
  - **attention** (e.g., focus on details, visual feedback).
  - positional encoding (e.g., relations between objects).
  - sequential analysis (e.g., evolving situations).
- **New forms** of learnable Computer Vision become possible!
  - e.g., interpret situations by objects in context.

What’s in it for *us*?

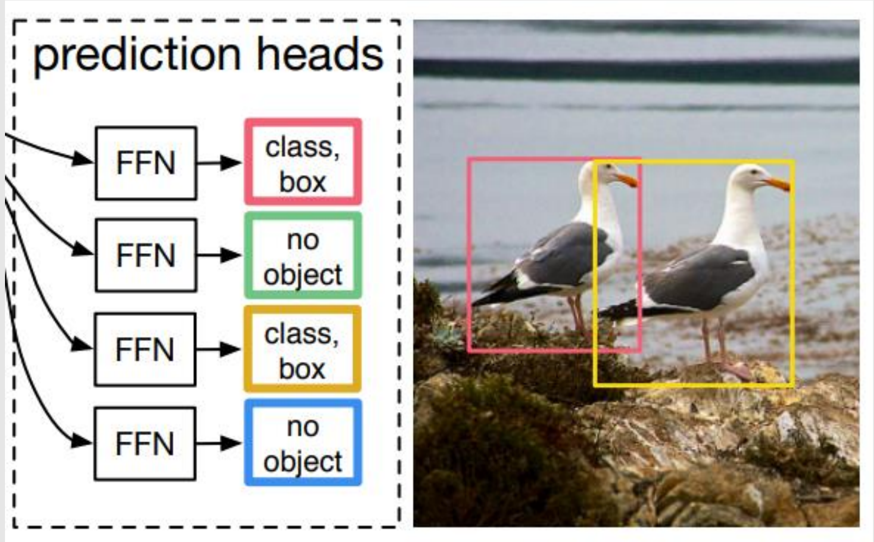
planting a seed for good afterthoughts and **new ideas** or already during & after this presentation!

Scope

Image classification

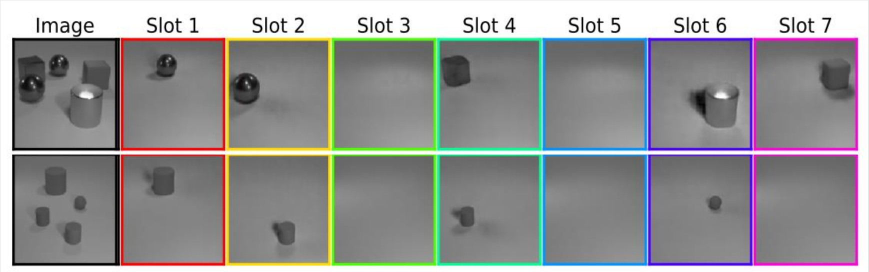
(Google AI)

Object detection



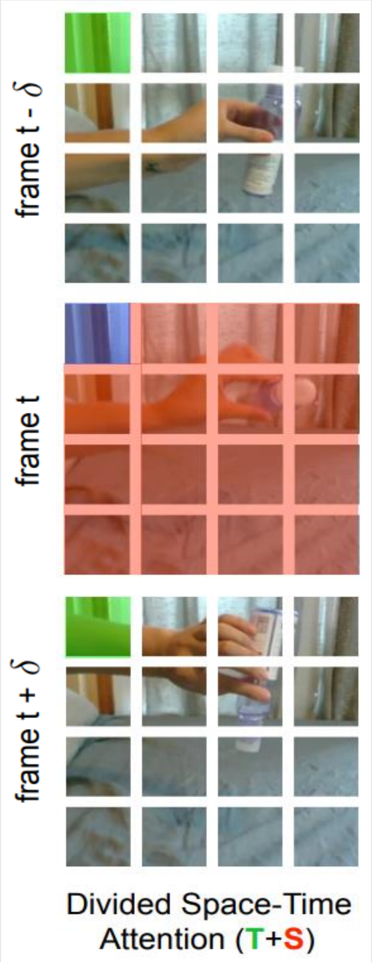
(Facebook AI)

Better generalization to new objects



(Google Brain A'dam)

Activity classification



(Facebook AI)

Today's ride

- Transformer
  - Model, Attention, Positional encoding, Training
- Vision
  - Image classification
  - Object detection
  - Few-shot generalization
  - Activity classification
- Summary & Discussion
- References
  - including further reading (advanced)

The focus will be more on the **ideas** and their potential **impact**.

Less on the implementation and results.

You can always check these yourself, via the references at each slide.

History of Transformers

"Attention is all you need" (Google)

2017

2018 - 2020

Introduction in various **Vision** Tasks
(scientific explorations, hybrid CNN-Transformers)

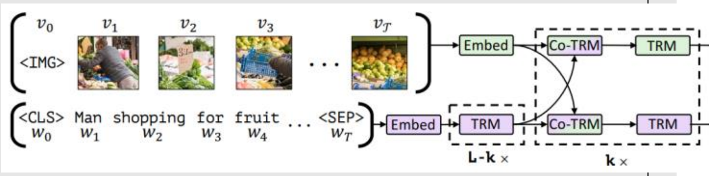
2021

Beating CNNs
on Large-scale Vision Tasks

Part I

Machine translation

Language & Vision



(highly relevant, also for II & DS,

but out-of-scope for today)

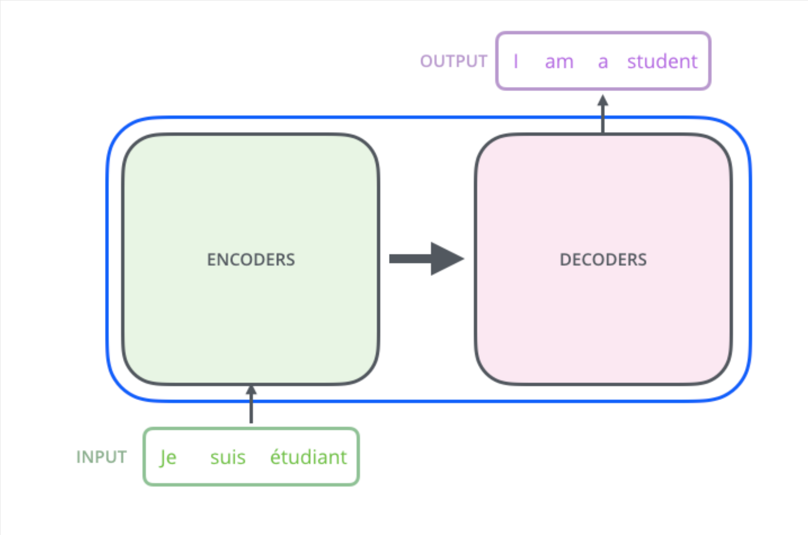
Part II

Computer Vision

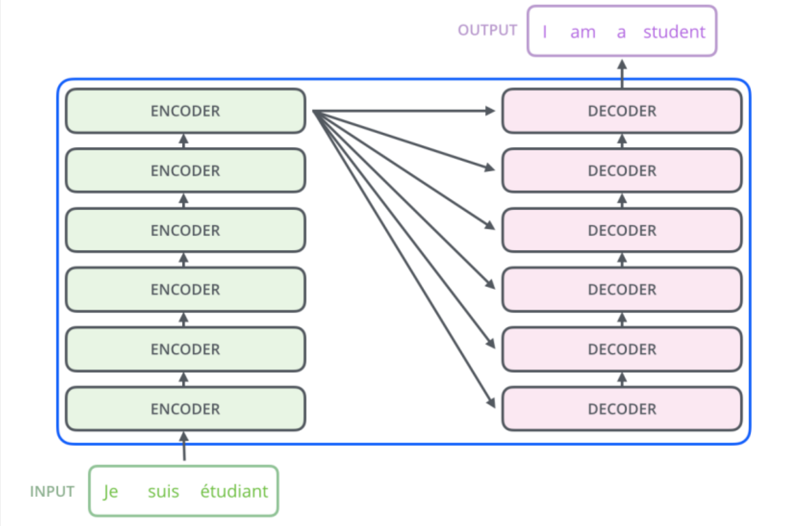
Architecture

- Natural Language Processing

model



stacked encoders/decoders

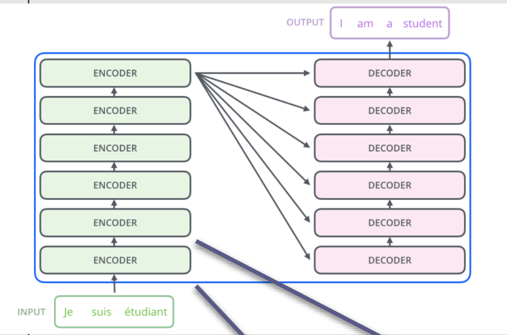


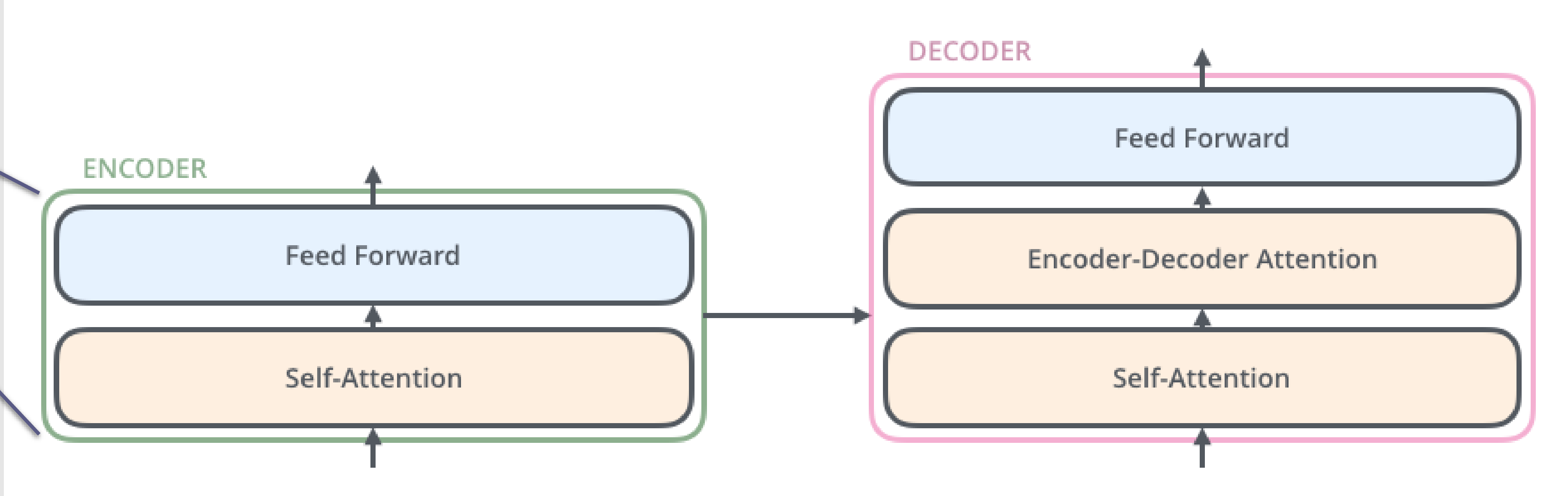
Machine translation

[The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)]

Encoder & Decoder

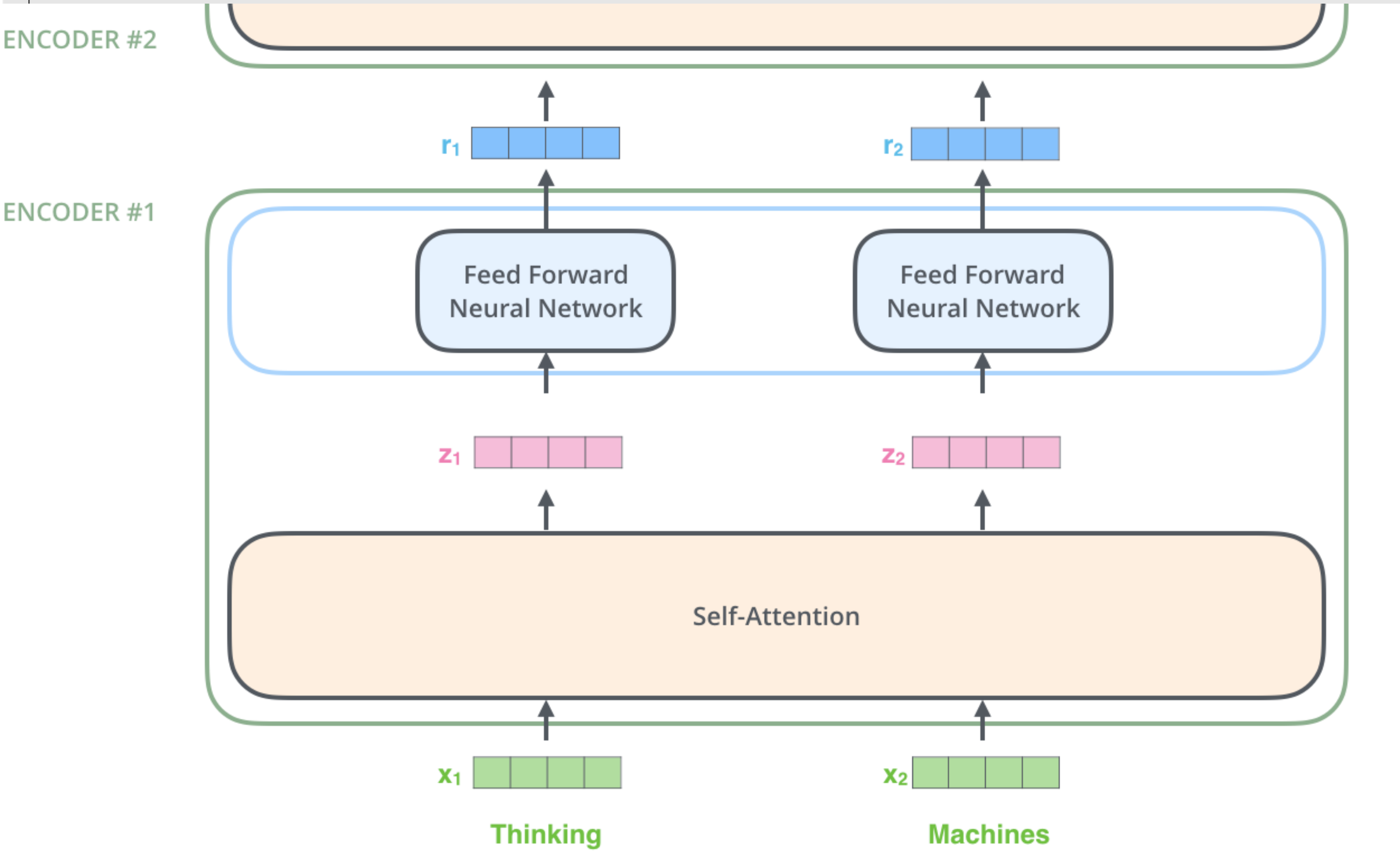
- each encoder/decoder in the stack has its own weights (no sharing)
- two main components:
  - Attention (complex)
  - Feed-forward (simple)





The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)

Architecture



intermediate representation

intermediate representation

attention for <u>one</u> word w.r.t. <u>all</u> other words

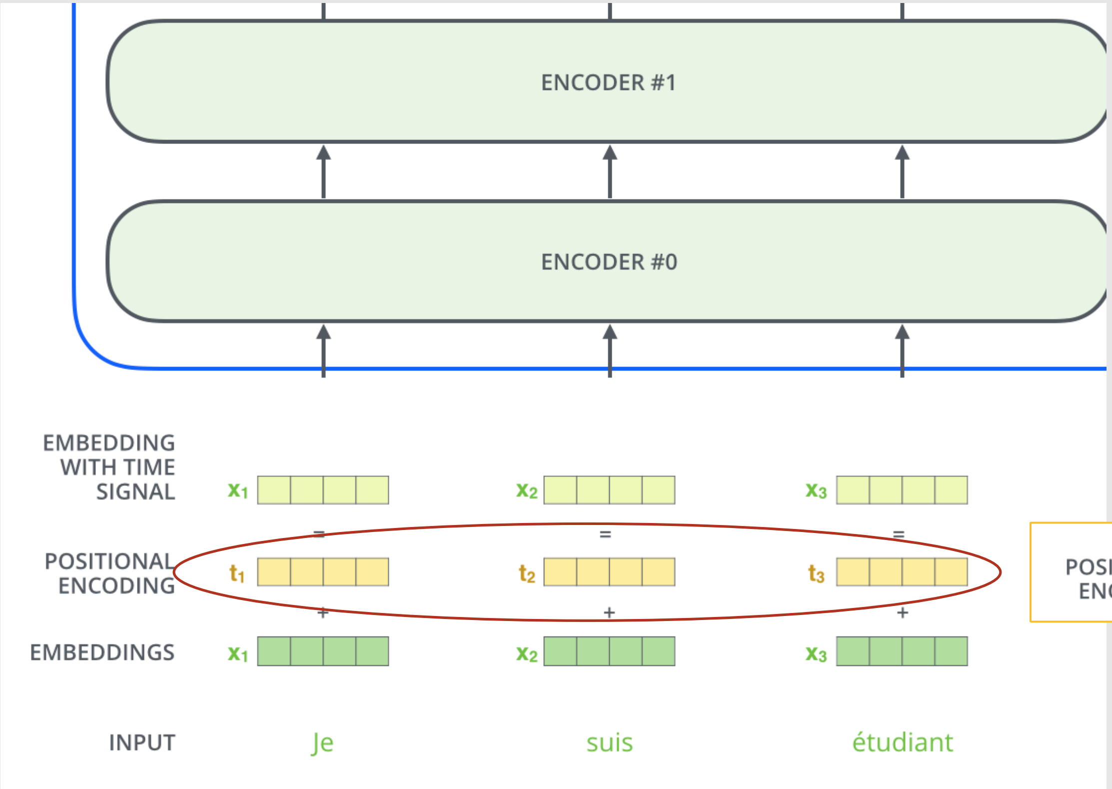
each word goes through its own path

at the input level, each word is encoded by a vector Xi of 512D by a label embedding (e.g., word2vec)

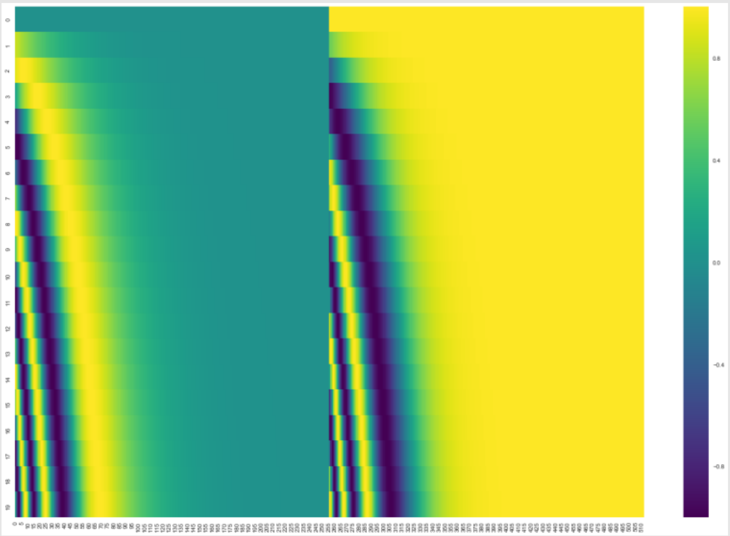
[The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)]

Positional encoding

- to have a clue about order and position of each word

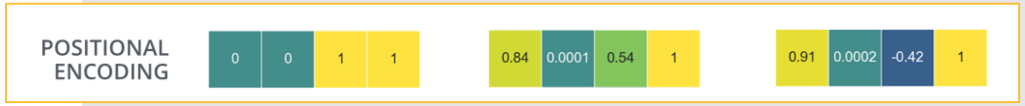


sentence length (20) ↑



→ word dimensionality (512D)

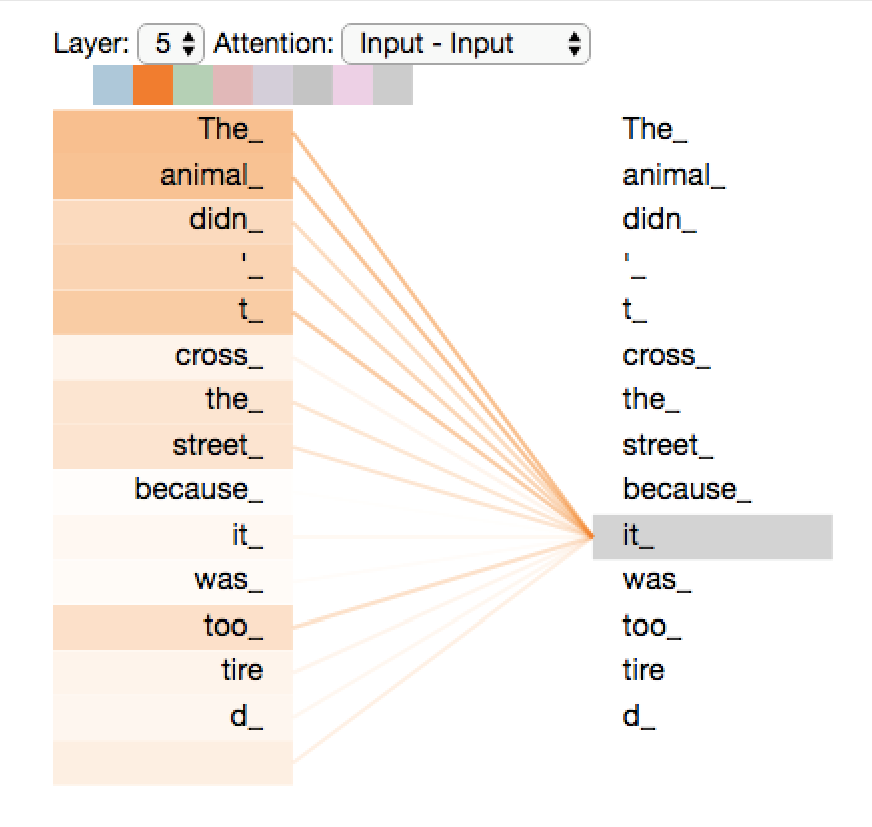
(example)



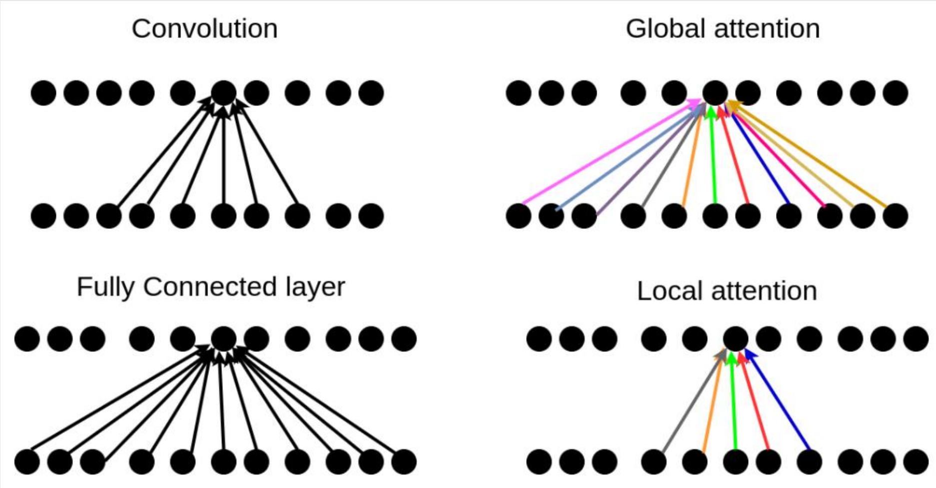
[The Illustrated Transformer](#)

Self-Attention

effect of other words on current word

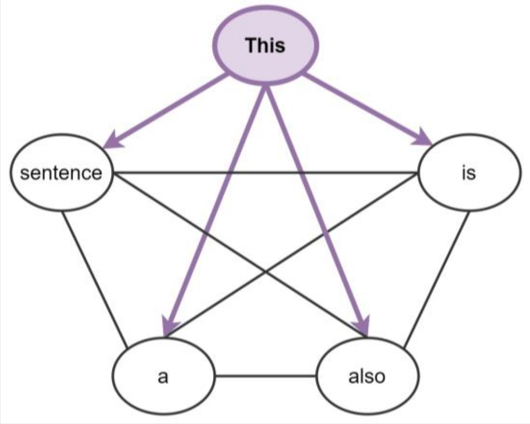


many variants of attention



the original Transformer uses Global Attention

Transformers are Graph Neural Networks



sentence = fully-connected graph of words

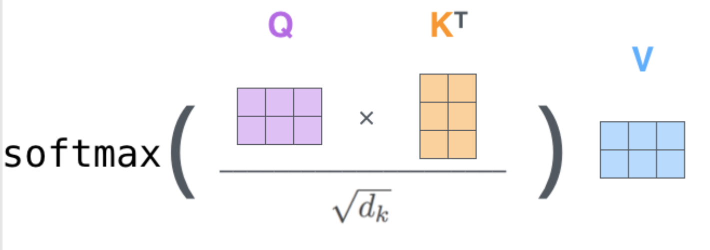
Graph Attention network (GAT)

Transformers are Graph Neural Networks | NTU Graph Deep Learning Lab   How Attention works in Deep Learning

Attention: Query, Key, Value

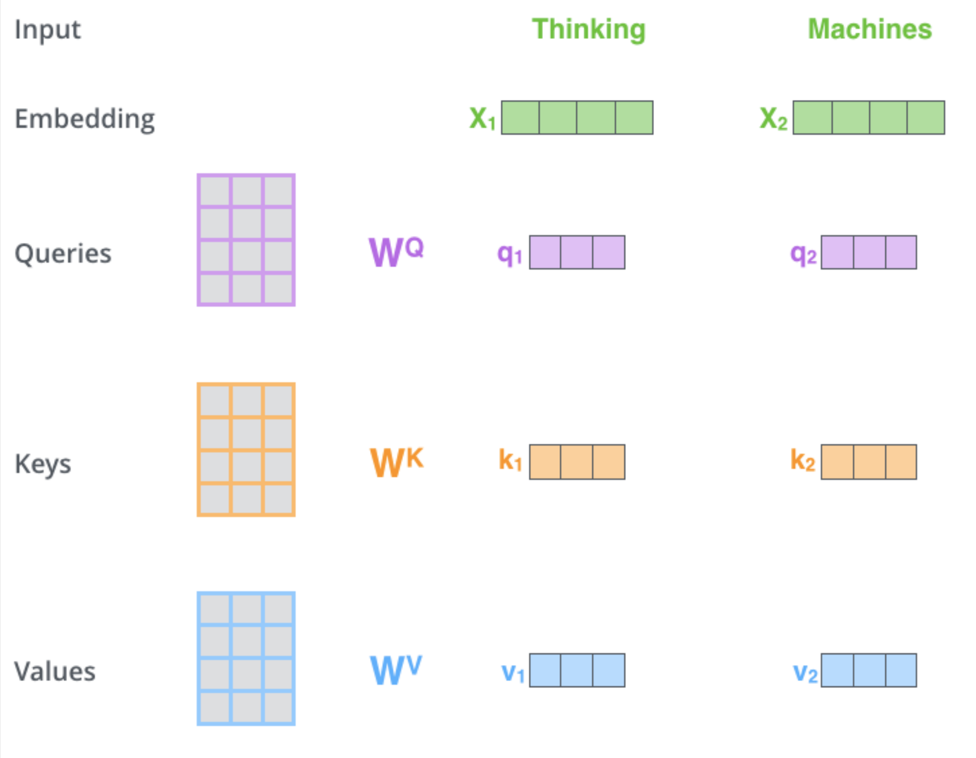
Game-changer proposed in Transformer:

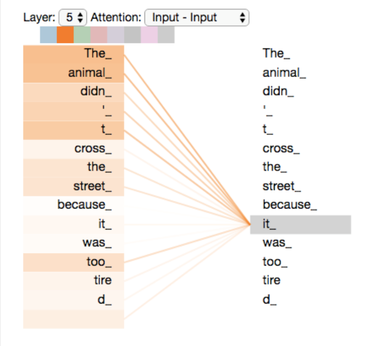
computation of Attention by Query, Key, Value



= equation from paper

Matrices Wq, Wk and Wv are learned during training

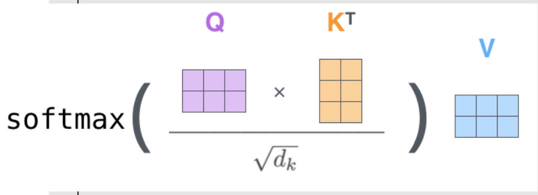


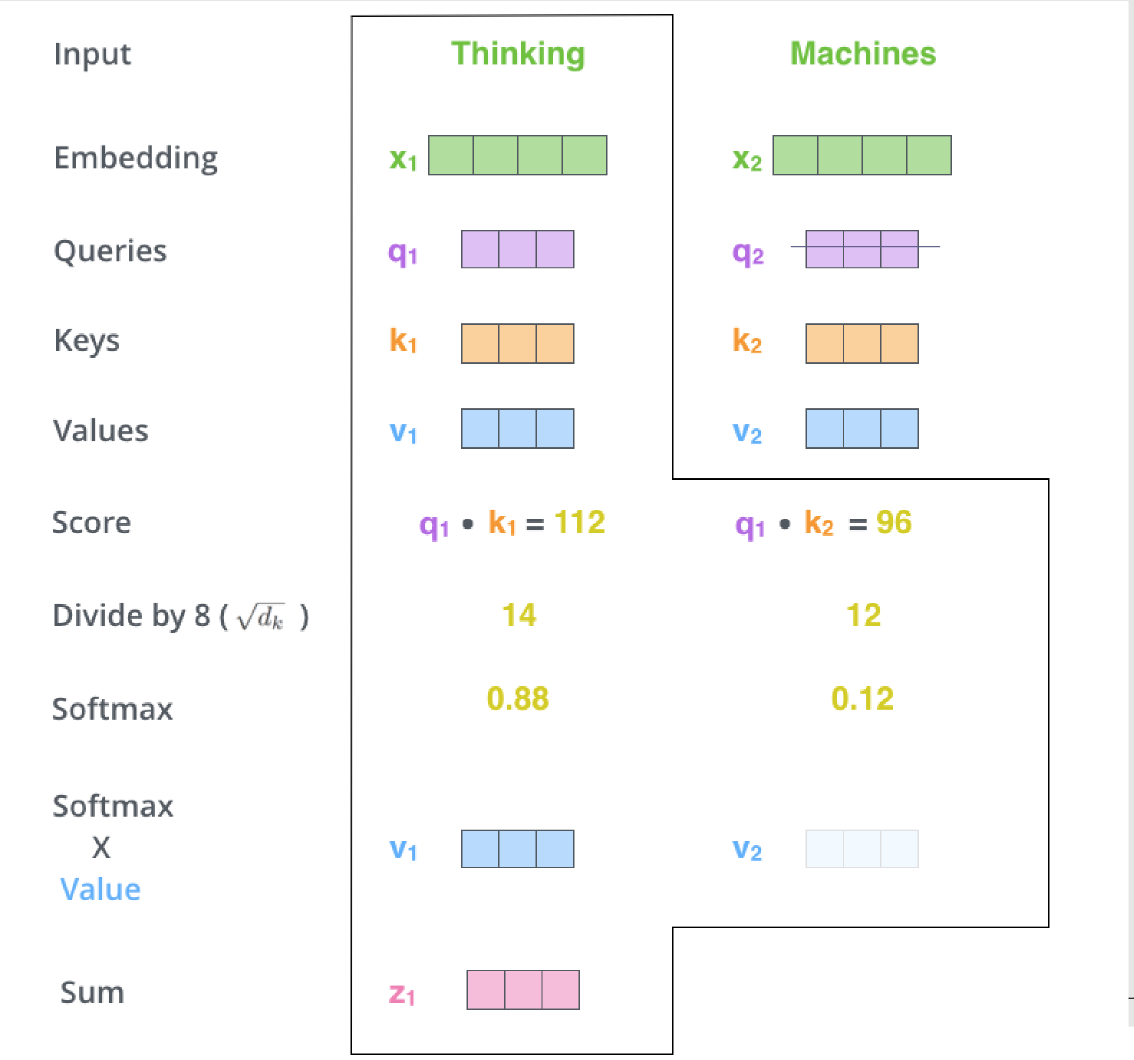


It focuses on the **other** words, and learns their importance, to understand the current word better.

[The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)]

Attention computation





= word embedding (e.g., word2vec)

by multiplication with Wq, Wk, Wv

(implementation detail to stabilize training)

(always divide same amount of attention)

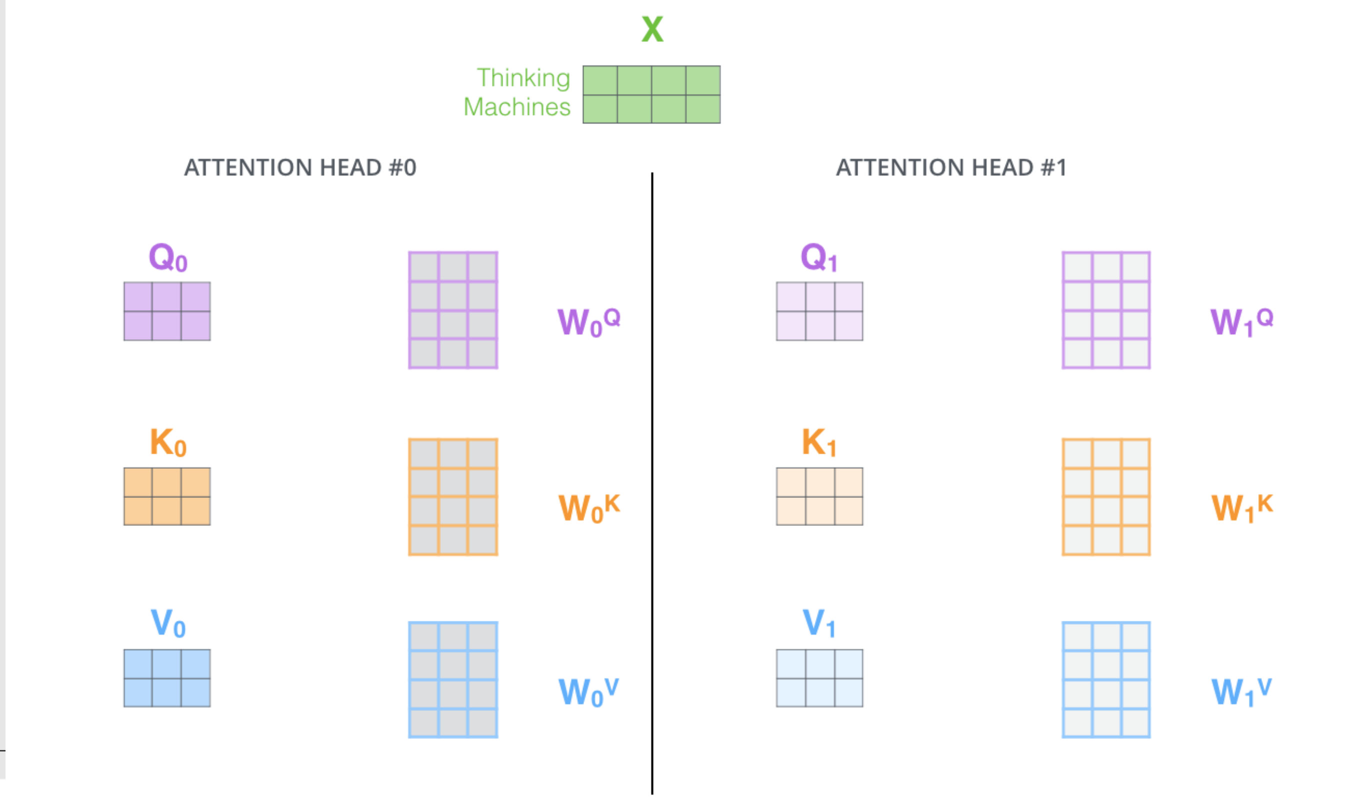
linear transformation to obtain new representation after the self-attention

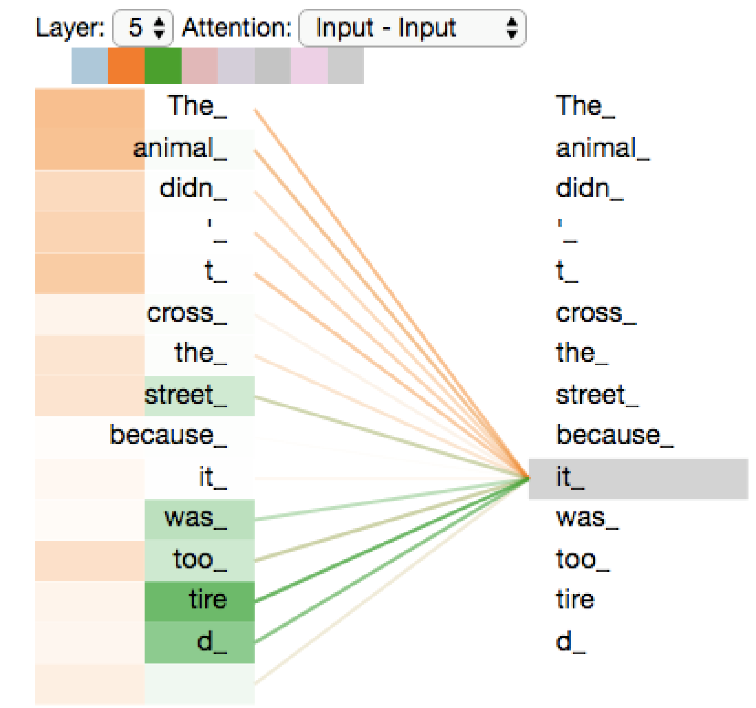
combined representation of word and other words

[The Illustrated Transformer]

Multi-head attention

- each head learns a different focus
- "representation subspaces"



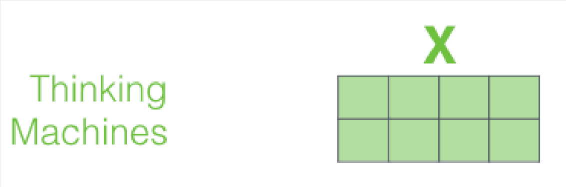


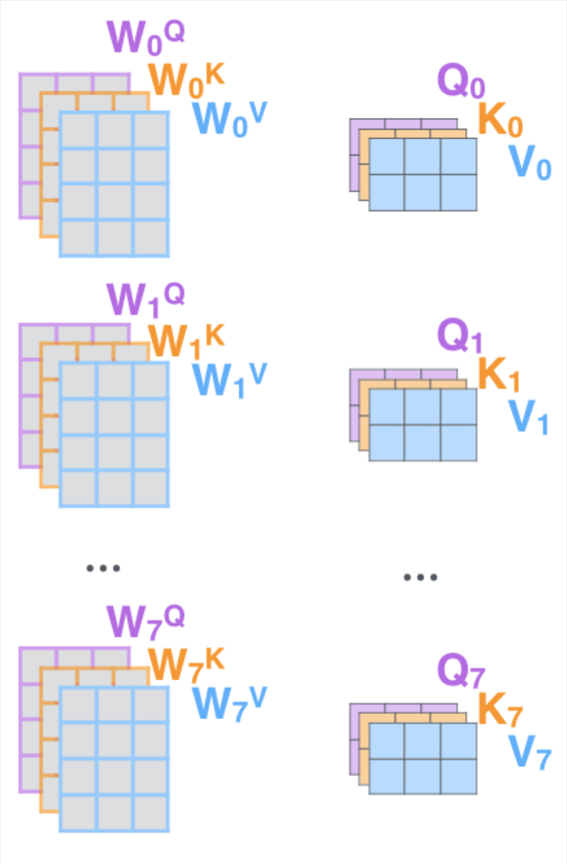
original Transformer has 8 attention heads

representation = concatenation of the vectors from all attention heads

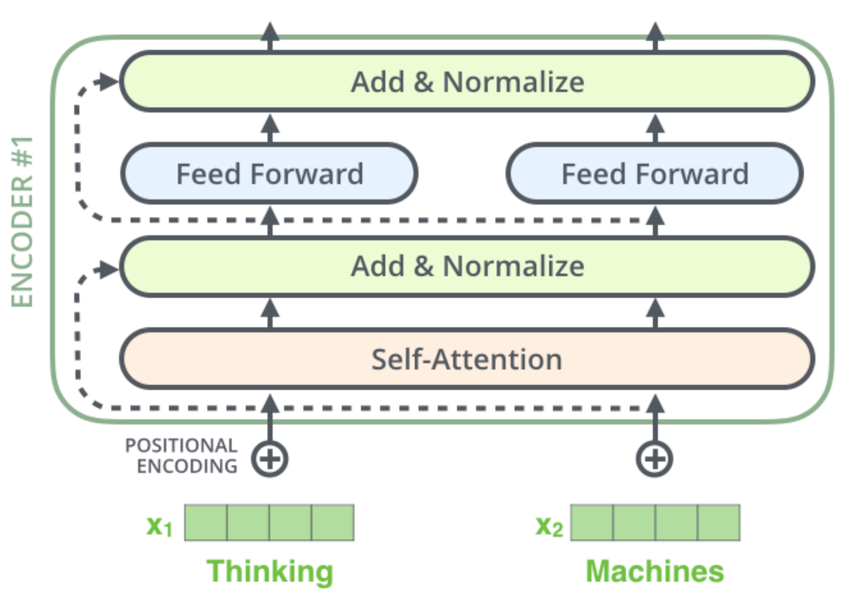
[The Illustrated Transformer]

Residual & Parallel





matrix computations to process all words at **once**



residual connections:

- like the standard Resnet-based architectures
- to enable **deep** architectures (i.e., many stacked encoders and decoders)

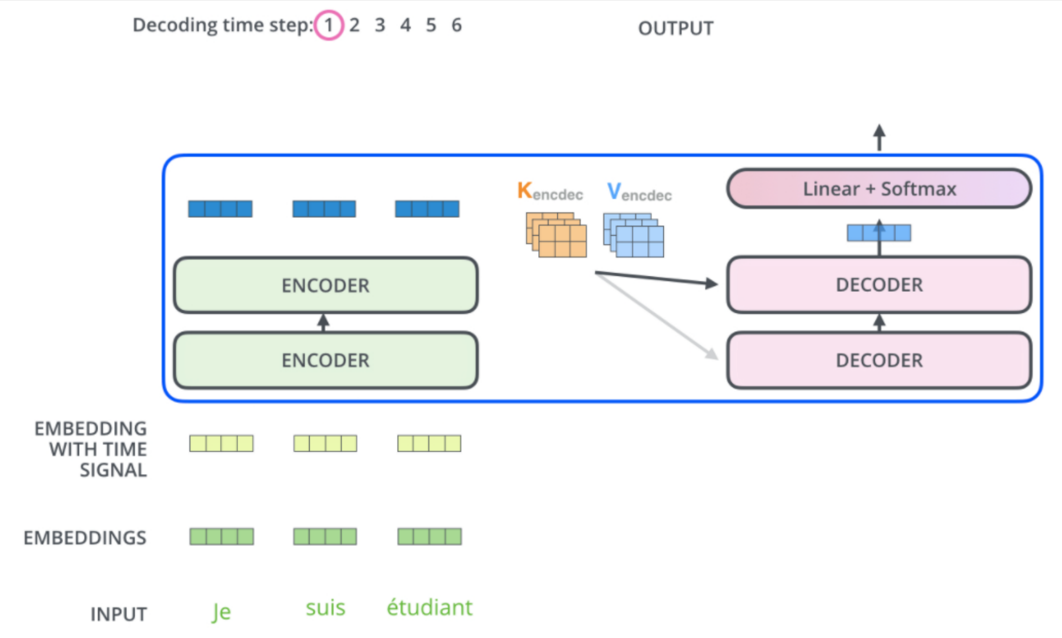
The Illustrated Transformer

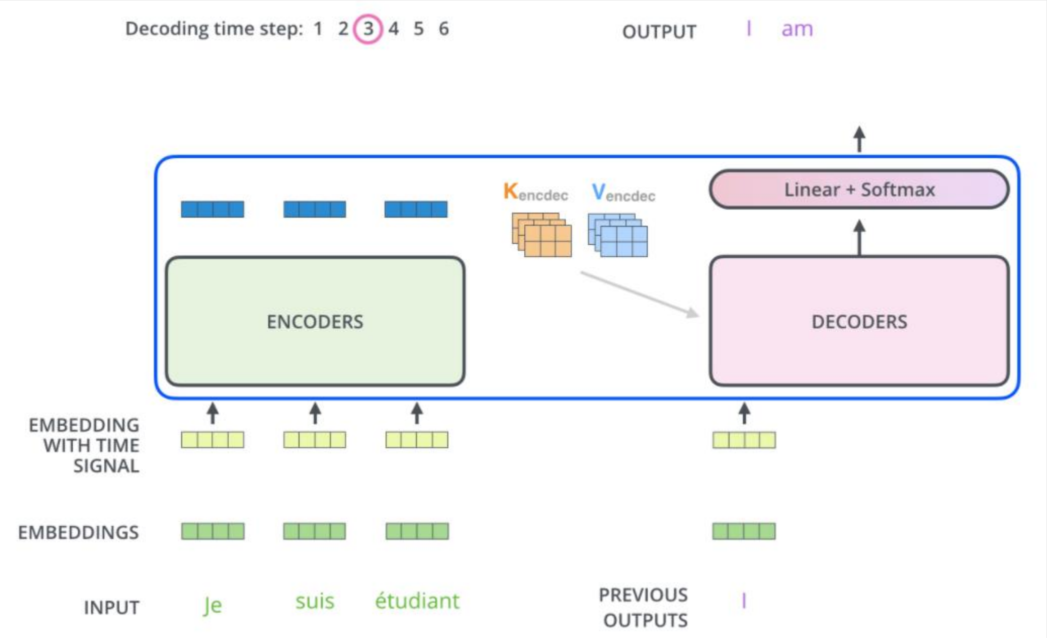
Decoding

- output of the encoder stack →

  connected to each of the decoders

- decoders are sequential
- word prediction is one-by-one
- until prediction of < end-of-sequence >

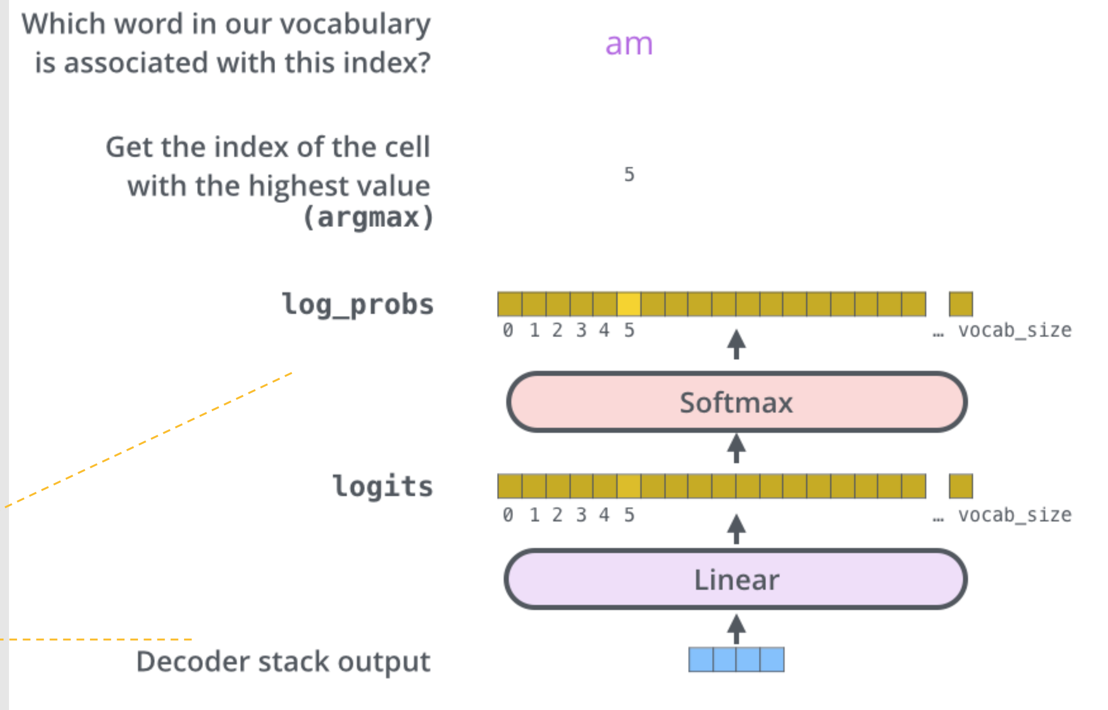


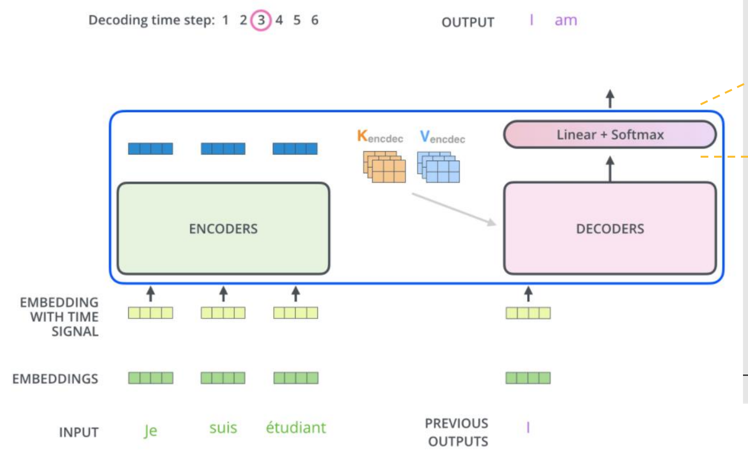


- decoders only look at earlier words
- masking future positions via (-inf) before softmax (so they don't count)

The Illustrated Transformer

Predicting each word





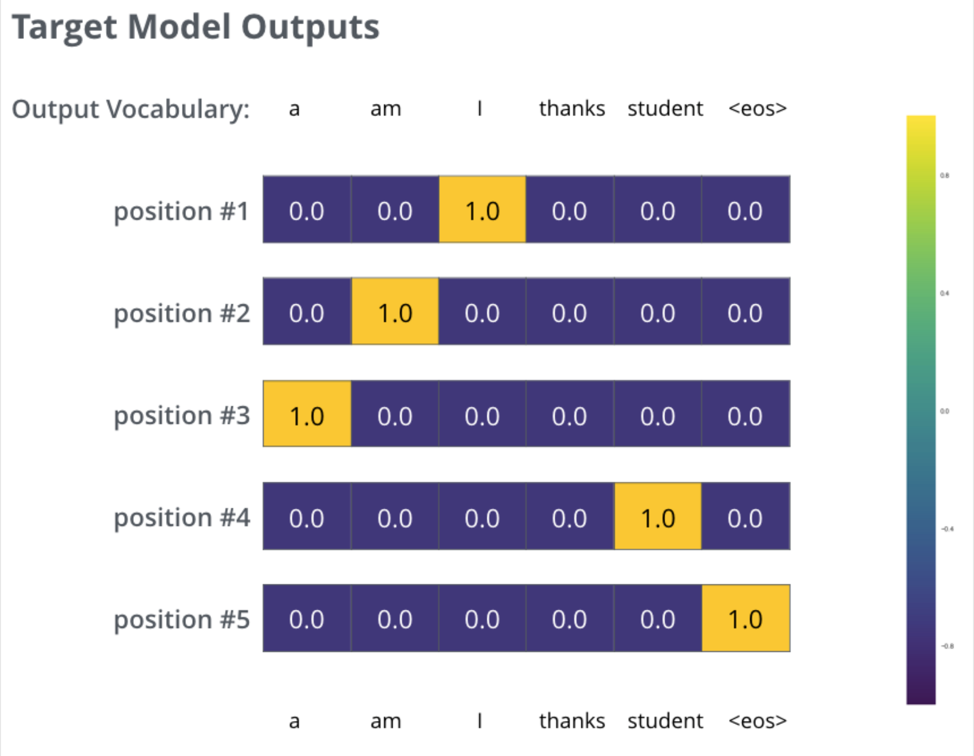
- huge vocabulary
- including

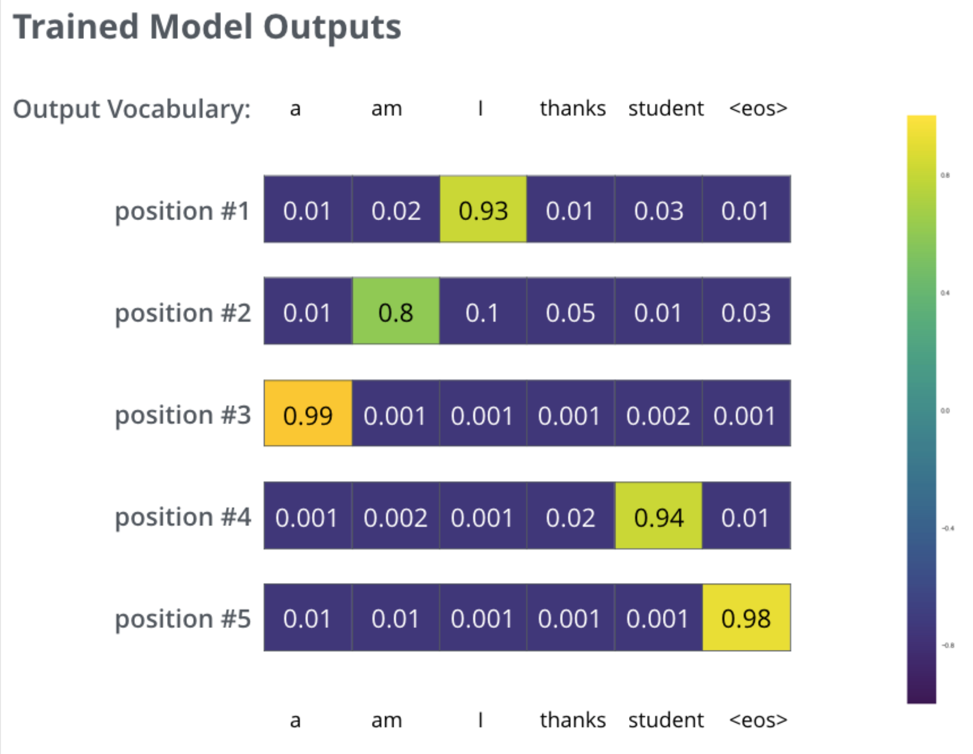
  <end-of-sentence>

[The Illustrated Transformer]

Training

- loss on the output set of words by standard cross-entropy in end-to-end training scheme

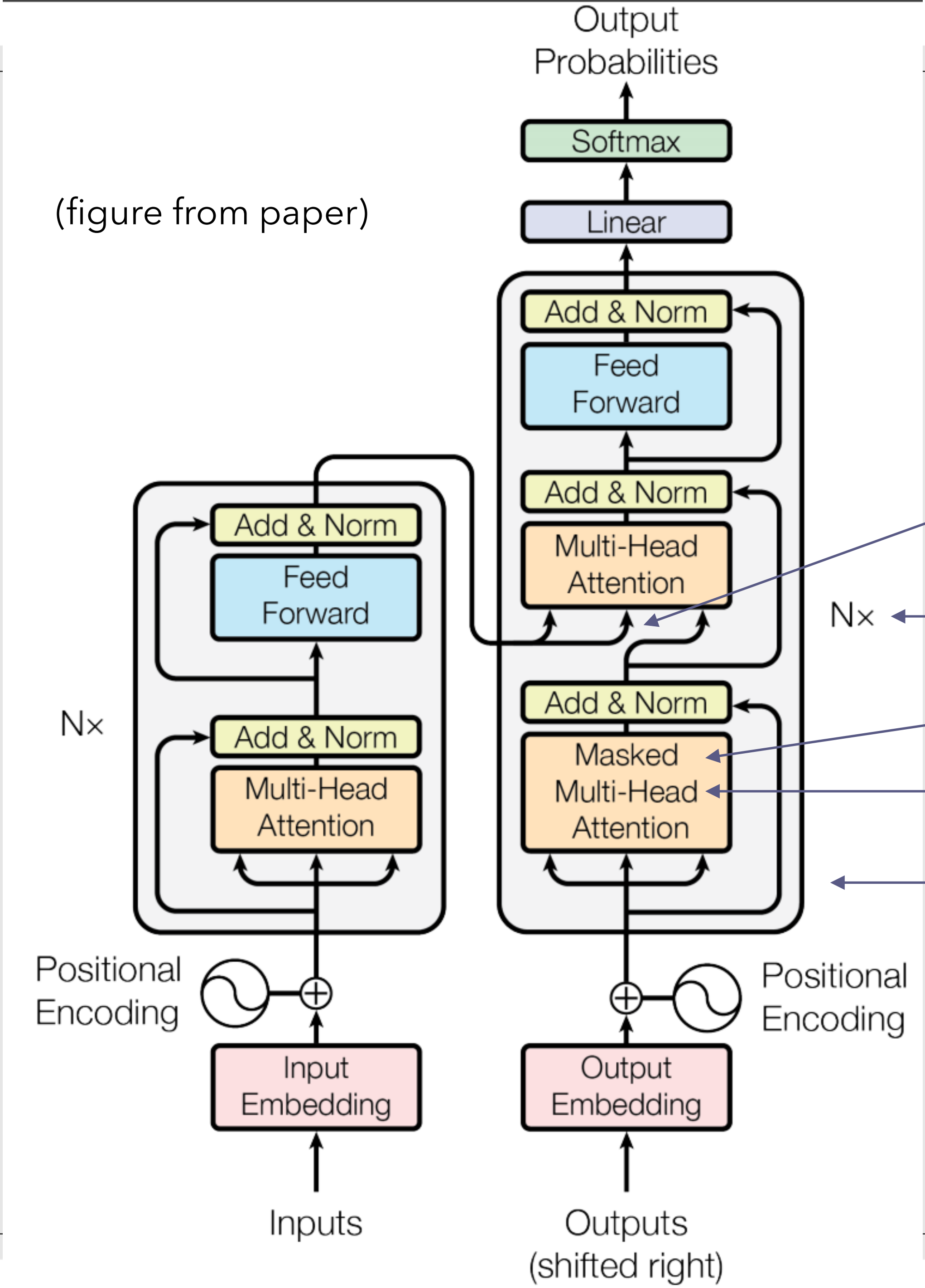




[The Illustrated Transformer] [Transformers Explained Visually (Part 2): How it works, step-by-step | by Ketan Doshi | Towards Data Science]

Transformer

(figure from paper)



= mixing encoded words and already predicted words

= stack

= masking the future

= multiple attention “subspaces”

= residual connection enabling deep stacks

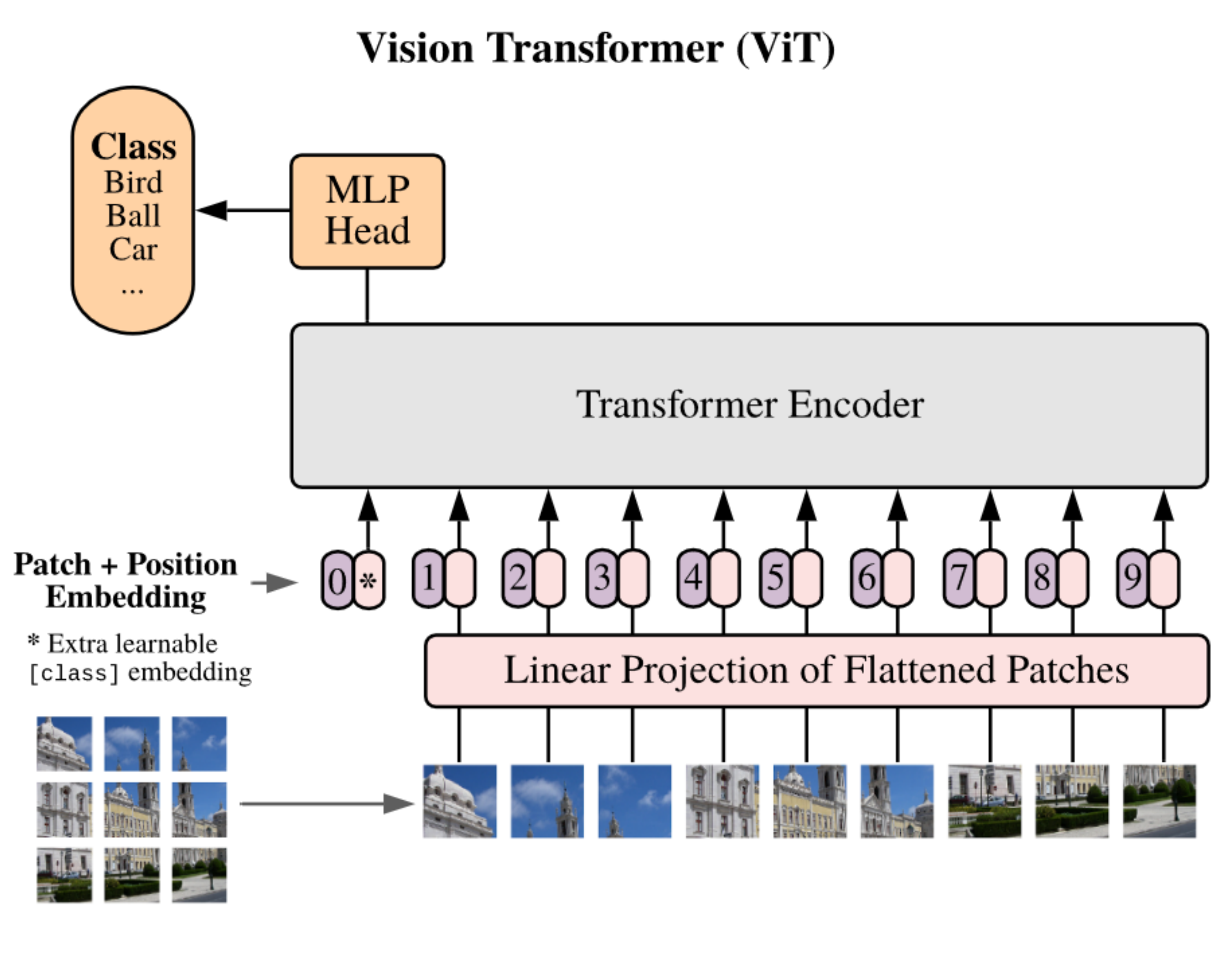
= temporal pattern for each word

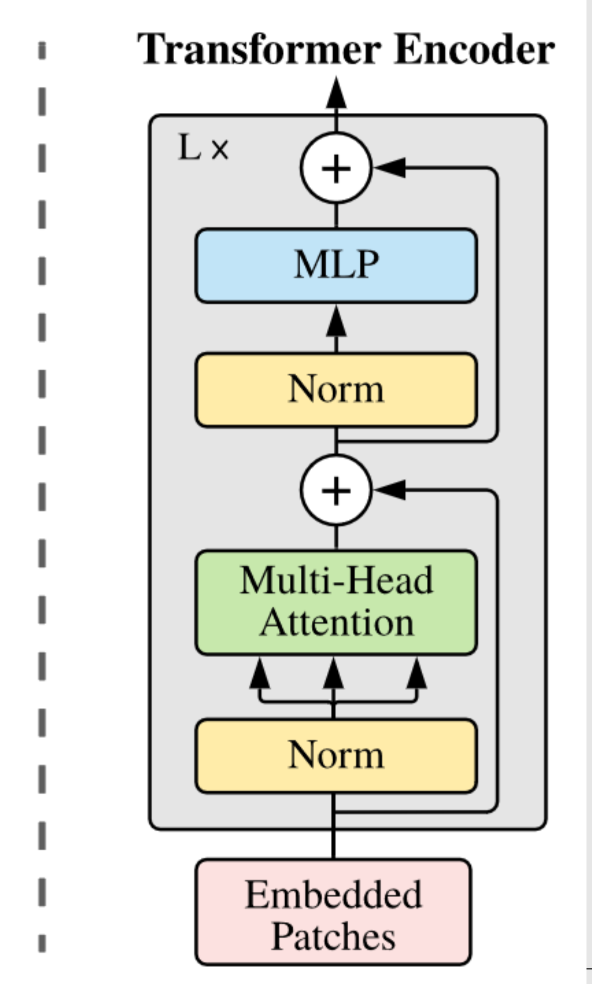
[1706.03762] Attention Is All You Need (arxiv.org)

Translating these ideas into Computer Vision

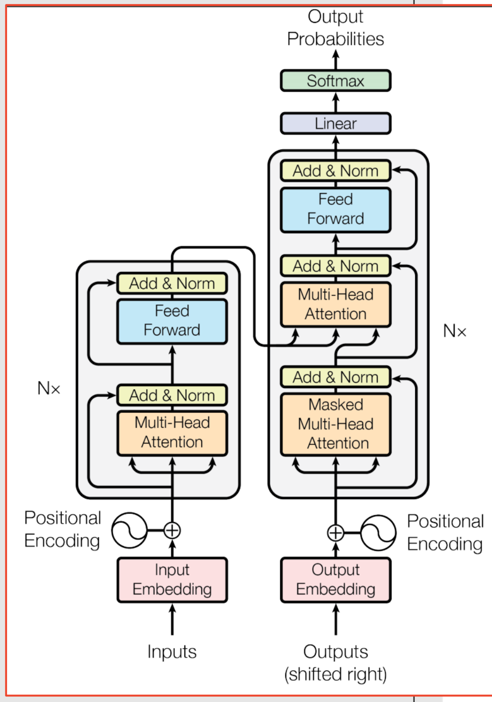
The Concept

no decoder



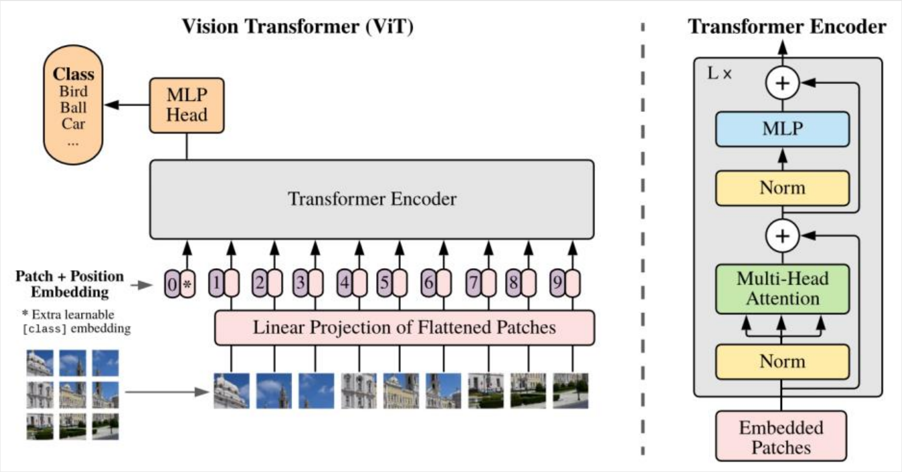


original



[[2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arxiv.org)]

Vision Transformer (ViT)



- first full Transformer architecture for Vision
- 16 x 16 patches as ‘words’
    - each patch = 16x16x3 (=768d)
- lack inductive biases by CNN (translation)
- has other inductive bias: permutation invariance
- huge pre-training (Imagenet doesn’t suffice)

= hybrid

CNN-Transformer

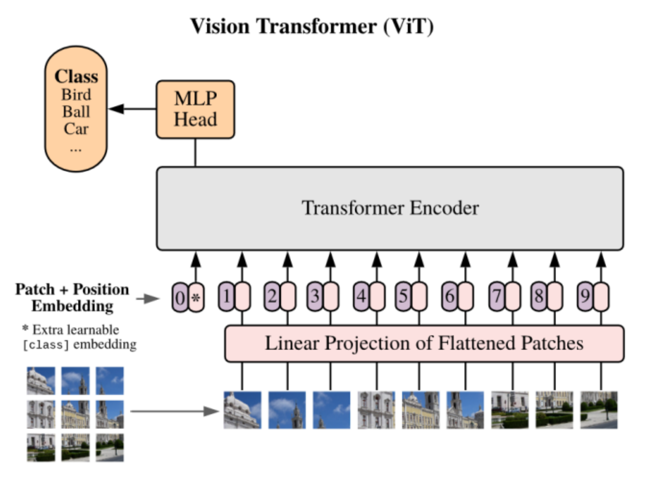
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	–
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	–
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	–
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	–
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

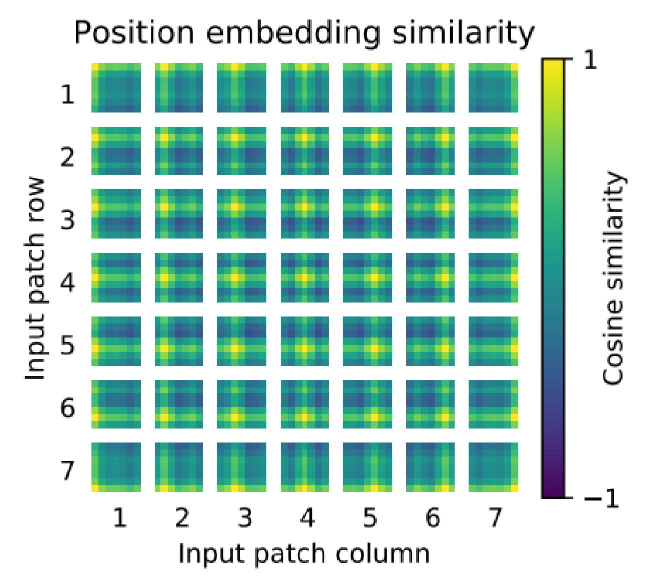
[2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arxiv.org)

Vision Transformer (ViT)

- Specifically, if ViT is trained on datasets with more than 14M images it can approach or beat state-of-the-art CNNs.
- If not, you better stick with ResNets or EfficientNets.

- Even though many positional embedding schemes were applied, no added value was found
- Therefore: Learnable position embedding

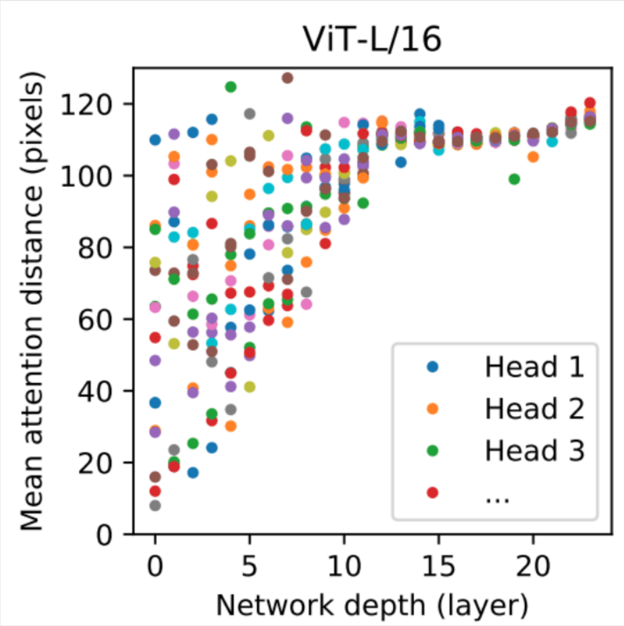


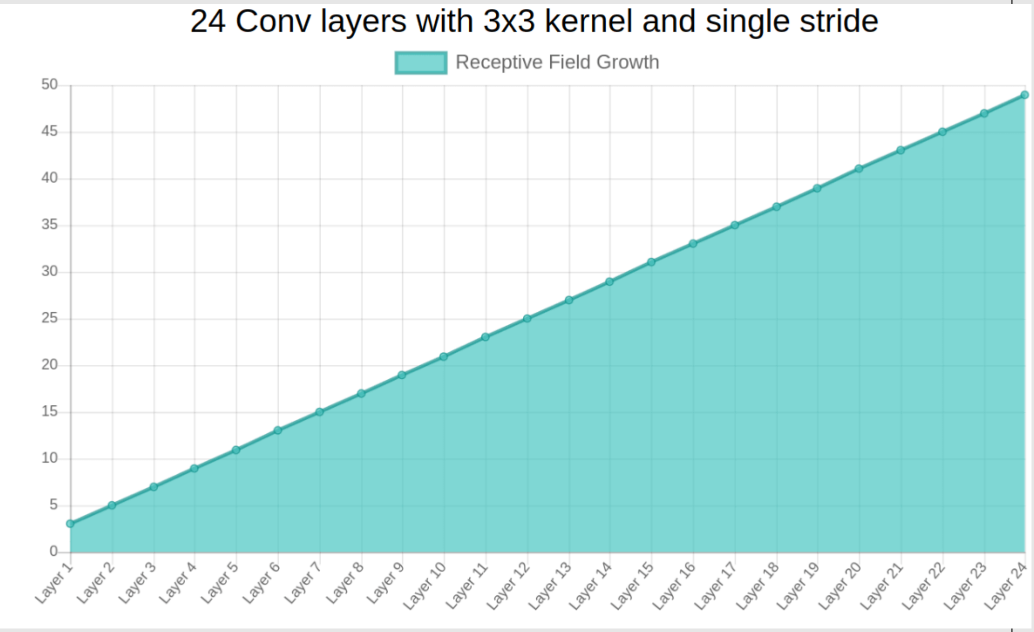


https://theaisummer.com/vision-transformer/ [2010.11929] An Image is Worth 16x16 Words

Long-range relations!

= attention heads (CNN)

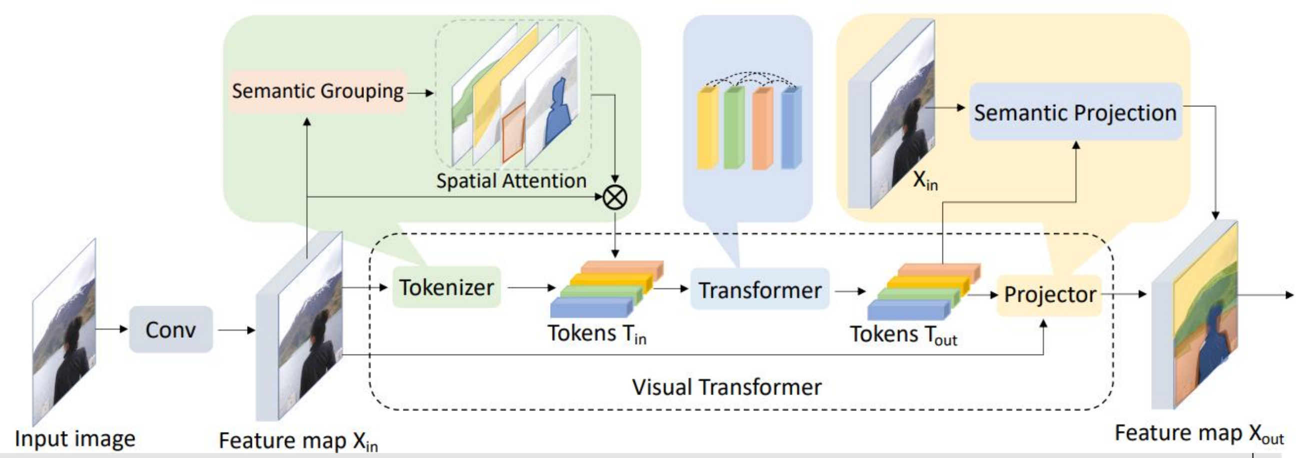




https://theaisummer.com/vision-transformer/ [2010.11929] An Image is Worth 16x16 Words

Visual Transformer (≠ViT)

- hybrid model of CNN (= tokens) &
- Transformer (= model relations between tokens)



Tokens-To-Token Vision Transformers (T2T-ViT)

- enable training on Imagenet only
- deep-narrow architecture to capture local image features (which the ViT fail to do)
- tokens-to-token model to capture features in neighborhoods

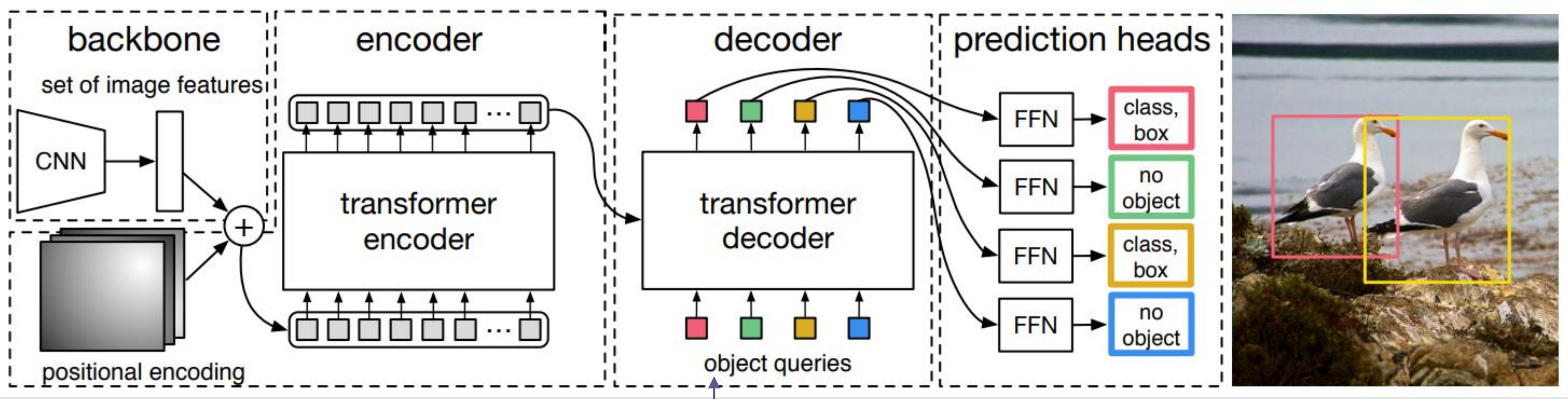
Data-efficient Image Transformer

- enable training on Imagenet only
- trained in ~3 days
- by student-teacher setup with CNN as a teacher

[2006.03677] Visual Transformers: Token-based [2101.11986] Tokens-to-Token ViT [2012.12877] Training data-efficient image transformers

Object Detection (DETR)

- detection transformer: relations between objects (co-occurrence!)
- decoders: ‘translate’ representations to boxes with labels
- fixed-size set of N predictions (N >> #objects, many ‘no object’ predictions)
- end-to-end training; removing hand-designed components (e.g., anchors, non-max suppression)

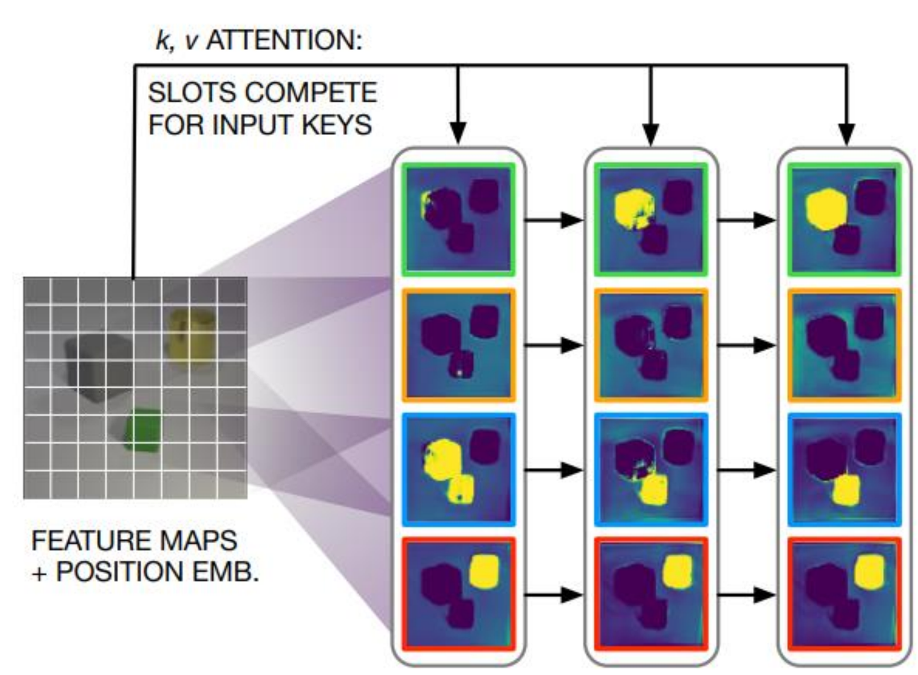


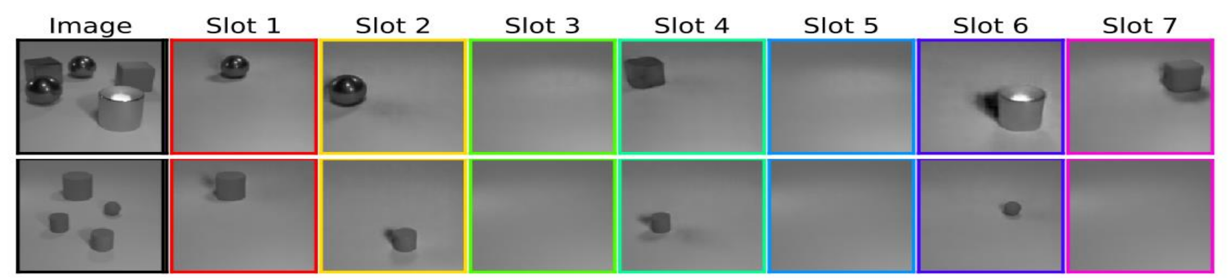
= learned object priors (kind of anchors)

[2005.12872] End-to-End Object Detection with Transformers (arxiv.org)

Object-centric: Slot Attention

- object centric: localized attributes
  - standard DL will learn spurious correlations:
    - e.g., yellow → contains cube (=spurious, coincidental)
  - force learning of localized "**slot**" for "gray + cube", "yellow + cylinder"
    - cutting the spurious correlation "yellow ~ cube"
    - disentangle!
- slots
  - compete for explaining parts of the input via a softmax-based attention mechanism
  - inputs can be pixels, CNN, etc.
- achievement: better generalization
  - to new scenes and objects

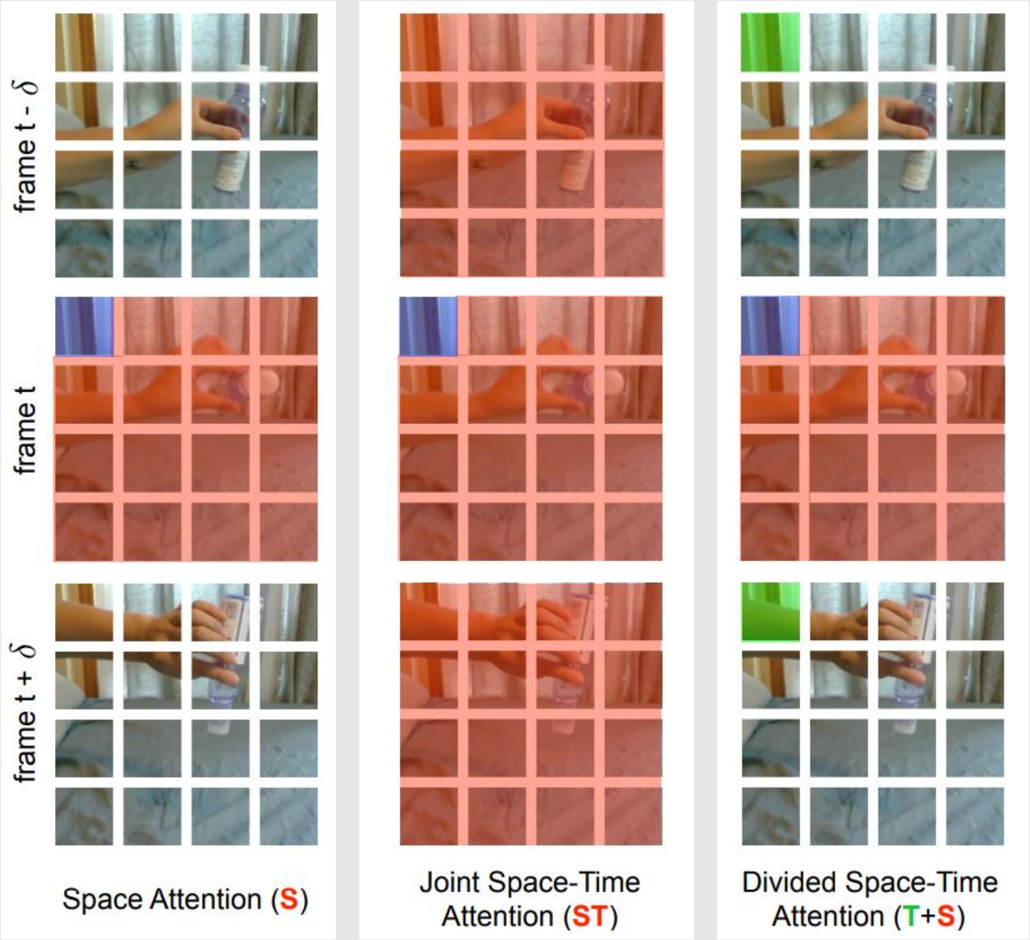




[2006.15055] Object-Centric Learning with Slot Attention (arxiv.org)

Space-time Transformer

- TimeSformer
- video
- space-time attention



"divided S-T attention"

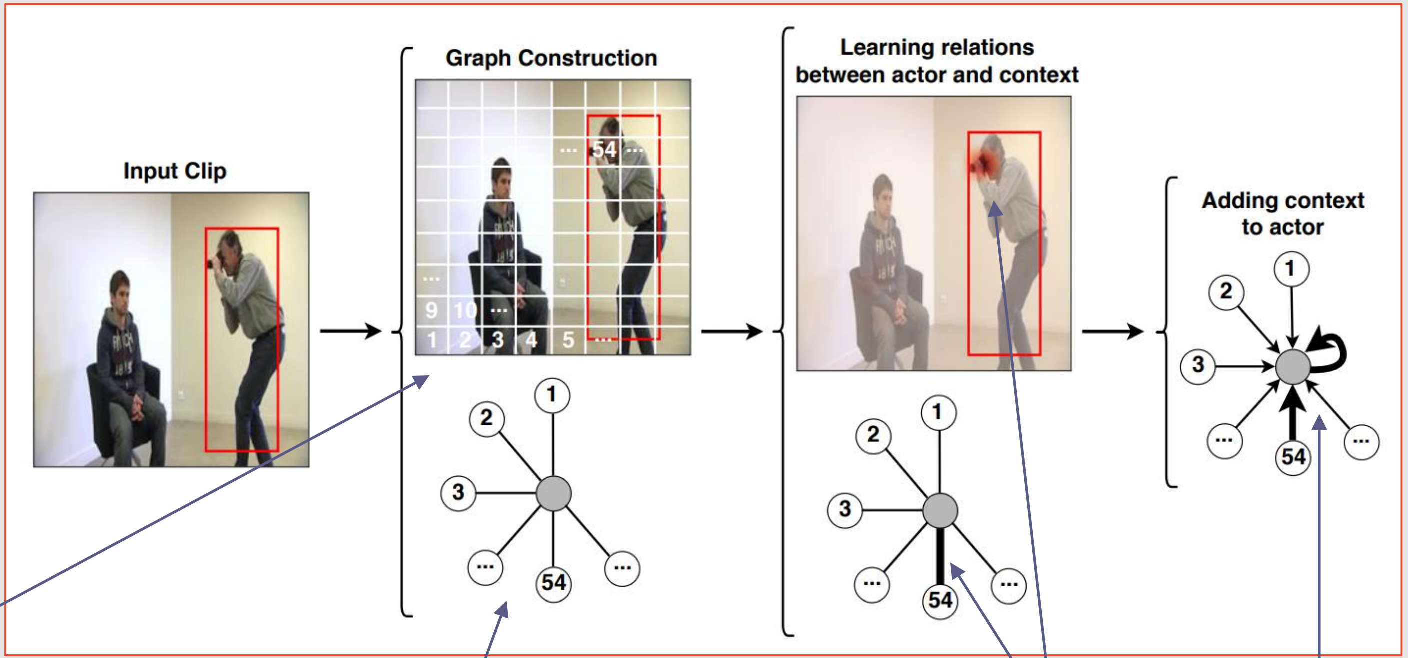
works best

[2102.05095] Is Space-Time Attention All You Need for Video Understanding? (arxiv.org)

Human-Object Interactions

- attention (= actor * context)
- multi-heads (=2)
- positional encoding

- largely CNN-based (video, I3D)
- not fully Transformer (hybrid)
  - yet very useful and effective (perf. & vis.)



patches as 'words'

(encoded by I3D)

interactions actor (red) vs. context (grid)

(fully connected)

= learning attention

learned attention

sum

= repr.

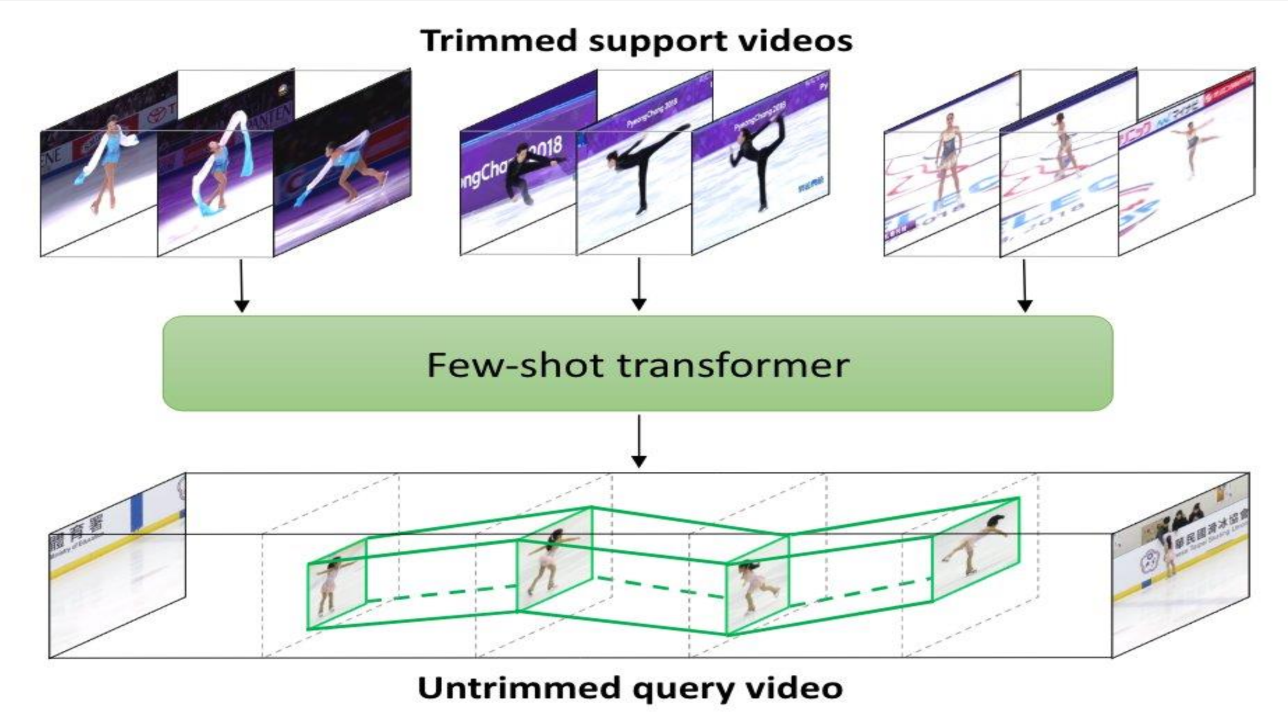
+= acc.

[Spot What Matters: Learning Context Using GCN]

Summary & Discussion

Summary

- Transformers are “here to stay”, also for Vision
  - with a large community working on its developments



(appeared yesterday,
CVPR 2021, UvA)

Summary

- Transformers are “here to stay”, also for Vision
  - with a large community working on its developments
- Many opportunities for applications that we’re working on already
  - objects in context (scene & situation understanding)
  - spatially distant relations / interactions (sports analysis)
- New possibilities
  - long range temporal interactions (scenario recognition)
- Inspiration for new components
  - attention, positional encoding, modeling patches & frames as a sequence
- Training can be difficult
  - not as efficient as finetuning CNN, but steps are being made
  - no common best practices yet, but that will come

Tnx for *your* Attention ☺

Hope it was useful!

(Some opportunities for Intelligent Imaging on next slide)

Opportunities for Intelligent Imaging

- long-range interactions in the image
  - interesting! e.g., sport: location / relation between (distant) players
  - can we enforce sparsity? (Wouter) – maybe by slots? see next point
- better generalization by slot attention
  - few-shot learning
- relations between objects
  - scene graphs
  - can we include prior knowledge? (Fieke)
- other applications?

References – Model

- The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)

  [best starting point for introduction]
- [1706.03762] Attention Is All You Need (arxiv.org) [original paper]
- Transformers Explained Visually (Part 2): How it works, step-by-step | by Ketan Doshi | Towards Data Science [masking]

- Attention? Attention! (lilianweng.github.io) [attention mechanism: key, value, query]
- How Attention works in Deep Learning: understanding the attention mechanism in sequence models | AI Summer (theaisummer.com)
- CSC421/2516 Lecture 16: Attention (toronto.edu) [incl. image attention in image-caption models]

- Transformers are Graph Neural Networks | NTU Graph Deep Learning Lab [sentences are fully-connected word graphs]

References – Vision

- [2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arxiv.org)
- [2006.03677] Visual Transformers: Token-based Image Representation and Processing for Computer Vision (arxiv.org)
- [2101.11986] Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet (arxiv.org)
- [2012.12877] Training data-efficient image transformers & distillation through attention (arxiv.org)

- [2005.12872] End-to-End Object Detection with Transformers (arxiv.org) [object detection, DETR]
- [2006.15055] Object-Centric Learning with Slot Attention (arxiv.org) [object segmentation]

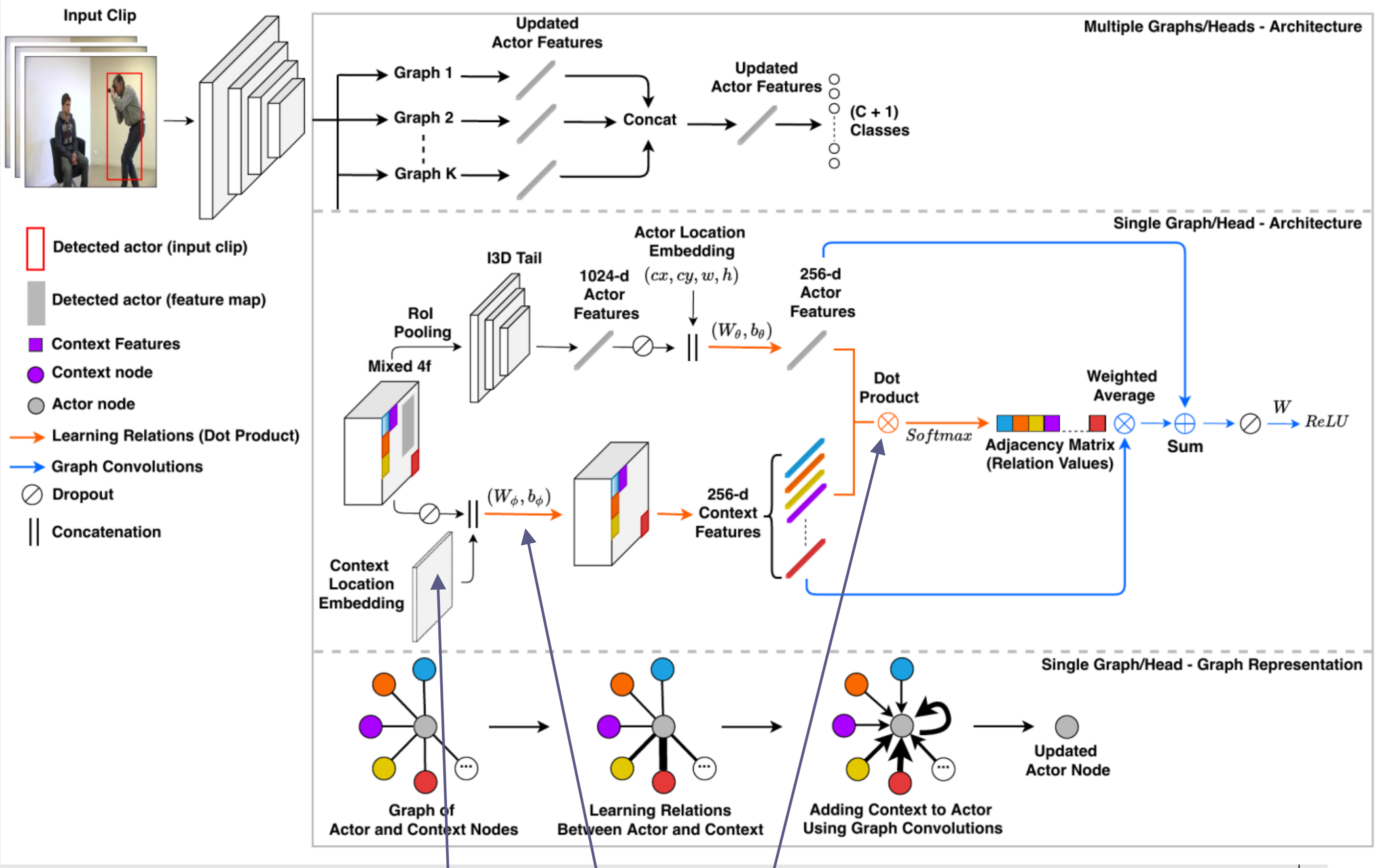
- GitHub - micts/acgcn: Code for the paper "Spot What Matters: Learning Context Using Graph Convolutional Networks for Weakly-Supervised Action Detection" [our work]
- [2102.05095] Is Space-Time Attention All You Need for Video Understanding? (arxiv.org) [video, multi-frame, TimeSformer]
- [2103.01209] Generative Adversarial Transformers (arxiv.org) [image generation, Gansformer]

References – Advanced (Fieke, Raimon, Wouter)

- [2103.14030] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (arxiv.org)
- [2006.08084] Neural Execution Engines: Learning to Execute Subroutines (arxiv.org) [learning programming by sequences of smaller routines]
- Neuro-Symbolic Deductive Reasoning for Cross-Knowledge Graph Entailment (aaai-make.info) [infer the set of all facts that are a logical consequence of current and potential facts of a knowledge graph]
- paper4.pdf (ceur-ws.org) [named entities and relations]
- [2002.05544] Superpixel Image Classification with Graph Attention Networks (arxiv.org) [beyond rectangular-gridded images, such as 360-degree field of view panoramas]
- Reasoning-RCNN: Unifying Adaptive Global Reasoning Into Large-Scale Object Detection (thecvf.com) [scaling DETR to Visual Genome sized datasets with >1000 object classes]
- [2103.04037] Perspectives and Prospects on Transformer Architecture for Cross-Modal Tasks with Language and Vision (arxiv.org) [visuolinguistic cross-modal tasks]
- [1908.02265] ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks (arxiv.org)
- [2102.10772] Transformer is All You Need: Multimodal Multitask Learning with a Unified Transformer (arxiv.org)

Appendix

Action Detection



= actor-object interaction

= pos. emb.

= learnable attention towards object

[Spot What Matters: Learning Context Using GCN]