Transformers: Attention is all you need!

Florian LEMARCHAND
Univ. Rennes, INSA Rennes, IETR - UMR CNRS 6164

VAADER “AI for Image” Reading Group – 2020-21-01

Let's break the ice!

Who is behind that microphone?

- Florian Lemarchand
- INSA Rennes Engineer« Electronique et Informatique Industrielle (EII) », 2018
- PhD Student since October 2018:
  - Lab: "Institut D'Electronique et des Technologies du numéRique de Rennes" (IETR)
  - Team: " Video Analysis and Architecture Design for Embedded Resources " (VAADER)
  - PhD Founded by " Pole d'Excellence Cyber " (PEC) → Bretagne council and French ministry of armed forces
  - Advisors : Erwan Nogues and Maxime Pelcat
  - PhD Subject:
    - "Recognition of Images and Intercepted Signal using Artificial Intelligence "
  - Technical Domains :
    - Image Restoration
    - Machine (Deep) Learning
- More information on my research webpage!
- Hobbies: Sport(s) and Nature, especially those two associated

The Paper

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

- Paper[1] presented at NeurIPS 2017, 16k citations
- 60 + submissions with “transformer” in the title at ICLR21
- Highlighted by GPT-3 [2], state of the art of language models

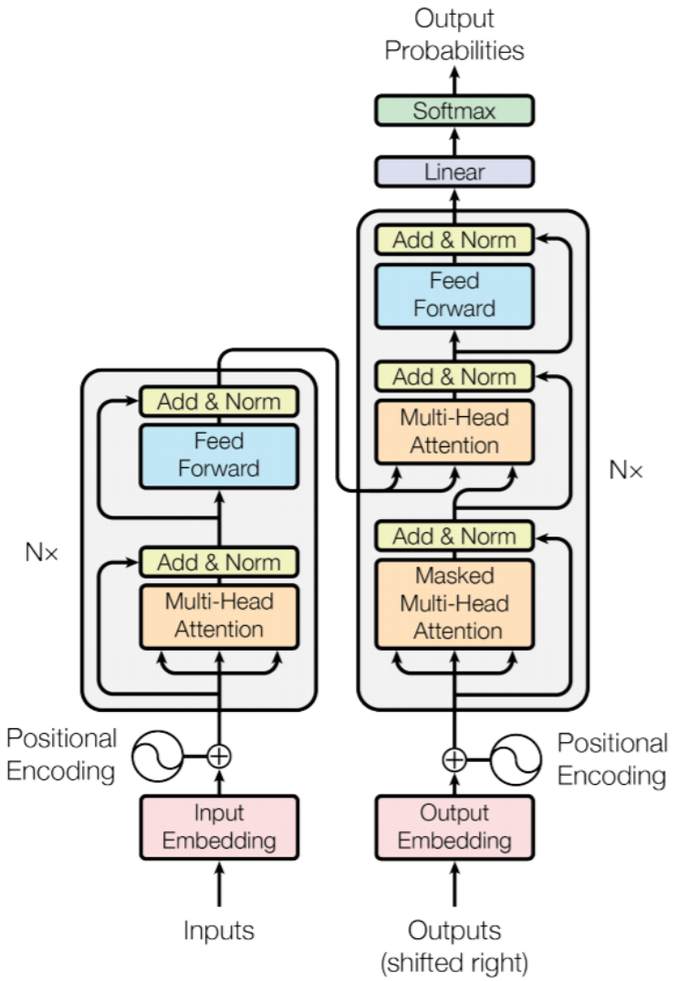


Figure 1: The Transformer - model architecture.

[1] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
[2] Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

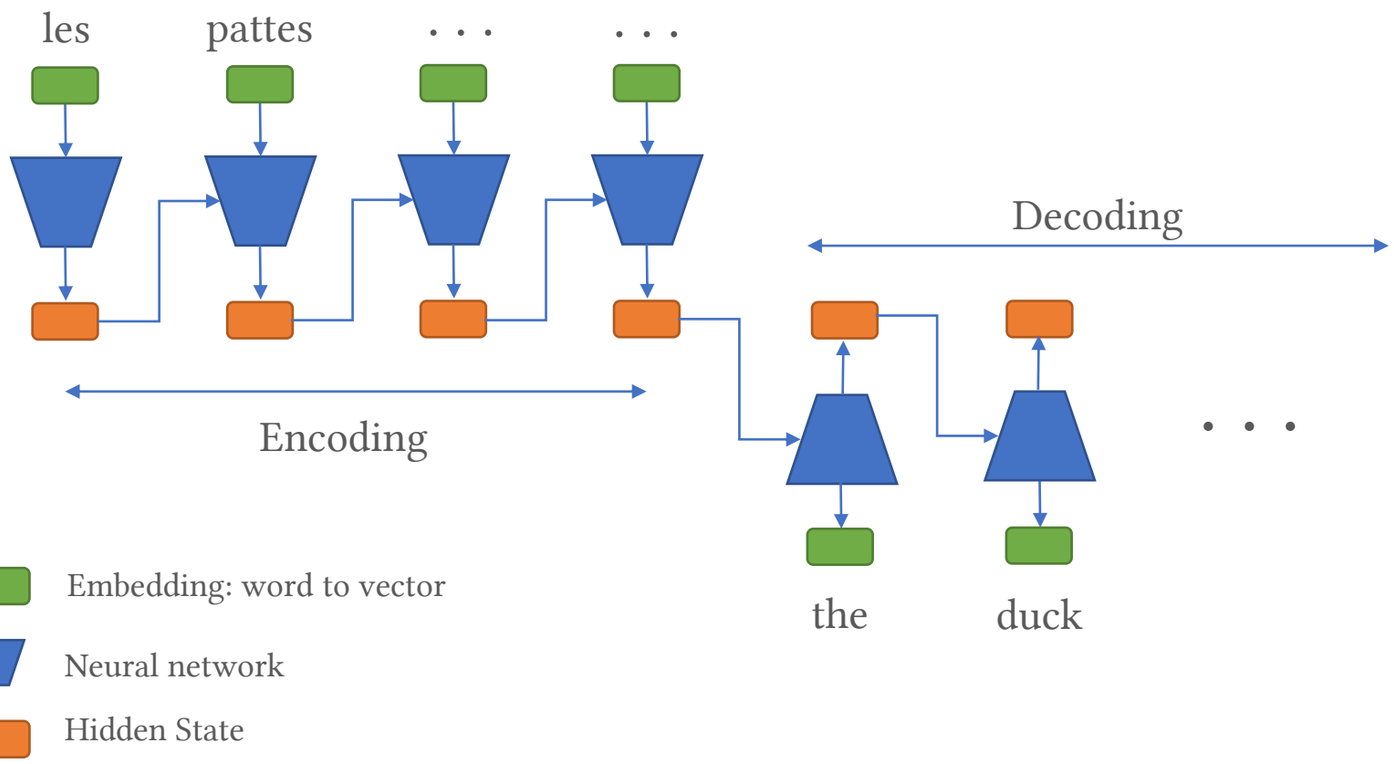
Context

Sequence to Sequence (Seq2Seq) models

French to English

Les pattes de canard → The duck paws

Recurrent Neural Network (RNN)



Issues:

- Problem of long memory
  - Decode only using current hidden state
  - Addressed through attention mechanisms
- Sequential processing
  - Decoding a state requires processing previous ones
- Complicated to train:
  - Long gradient path → vanishing gradients

The Paper

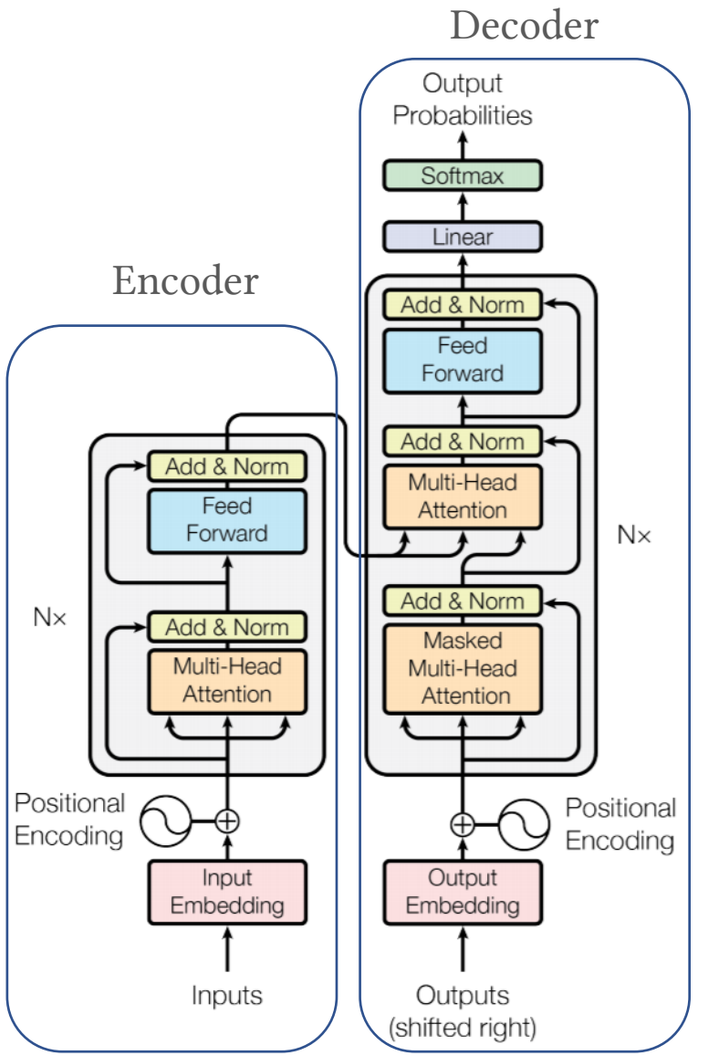


Figure 1: The Transformer - model architecture.

The Paper

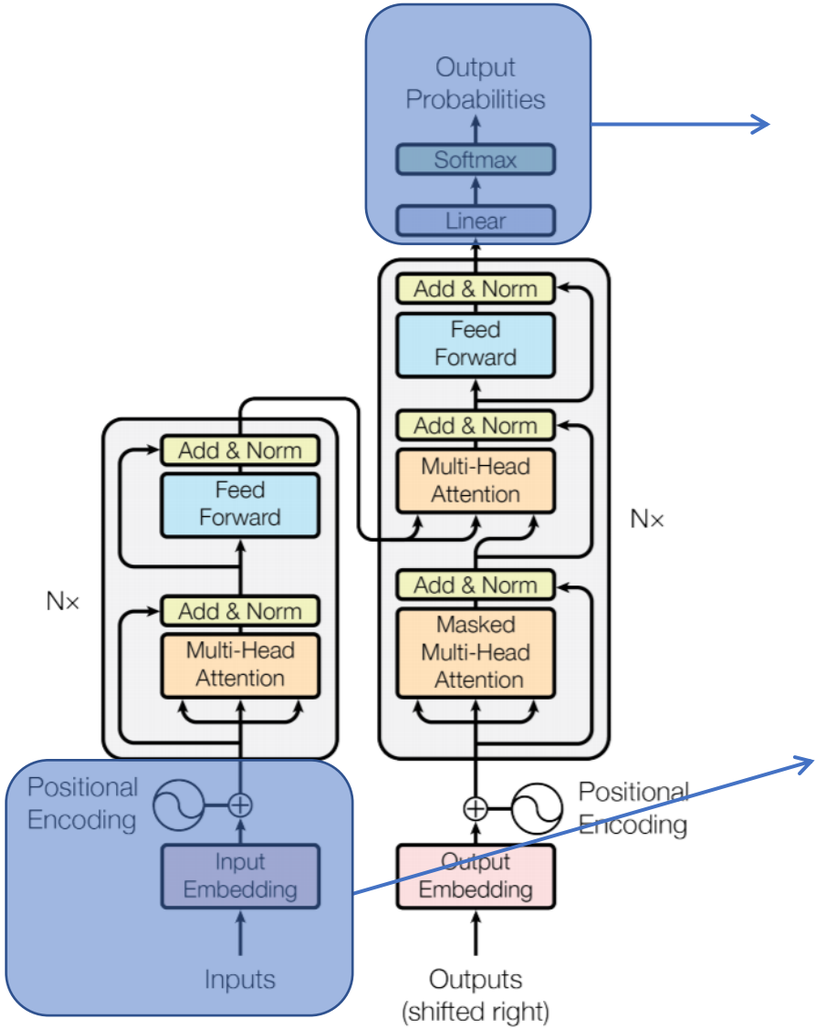
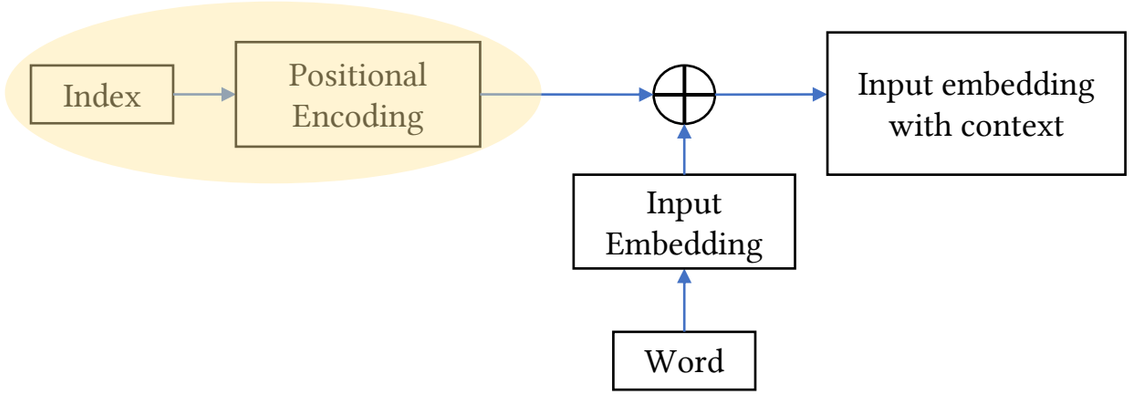


Figure 1: The Transformer - model architecture.

Output "probability" vector to predict the next word of the sequence

Enable position in sequence coding to avoid recurrence



Word	This	presentation	is	amazing
Index	0	1	2	3

The Paper

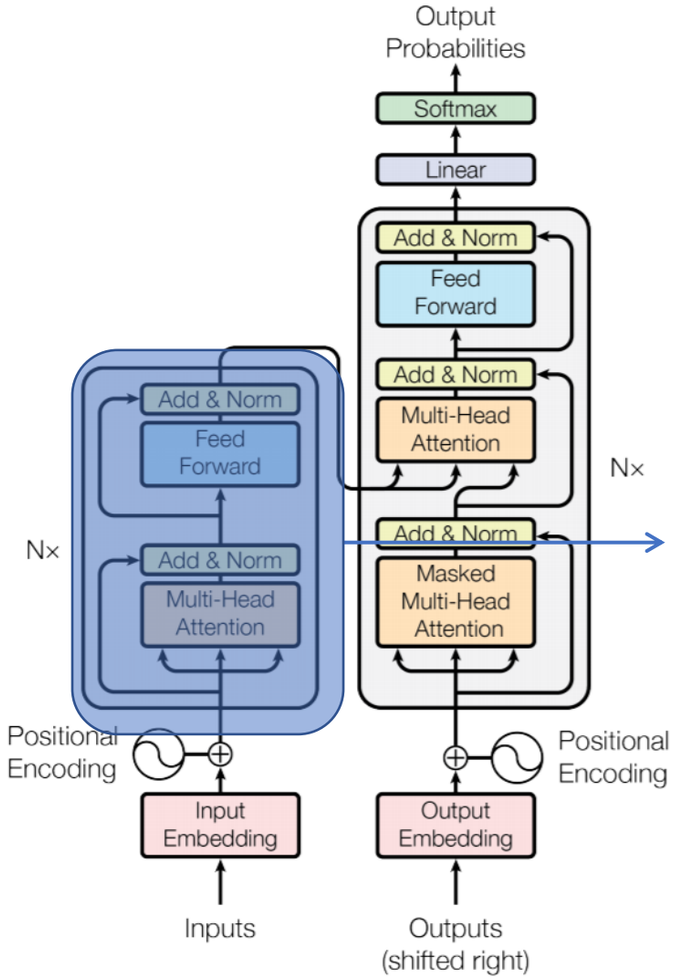


Figure 1: The Transformer - model architecture.

Self-Attention Block

Focus on the most important part of the input

This
Presentation
is

Related?

This	presentation	is	amazing
This	presentation	is	amazing
This	presentation	is	amazing

$[1 \quad \cdots \quad 0.1]$

$[0.5 \quad \cdots \quad 0.05]$

$[0.01 \quad \cdots \quad 0.5]$

Extract several attention vector per word (multi-head), concatenate all vectors and merge them to get the final descriptor for a word

The Paper

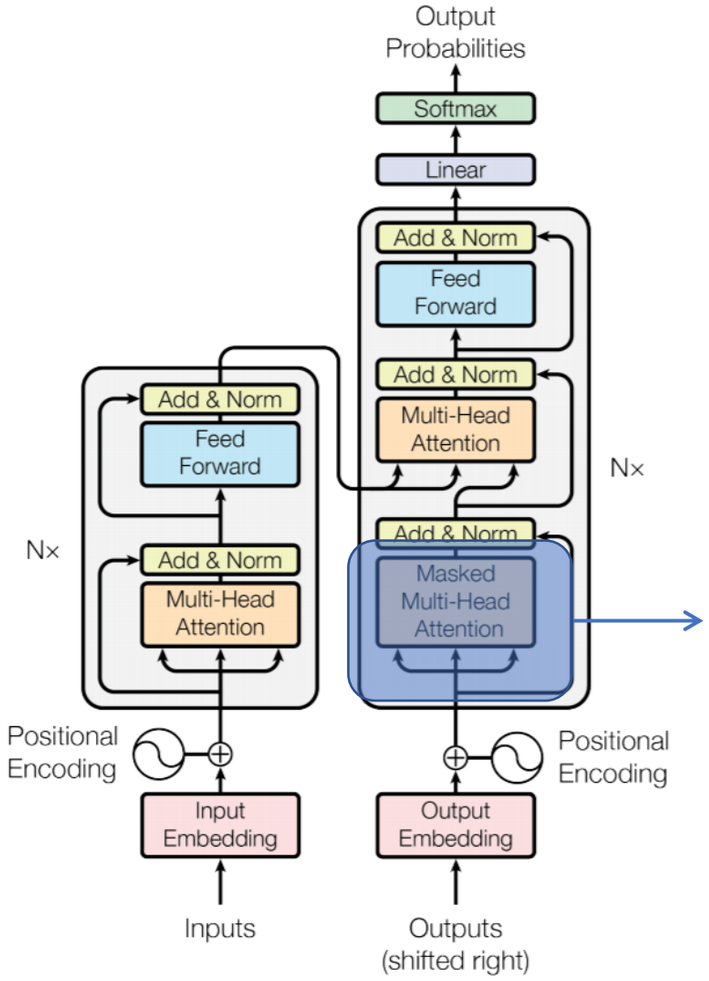


Figure 1: The Transformer - model architecture.

Masked Self-Attention Block

Focus on the most important part of the target but using only previous words

Cette
Présentation
est

Related?

Cette	**présentation**	**est**	**incroyable**
Cette	**présentation**	**est**	**incroyable**
Cette	**présentation**	**est**	**incroyable**

$[1 \quad \cdots \quad 0]$

$[0.5 \quad \cdots \quad 0]$

$[0.02 \quad \cdots \quad 0]$

The Paper

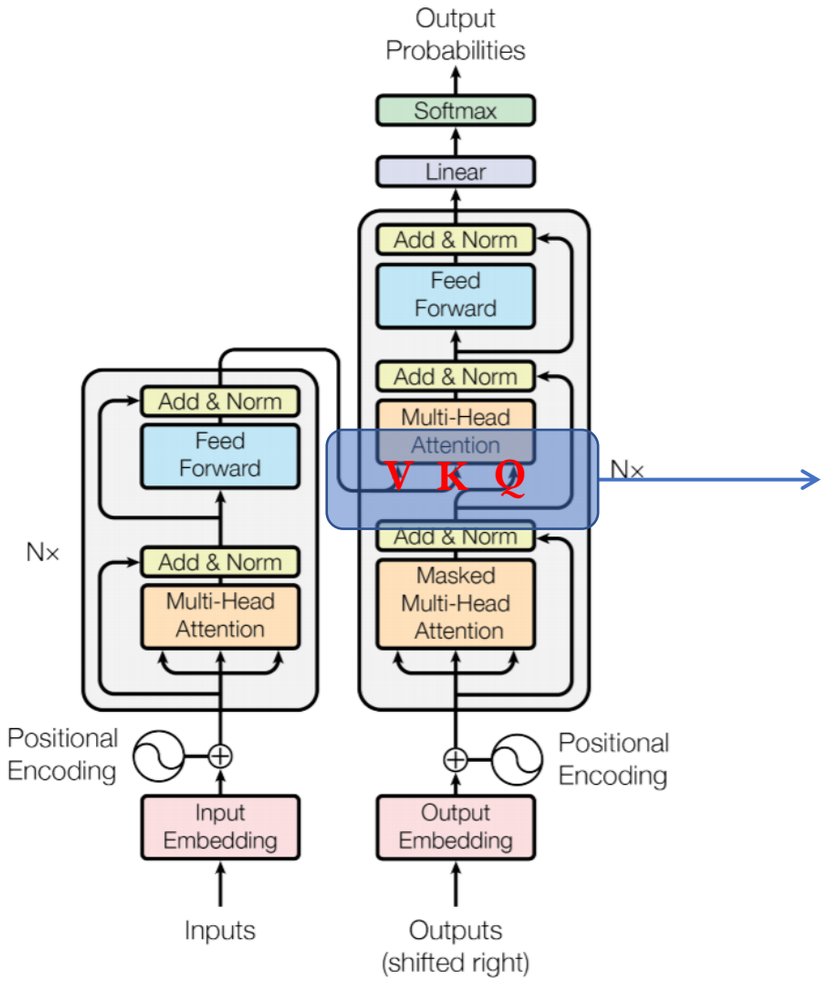


Figure 1: The Transformer - model architecture.

Keys Values Queries

- Encoder learns representative values V to describe the input.
  - Feature extraction
- The values are indexed and can be accessed using Keys.
- Decoder outputs Queries to "question" the encoder part
- The central Multi-Head Attention links input values V (using keys K) with output queries Q to define which values are interesting to predict the output word

Experiments

Training

- WMT English-German → 4,5 M sentence pairs , 37k tokens (unique words)
- WMT English-French→ 36 M sentence pairs , 32k tokens (unique words)
- Trained 3,5 days on 8 NVIDIA P100 GPUs

- Use of model averaging, 5 or 20 last models averaged depending on the settings
- Metric: BLEU[1] (Bilingual Evaluation Understudy) → evaluation metric for machine translation

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

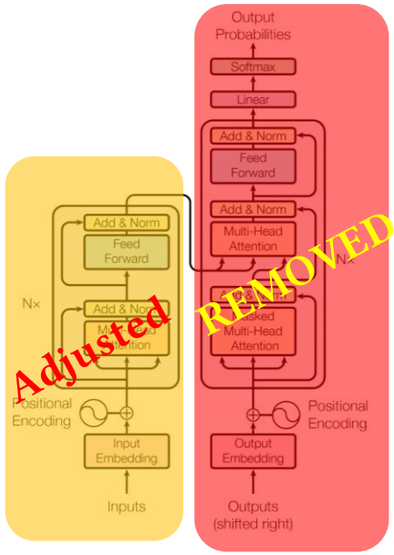
[1] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

Application to Image Processing

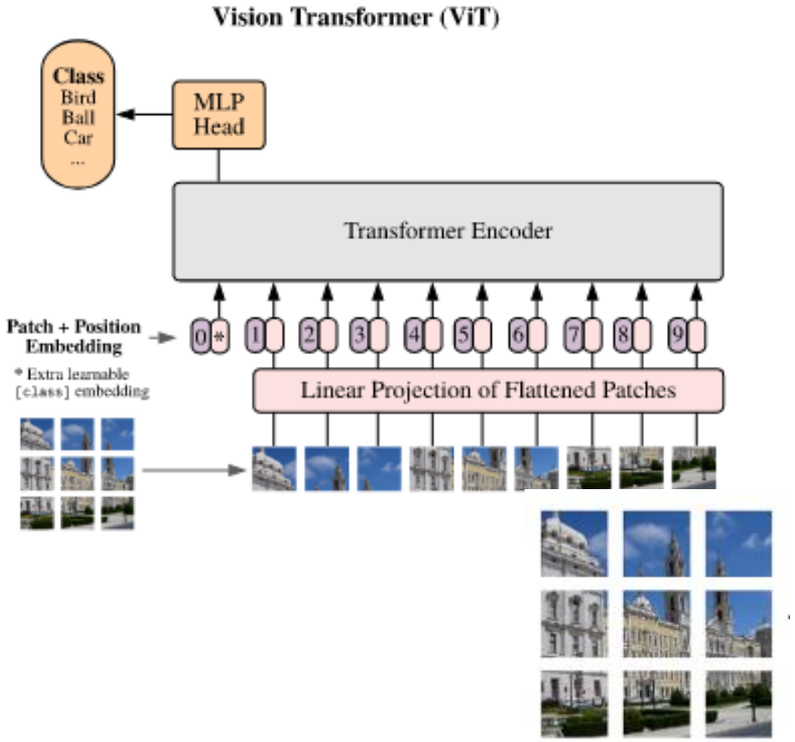
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

- Paper [1] to be presented at ICLR 2021 (oral presentation)
- Architecture called ViT for Vision Transformers



- Attention is quadratic operation
  - Attention score computed between each pair of inputs
  - Impossible to use an input per pixel as is
- Using patch as word instead of pixel reduce the number of connections



[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

Application to Image Processing

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

Experiments

Training

- Pre-trained on massive datasets live JFT-300M
- Then fine-tuned on the training sets of most known benchmarks

Gigantism of the model

- Trained thousands of days on TPUs ???
- Best model have 632M parameters

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

Table 1: Details of Vision Transformer model variants.

	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

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

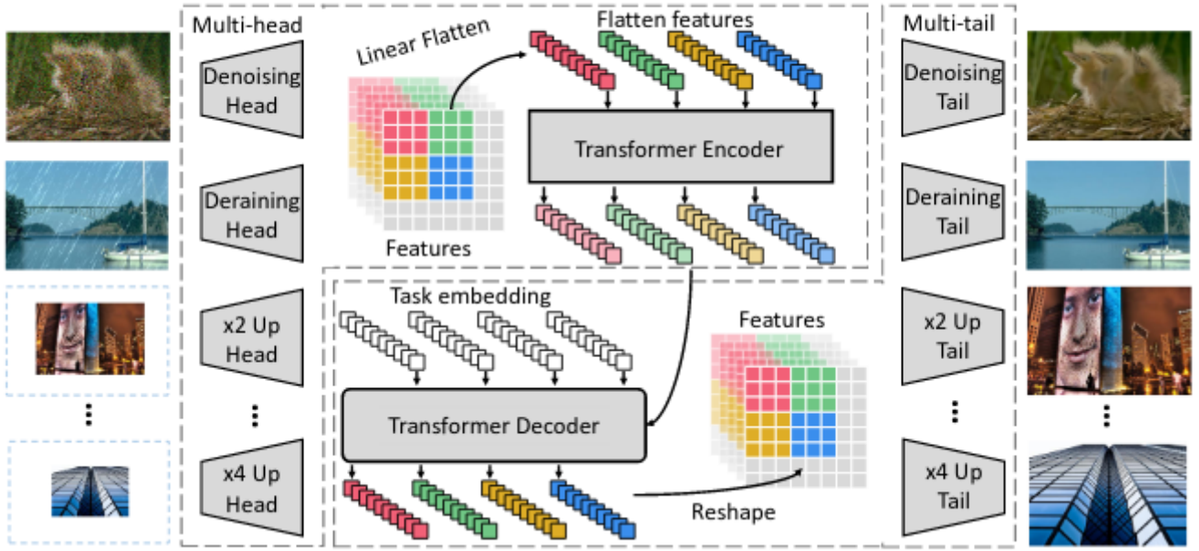
Application to Image Processing

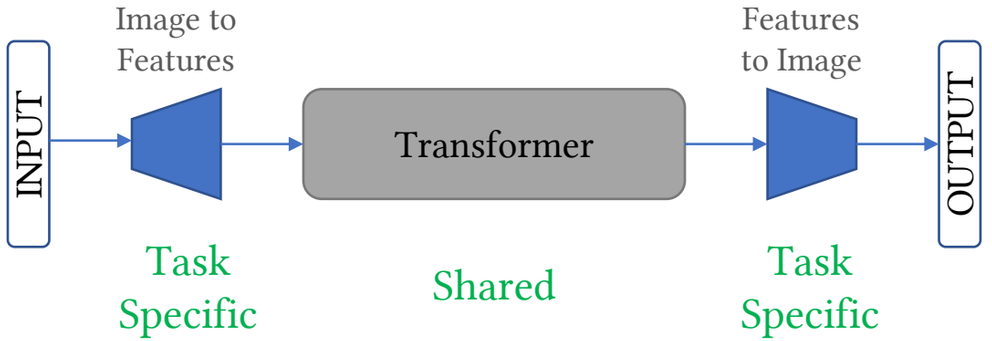
Submitted to CVPR 2021?

Pre-Trained Image Processing Transformer

Hanting Chen[1,2], Yunhe Wang[2]*, Tianyu Guo[1,2], Chang Xu[3], Yiping Deng[4],
Zhenhua Liu[2,5,6], Siwei Ma[5,6], Chunjing Xu[2], Chao Xu[1], Wen Gao[5,6]

[1] Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University. [2] Noah's Ark Lab, Huawei Technologies.
[3] School of Computer Science, Faculty of Engineering, The University of Sydney. [4] Central Software Institution, Huawei Technologies.
[5] Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University. [6] Peng Cheng Laboratory.
{htchen, tianyuguo, liu-zh, swma, wgao}@pku.edu.cn, c.xu@sydney.edu.au
{yunhe.wang, yiping.deng, xuchunjing}@huawei.com, xuchao@cis.pku.edu.cn





Training Strategy

for batch with imagenet
- select one task at every batch
- train specific head/tail and shared transformer

Fine-tuning on specific task:
- drop unnecessary heads/tails

[1] Chen, Hanting, et al. "Pre-trained image processing transformer." arXiv preprint arXiv:2012.00364 (2020).

Application to Image Processing

Pre-Trained Image Processing Transformer

Experiments

Training

- Pre-trained on 6x Imagenet → 60 M 48x48 patches
- Then fine-tuned on the training set of the desired task

Training & Fine-tuning. We use 32 Nvidia NVIDIA Tesla V100 cards to train our IPT model using the conventional Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ for 300 epochs on the modified ImageNet dataset. The initial learning rate is set as $5e^{-5}$ and decayed to $2e^{-5}$ in 200 epoch with 256 batch size. Since the training set consists of different

ation. After pre-training on the entire synthesized dataset, we fine-tune the IPT model on the desired task (*e.g.*, $\times 3$ single image super-resolution) for 30 epochs with a learn-

Table 4. Generation ability of our IPT model on color image denoising with different noise levels. Best and second best results are **highlighted** and underlined.

Method	BSD68		Urban100	
	10	70	10	70
CBM3D [14]	35.91	26.00	36.00	26.31
TNRD [12]	33.36	23.83	33.60	22.63
DnCNN [69]	36.31	26.56	36.21	26.17
MemNet [51]	N/A	25.08	N/A	24.96
IRCNN [70]	36.06	N/A	35.81	N/A
FFDNet [71]	36.14	26.53	35.77	26.39
RDN [77]	36.47	26.85	36.69	27.63
IPT (ours)	**38.30**	**28.21**	**39.07**	**28.80**

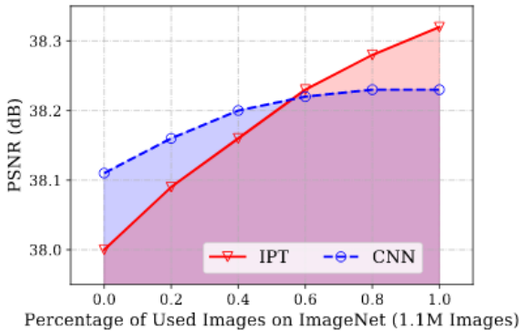


Figure 6. The performance of CNN and IPT models using different percentages of data.

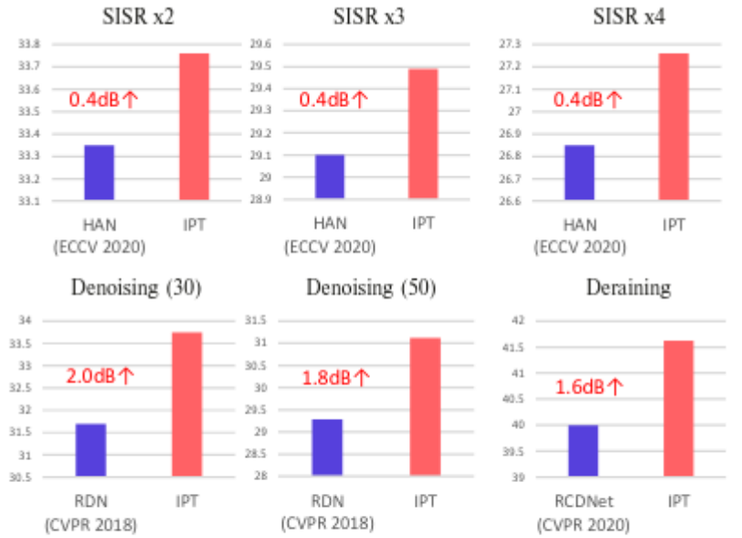


Figure 1. Comparison on the performance of the proposed IPT and the state-of-the-art image processing models on different tasks.

Application to Image Processing

Transformers in Vision: A Survey

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir,
Fahad Shahbaz Khan, and Mubarak Shah

Task	Input Data Type
Image Classification	2D Image
Image Classification	2D Image
Object Detection	2D Image
Object Detection	2D Image
Low Shot Learning	2D Image
Image Colorization	2D Image
Action Recognition	Skeleton
Super-resolution	2D Image
Multi-Model Learning	2D Images
3D Classification/Segmentation	CAD models, 3D object part segmentation
3D Mesh Reconstruction	2D Image
Vision and Language Navigation	Instruction text + RGBD panorama + Topological Environment Map
Referring Image Segmentation	2D Image + Language expression
Video Classification	Audio-Visual

[1] Khan, Salman, et al. "Transformers in Vision: A Survey." arXiv preprint arXiv:2101.01169 (2021).

Conclusion

Advantages:

- Break recurrence
  - Shorten the paths, better for gradient flow?
  - Better parallelization
- Less inductive bias than CNNs
  - In a CNN, limited receptive field
  - Attention mechanism enables a transformer to look at any part of the data/feature piece at (almost) any moment
- Better data-efficiency than CNNs? Perfs vs # of images

Drawbacks:

- Seems to require large datasets. Is it applicable on smaller datasets?
- Quadratic complexity for attention mechanism → transformers already old? Performers [1]

Open questions:

- Is gigantic pre-training on mainstream image processing tasks sufficient to target more specific tasks with limited data?
- Did you figure out that I cited an only already published paper?

Seems to be large models complicated to train

[1] Choromanski, Krzysztof, et al. "Rethinking attention with performers." arXiv preprint arXiv:2009.14794 (2020).