

Generative AI: Current Trends and Practical Applications

Kao Panboonyuen
panboonyuen.kao@gmail.com

<https://kaopanboonyuen.github.io>

Reference:

1. <https://pytorch.org/tutorials>
2. <https://stanford.edu/~shervine/teaching/cs-229/>
3. <http://introtodeeplearning.com/>
4. <https://www.simplilearn.com/tutorials/deep-learning-tutorial/introduction-to-deep-learning>
5. <https://www.geeksforgeeks.org/introduction-deep-learning/>

Outlines

- About Me
- Introduction to Generative AI
- Key Trends in Generative AI
- Applications of Generative AI
- Challenges and Ethical Considerations
- Future Directions and Opportunities

About Me



Kao
Panboonyuen

kao-panboonyuen 

AI Research Scientist

Name: Teerapong Panboonyuen (P'Kao)

Contact: teerapong.pa@chula.ac.th

panboonyuen.kao@gmail.com

Education: Ph.D. in Computer Engineering, Chula

Position: AI Team Lead, MARS

PostDoc, Chula

Visiting Lecturer, KKU

Interests: Computer Vision, Deep Learning

Machine Learning, Remote Sensing



Teerapong Panboonyuen

Senior Research Scientist at MARS, Post-doc at Chula
Verified email at chula.ac.th - [Homepage](#)

[Artificial Intelligence](#) [Machine Learning](#) [Deep Learning](#) [Computer Vision](#) [Remote Sensing](#)

FOLLOW



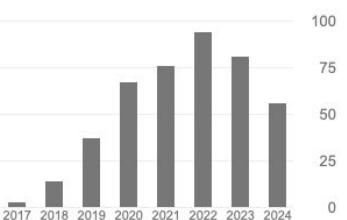
SIGN IN

Google Scholar

GET MY OWN PROFILE

Cited by

	All	Since 2019
Citations	429	411
h-index	9	9
i10-index	8	8



Public access [VIEW ALL](#)

0 articles [4 articles](#)

not available [available](#)

Based on funding mandates

TITLE	CITED BY	YEAR
Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields T Panboonyuen, K Jitkajornwanich, S Lawawirojwong, P Srestasathien, ... Remote Sensing 9 (7), 680	137	2017
Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning T Panboonyuen, K Jitkajornwanich, S Lawawirojwong, P Srestasathien, ... Remote Sensing 11 (1), 83	107	2019
An enhanced deep convolutional encoder-decoder network for road segmentation on aerial imagery T Panboonyuen, P Vateekul, K Jitkajornwanich, S Lawawirojwong Recent Advances in Information and Communication Technology 2017 ...	47	2018
Transformer-based decoder designs for semantic segmentation on remotely sensed images T Panboonyuen, K Jitkajornwanich, S Lawawirojwong, P Srestasathien, ... Remote Sensing 13 (24), 5100	45	2021
Object detection of road assets using transformer-based YOLOX with feature pyramid decoder on thai highway panorama T Panboonyuen, S Thongbai, W Wongweeranimit, P Santiamnont, ... Information 13 (1), 5	20	2021
Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network I Wichakam, T Panboonyuen, C Udomcharoenchaikit, P Vateekul MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok ...	18	2018
Semantic segmentation on medium-resolution satellite images using deep convolutional networks with remote sensing derived indices S Chantharaj, K Pornratthanapong, P Chitsinpchayakun, T Panboonyuen, ... 2018 15th International joint conference on computer science and software ...	13	2018

Featured Publications

≡ Google Scholar

Q SIGN IN

Teerapong Panboonyuen

Senior Research Scientist at MARS, Post-doc at Chula
Verified email at chula.ac.th - Homepage

Artificial Intelligence Machine Learning Deep Learning Computer Vision Remote Sensing

[FOLLOW](#)

GET MY OWN PROFILE

Cited by

All Since 2019

Citations 429 411
h-index 9 9
i10-index 8 8

TITLE CITED BY YEAR

Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields

137 2017

T Panboonyuen, K Jitjajawichit, S Lawawirojwong, P Srestasathien, ...

Remote Sensing 9 (7), 680

Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning

107 2019

T Panboonyuen, K Jitjajawichit, S Lawawirojwong, P Srestasathien, ...

Remote Sensing 11 (1), 83

An enhanced deep convolutional encoder-decoder network for road segmentation on aerial imagery

47 2018

T Panboonyuen, P Vateekul, K Jitjajawichit, S Lawawirojwong

Recent Advances in Information and Communication Technology 2017 ...

Transformer-based decoder designs for semantic segmentation on remotely sensed images

45 2021

T Panboonyuen, K Jitjajawichit, S Lawawirojwong, P Srestasathien, ...

Remote Sensing 13 (24), 5100

Object detection of road assets using transformer-based YOLOX with feature pyramid decoder for 3D highway panaroma

20 2021

T Panboonyuen, S Sirinaporn, W Wongweeraynit, P Sentiamont, ...

Information 13 (1), 5

Real-time polyp segmentation for colonoscopy video frames using compressed fully convolutional network

18 2018

I Wichiansri, T Panboonyuen, C Udomcharoenchait, P Vateekul

MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok ...

Semantic segmentation on medium-resolution satellite images using deep convolutional networks with remote sensing derived indices

13 2018

S Chanthai, K Ponnattanapong, P Chinsuphayakun, T Panboonyuen, ...

2019 15th International joint conference on computer science and software ...

Public access

VIEW ALL

0 articles 4 articles

not available available

Based on funding mandates

What Is Computer Vision? [Basic Tasks & Techniques]

Semantic
Segmentation

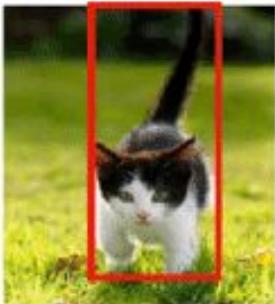


CAT GRASS
TREE

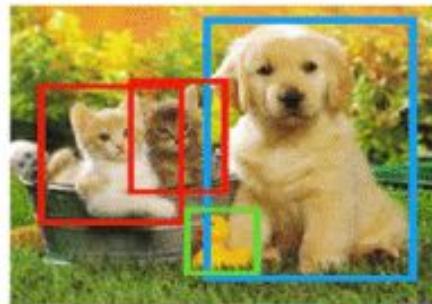
Classification



Classification
+ localization



Object detection



Instance
segmentation



No object
Just pixels

Single object

Multiple objects

CAT

CAT

CAT DOG DUCK

CAT CAT DOG DUCK

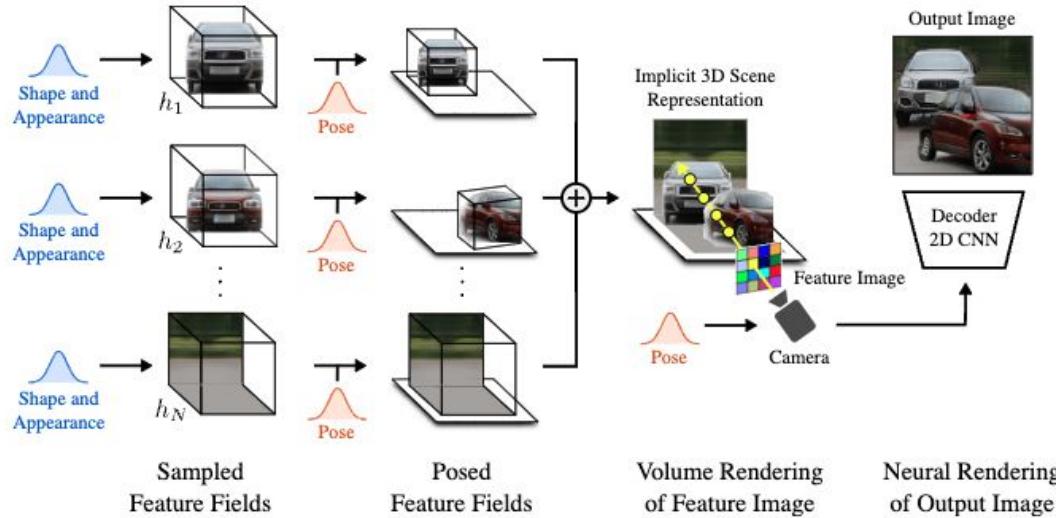
GIRAFFE

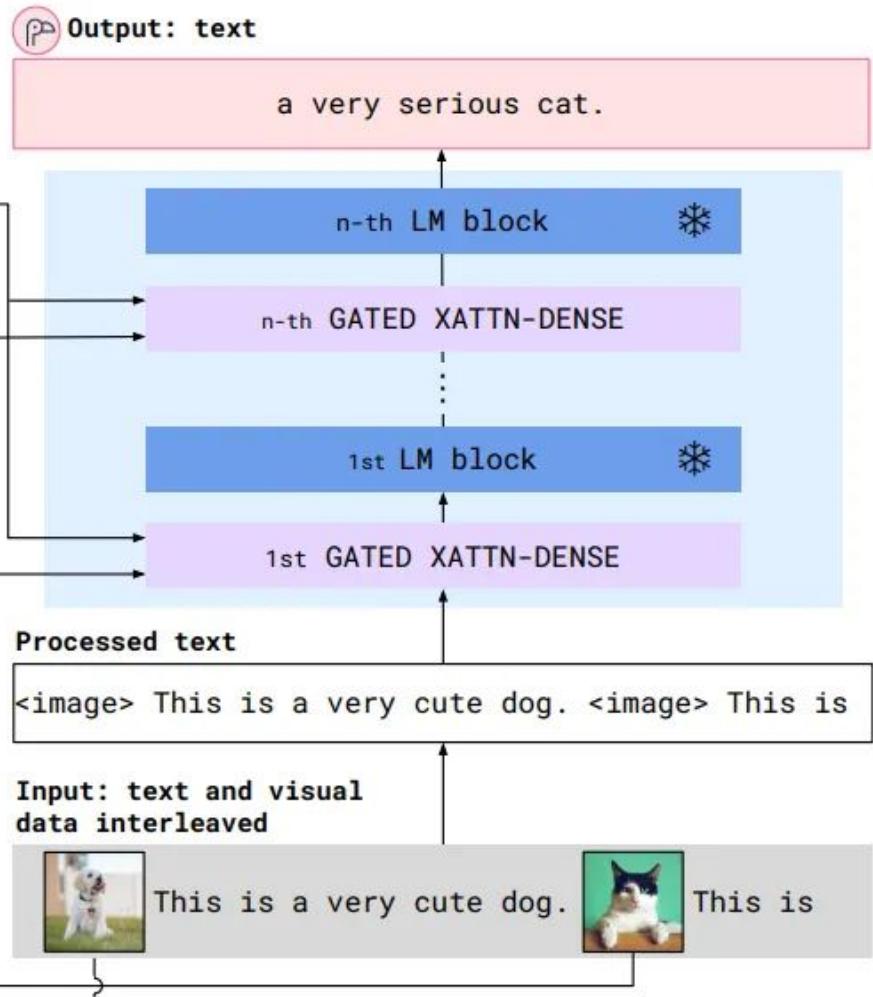
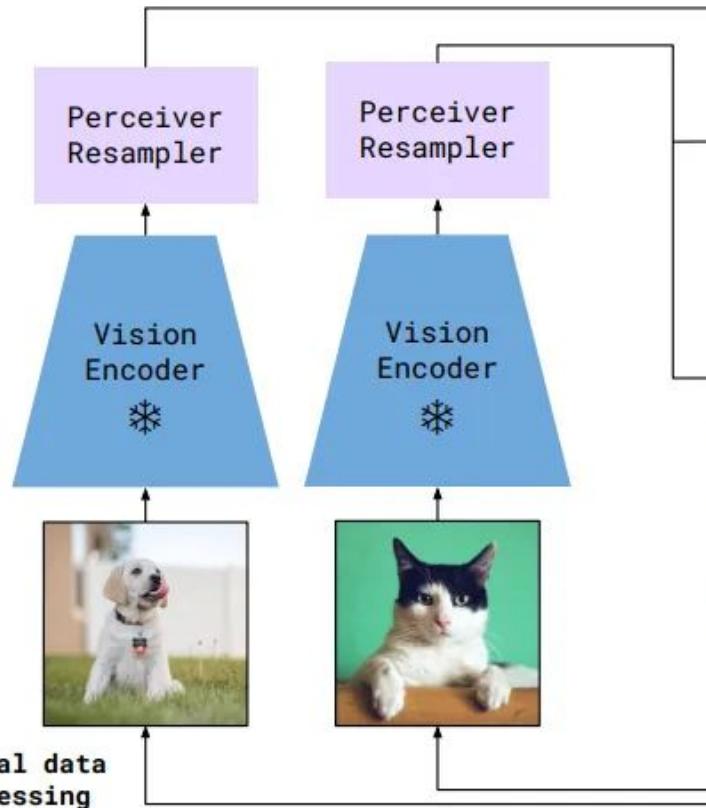
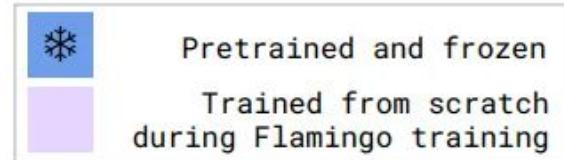
Representing Scenes as Compositional Generative Neural Feature Fields

Michael Niemeyer Andreas Geiger

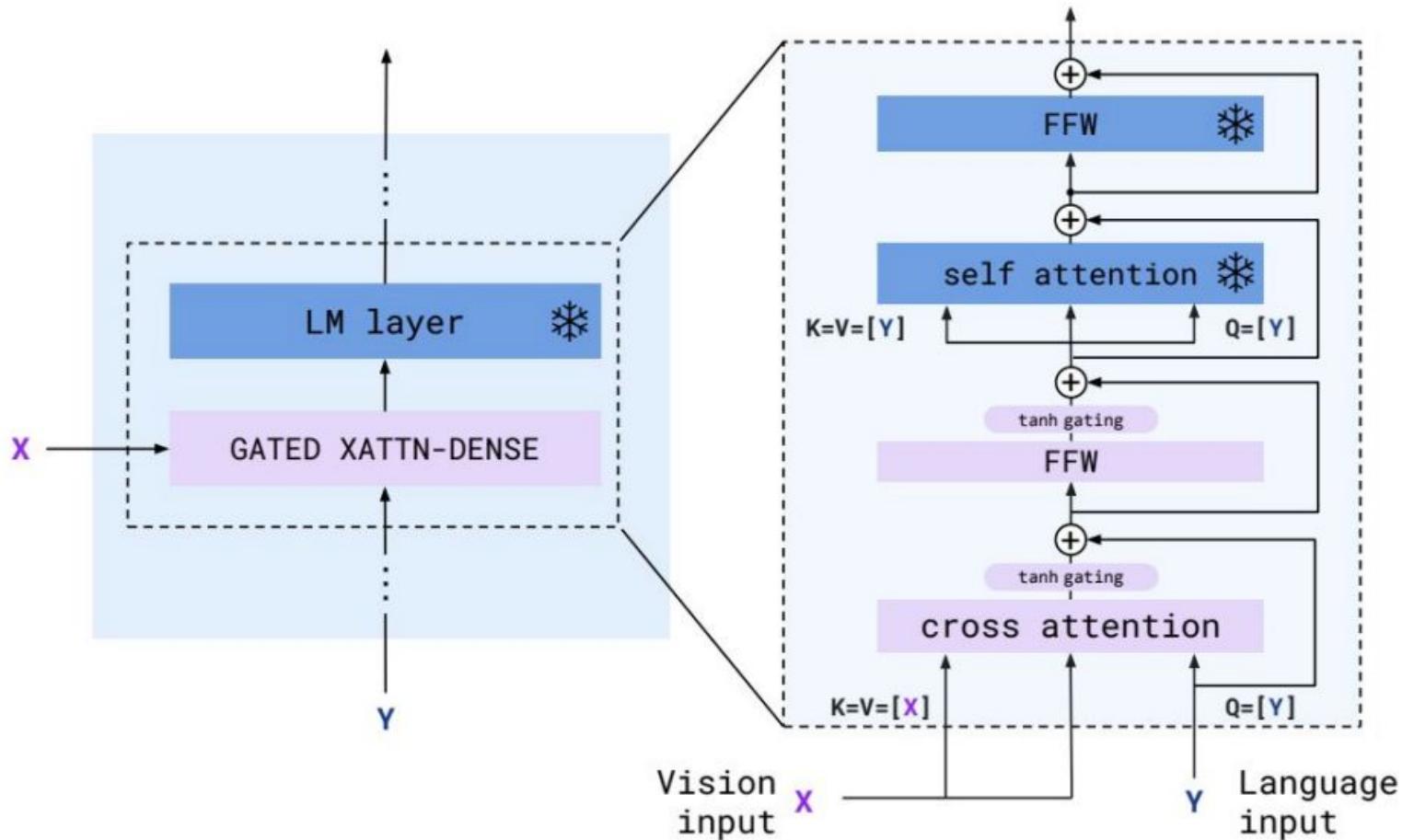
Max Planck Institute for Intelligent Systems and University of Tübingen

CVPR 2021 (oral, **best paper award**)





Gated XATTN-Dense layers



Understanding Multimodal



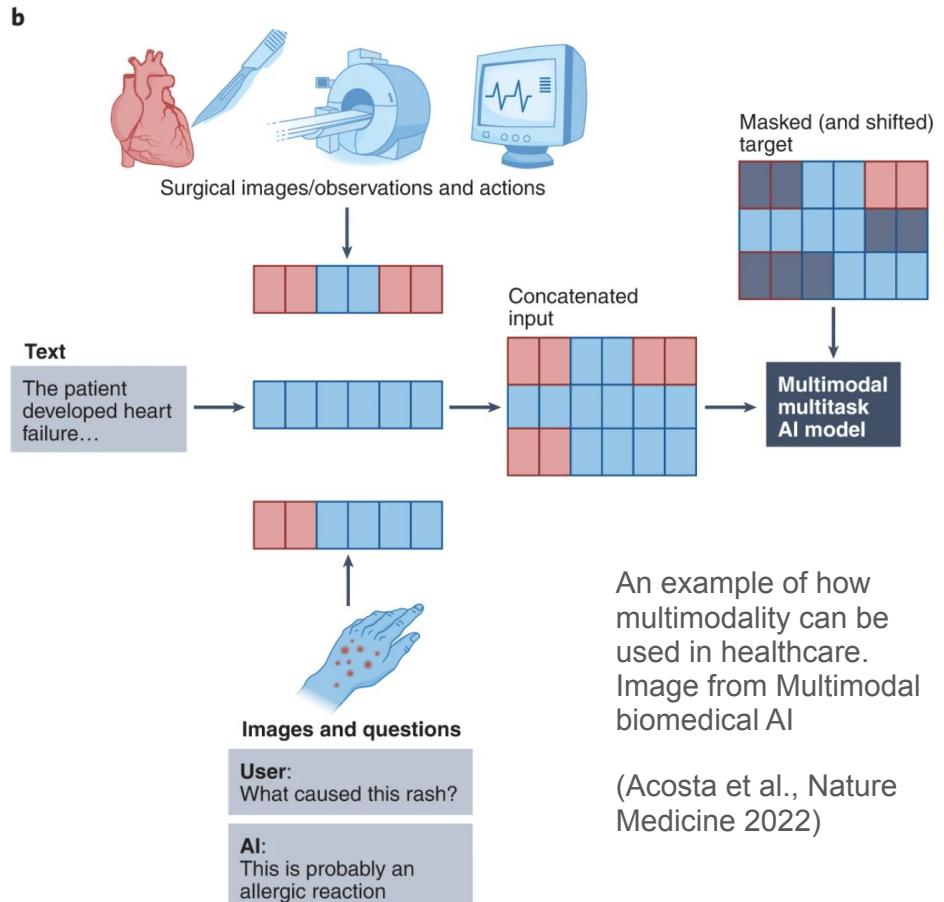
Chip Huyen ✅
@chipro

New blog post: Multimodality and Large Multimodal Models (LMMs)

Being able to work with data of different modalities -- e.g. text, images, videos, audio, etc. -- is essential for AI to operate in the real world.

This post covers multimodal systems in general, including Large Multimodal Models. It consists of 3 parts.

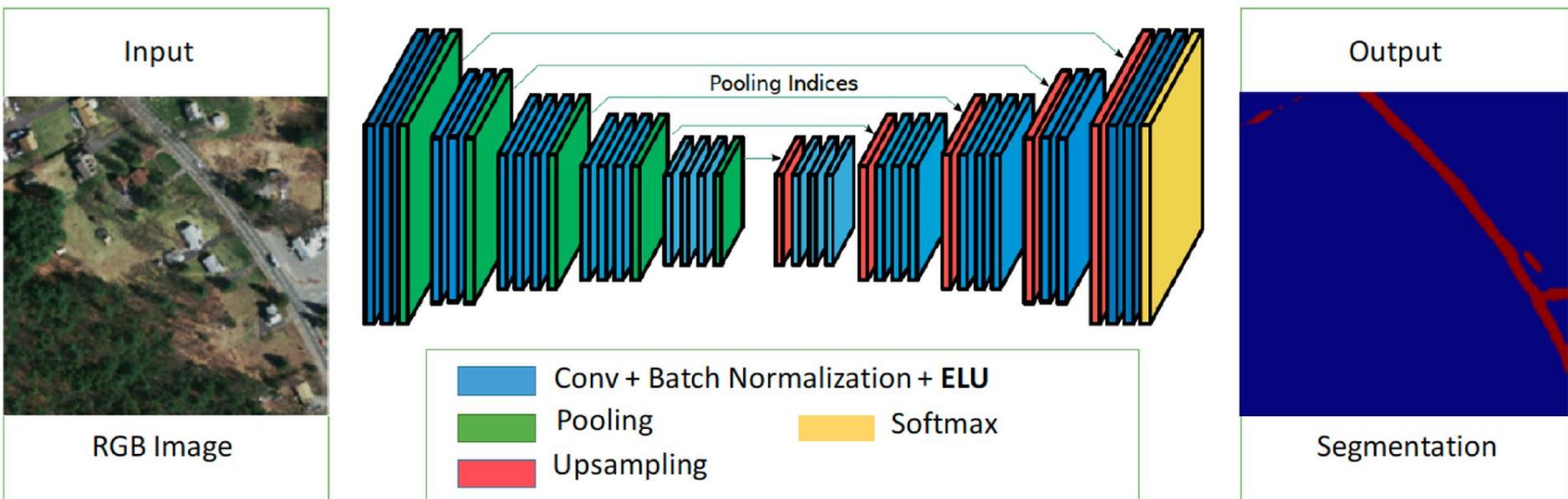
- * Part 1 covers the context for multimodality.
- * Part 2 discusses how to train a multimodal system, using the architectures of CLIP and Flamingo, and examples from GPT-4V.
- * Part 3 discusses some active research areas for LMMs, including generating multimodal outputs.



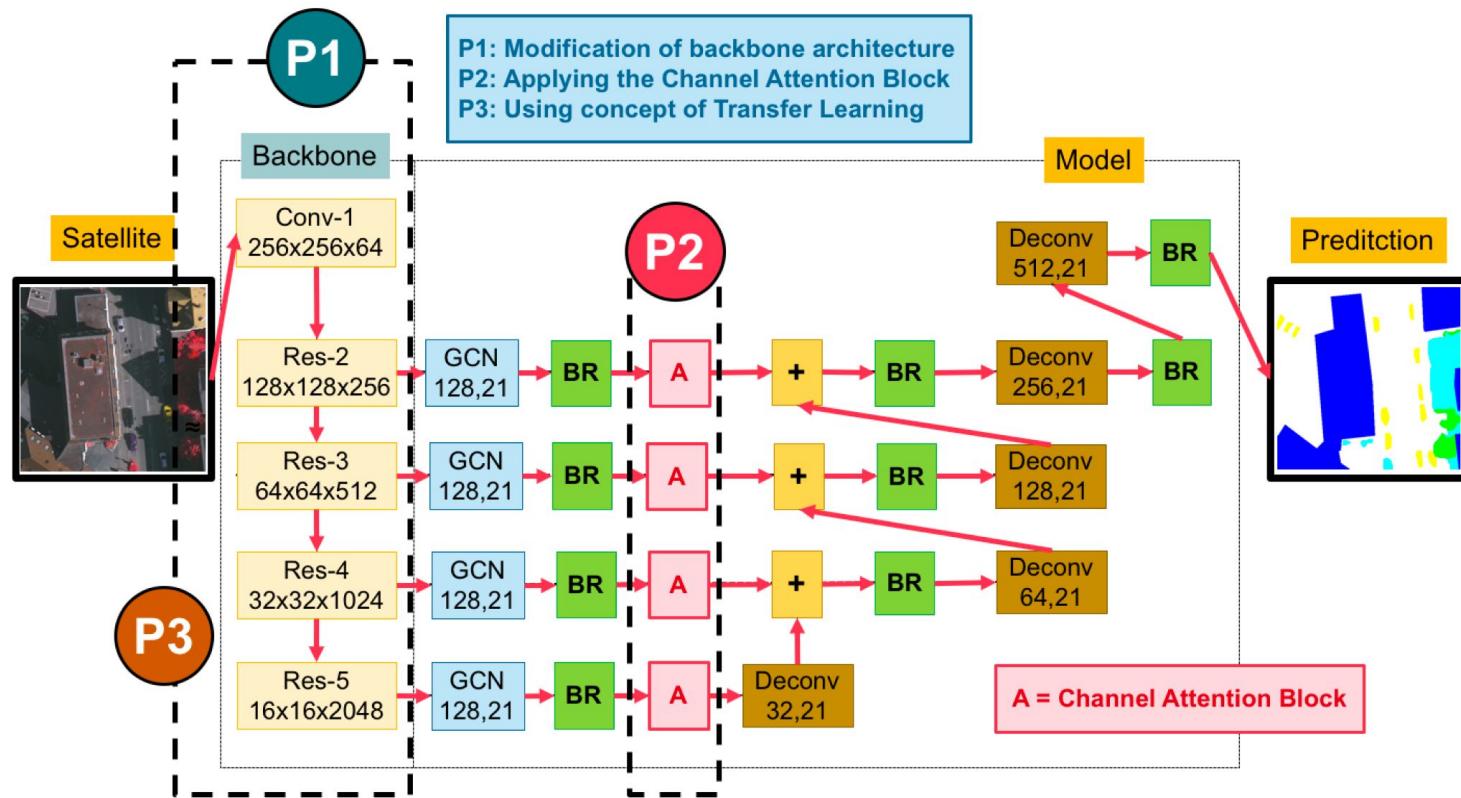
<https://huyenchip.com/2023/10/10/multimodal.htm>

<https://twitter.com/chipro/status/1711970025874321479/photo/1>

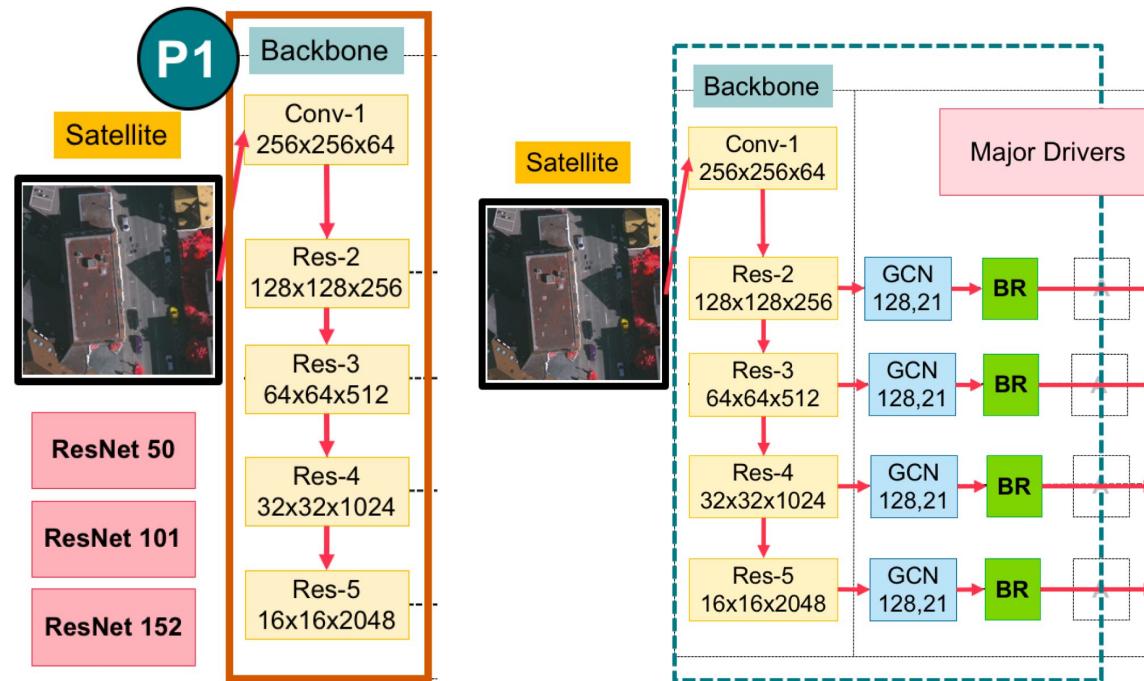
[1] Panboonyuen, Teerapong, et al. "Road segmentation of remotely-sensed images using deep convolutional neural networks with landscape metrics and conditional random fields." *Remote Sensing* 9.7 (2017): 680.



[2] Panboonyuen, Teerapong, et al. "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning." Remote Sensing 11.1 (2019): 83.

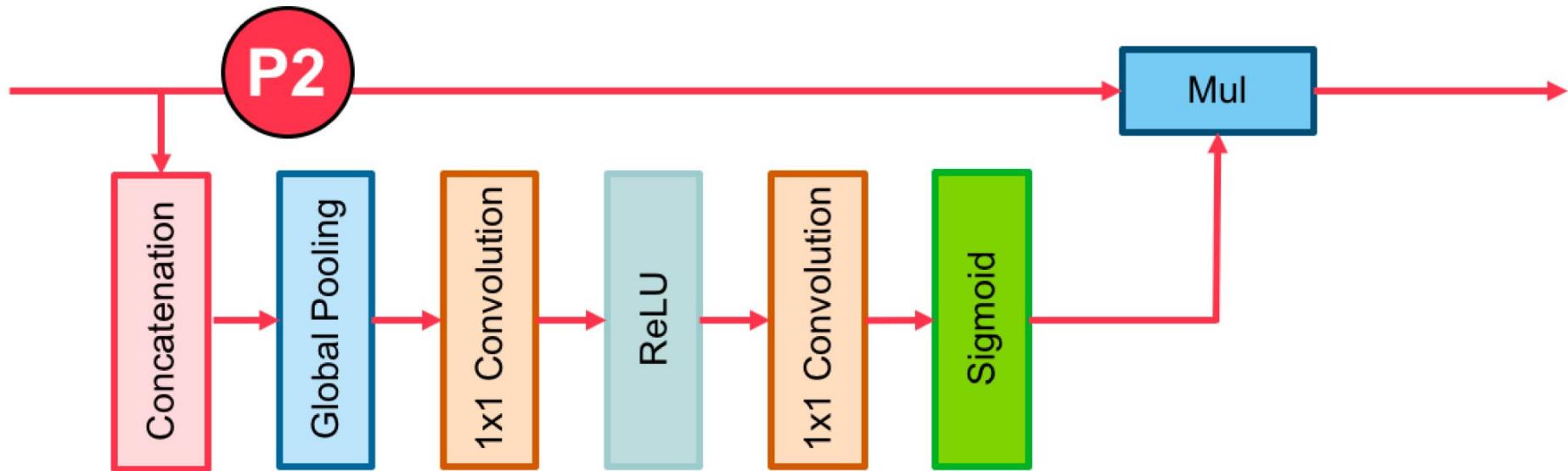


[2] Panboonyuen, Teerapong, et al. "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning." Remote Sensing 11.1 (2019): 83.



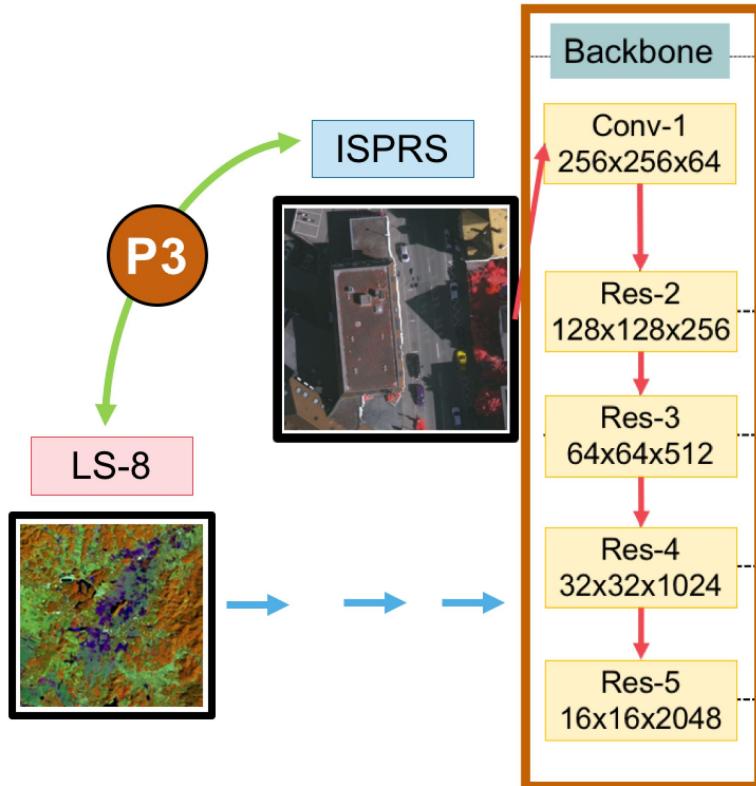
An overview of the whole backbone pipeline in (left) the main backbone with varying by ResNet50, ResNet 101, and ResNet 152; (right) the major drivers of my main classification network (composed of a global convolutional network (GCN) and a boundary refinement (BR) block).

[2] Panboonyuen, Teerapong, et al. "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning." Remote Sensing 11.1 (2019): 83.



- Components of the channel attention block.
- The red lines represent the downsample operators, respectively.
- The red line cannot change the size of feature maps.
- It is only a path for information passing.

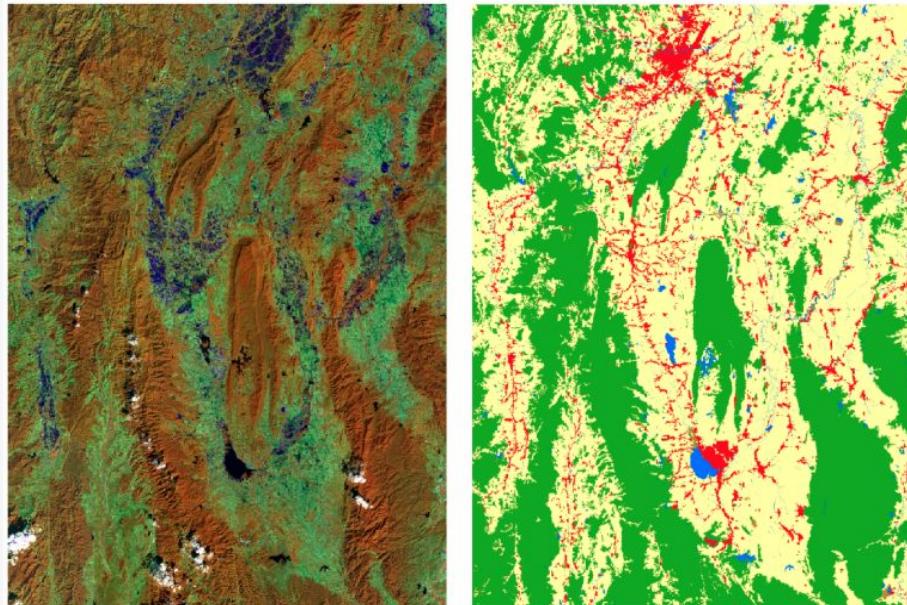
[2] Panboonyuen, Teerapong, et al. "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning." Remote Sensing 11.1 (2019): 83.



- The domain-specific transfer learning strategy reuses pre-trained weights of models between two datasets—very high (ISPRS) and medium (Landsat-8; LS-8) resolution images.

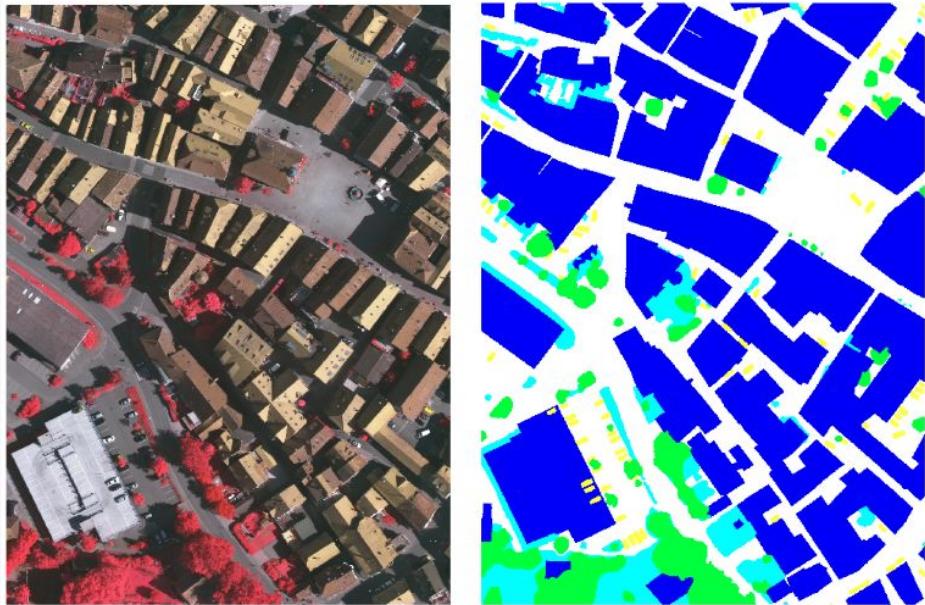
[2] Panboonyuen, Teerapong, et al. "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning." Remote Sensing 11.1 (2019): 83.

Figure 6. Sample satellite images from Nan, a province in Thailand (**left**), and corresponding ground truth (**right**). The label of medium resolution dataset includes five categories: agriculture (yellow), forest (green), miscellaneous (brown), urban (red), and water (blue).



X

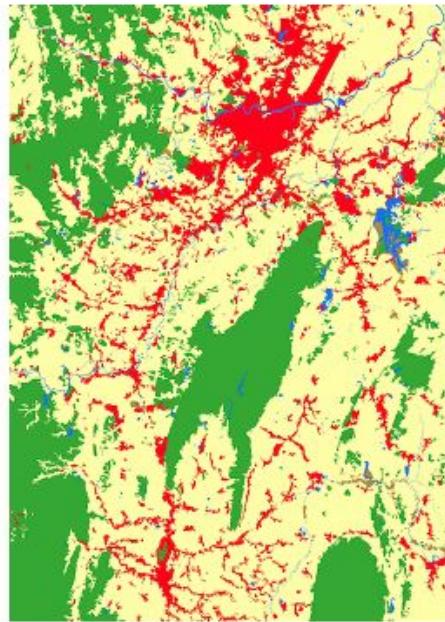
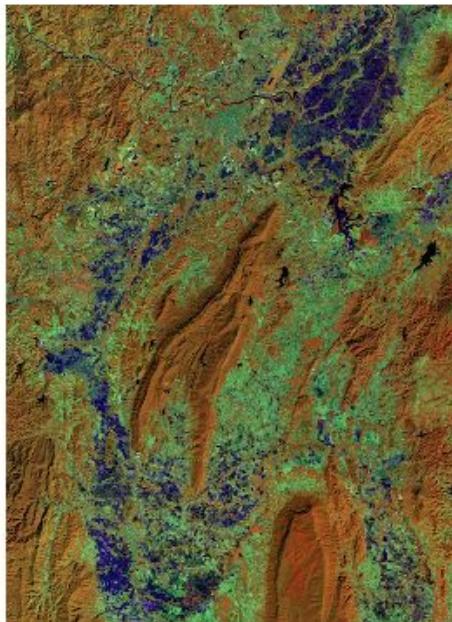
Figure 8. The sample input tile from **Figure 7 (left)** and corresponding ground truth (**right**). The label of the Vaihingen Challenge includes six categories: impervious surface (imp surf, white), building (blue), low vegetation (low veg, cyan), tree (green), car (yellow), and clutter/background (red).



X

Public and Private Corpora

Private corpus (GISTDA Nan Province Corpus)

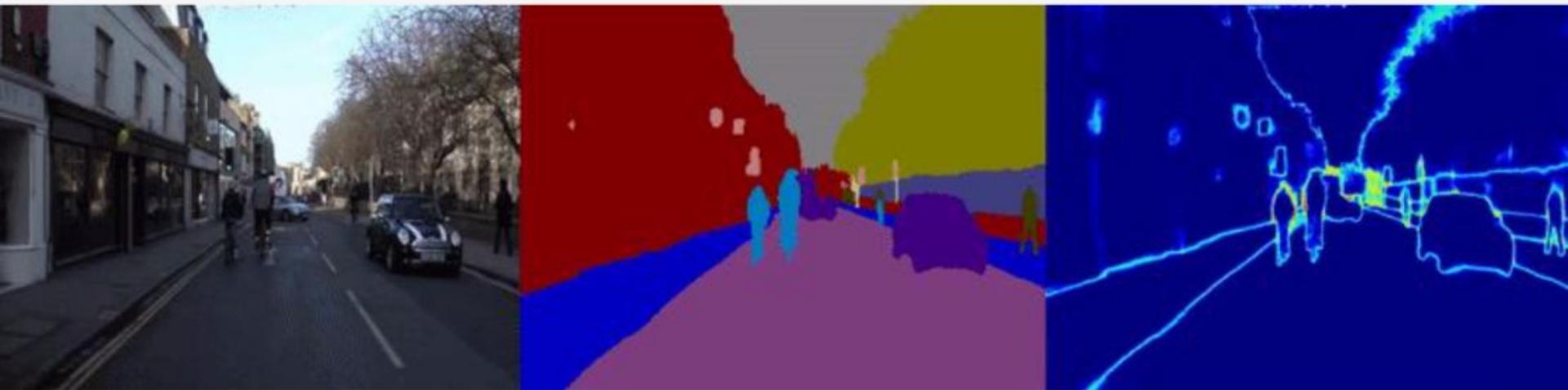


Color	Class
Yellow	Agriculture
Green	Forest
Brown	Miscellaneous
Red	Urban
Blue	Water

Predictions from Tiramisu on CamVid video stream.



Bayesian SegNet for probabilistic scene understanding



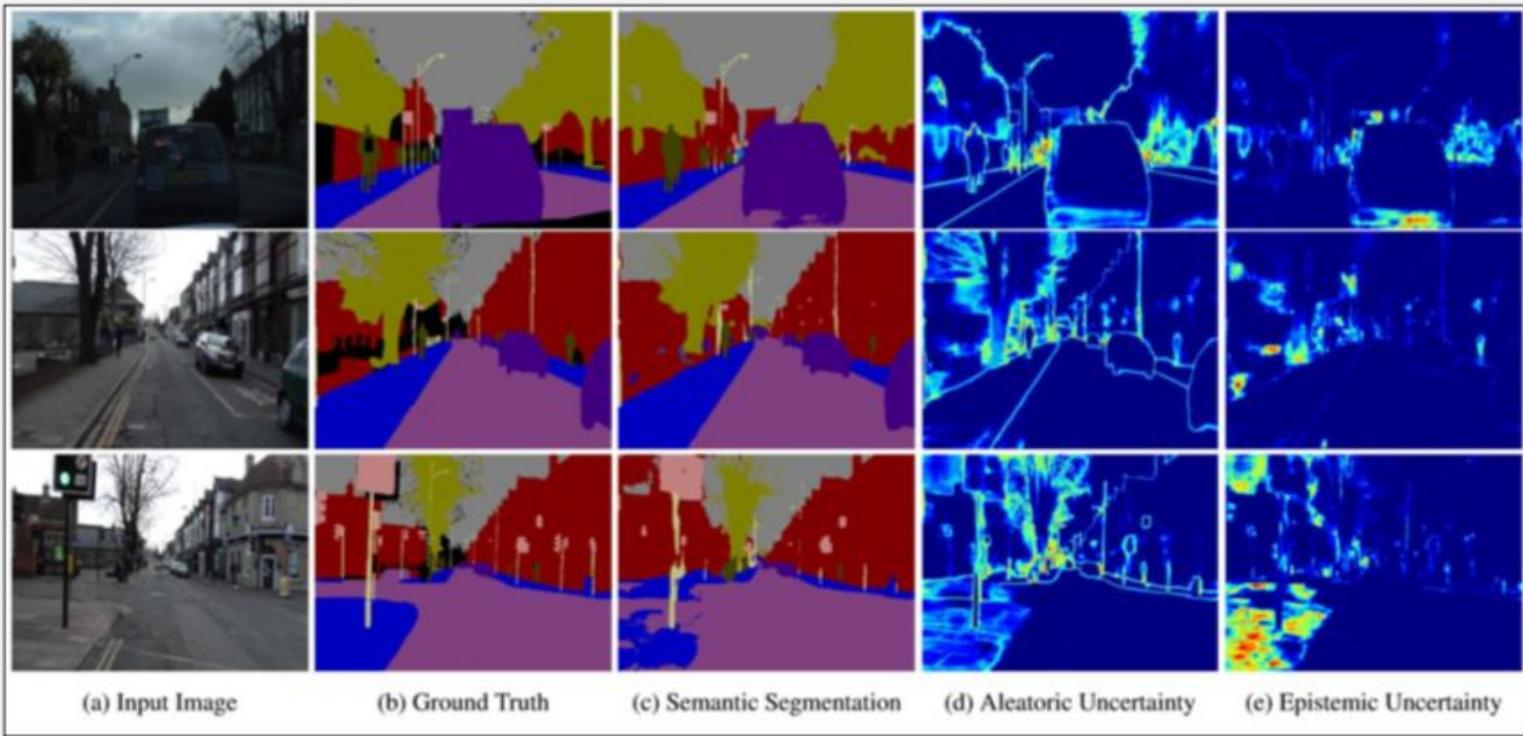
Input Image

Semantic Segmentation

Uncertainty

What kind of uncertainty can we model?

Epistemic uncertainty is *modeling uncertainty*
Aleatoric uncertainty is *sensing uncertainty*



Modeling Uncertainty with Bayesian Deep Learning



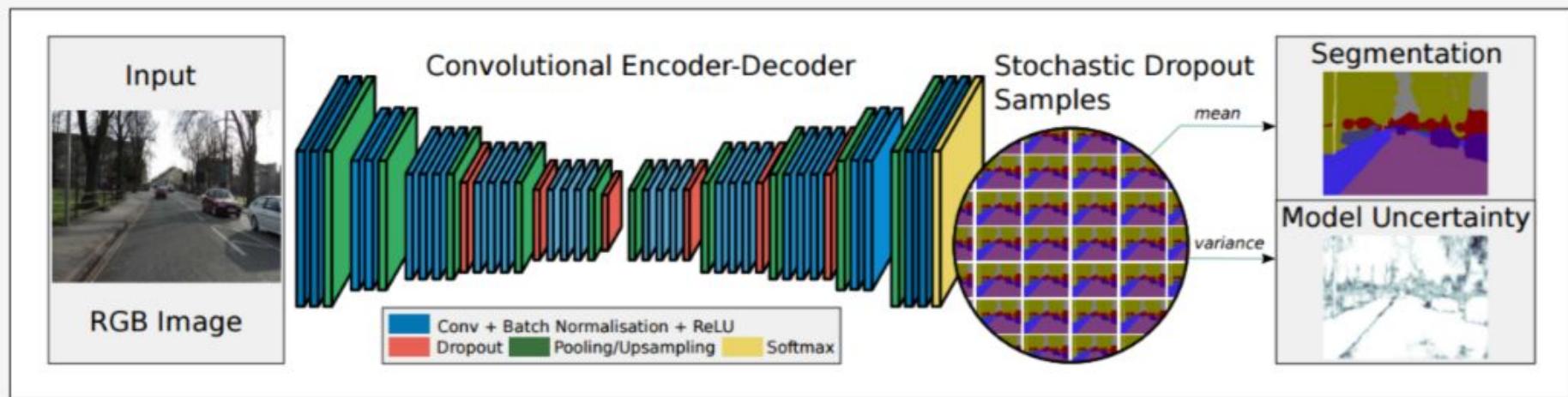
Deep learning is required to achieve state of the art results in computer vision applications but doesn't provide uncertainty estimates.

- **Bayesian neural networks** are a framework for understanding uncertainty in deep learning
- They have **distributions over network parameters** (rather than deterministic weights)
- Traditionally they have been **tricky to scale**

Modeling Epistemic Uncertainty with Bayesian Deep Learning

We can **model epistemic uncertainty** in deep learning models using Monte Carlo **dropout sampling** at test time.

Dropout sampling can be interpreted as **sampling from a distribution over models**.



Modeling Aleatoric Uncertainty with Probabilistic Deep Learning

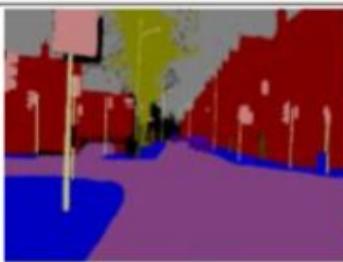
	Deep Learning	Probabilistic Deep Learning
Model	$[\hat{y}] = f(x)$	$[\hat{y}, \hat{\sigma}^2] = f(x)$
Regression	$Loss = \ y - \hat{y}\ ^2$	$Loss = \frac{\ y - \hat{y}\ ^2}{2\hat{\sigma}^2} + \log \hat{\sigma}^2$
Classification	$Loss = SoftmaxCrossEntropy(\hat{y}_t)$	$\hat{y}_t = \hat{y} + \epsilon_t \quad \epsilon_t \sim N(0, \hat{\sigma}^2)$ $Loss = \frac{1}{T} \sum_t SoftmaxCrossEntropy(\hat{y}_t)$

Semantic Segmentation Performance on CamVid

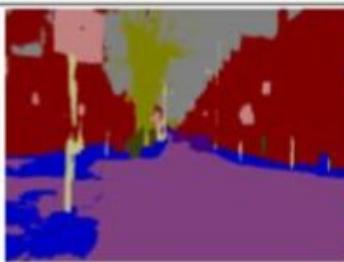
CamVid Results	IoU Accuracy
DenseNet (State of the art baseline)	67.1
+ Aleatoric Uncertainty	67.4
+ Epistemic Uncertainty	67.2
+ Aleatoric & Epistemic	67.5



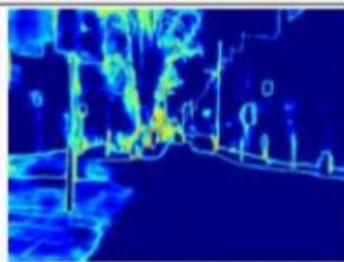
(a) Input Image



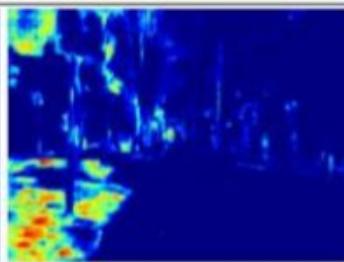
(b) Ground Truth



(c) Semantic Segmentation

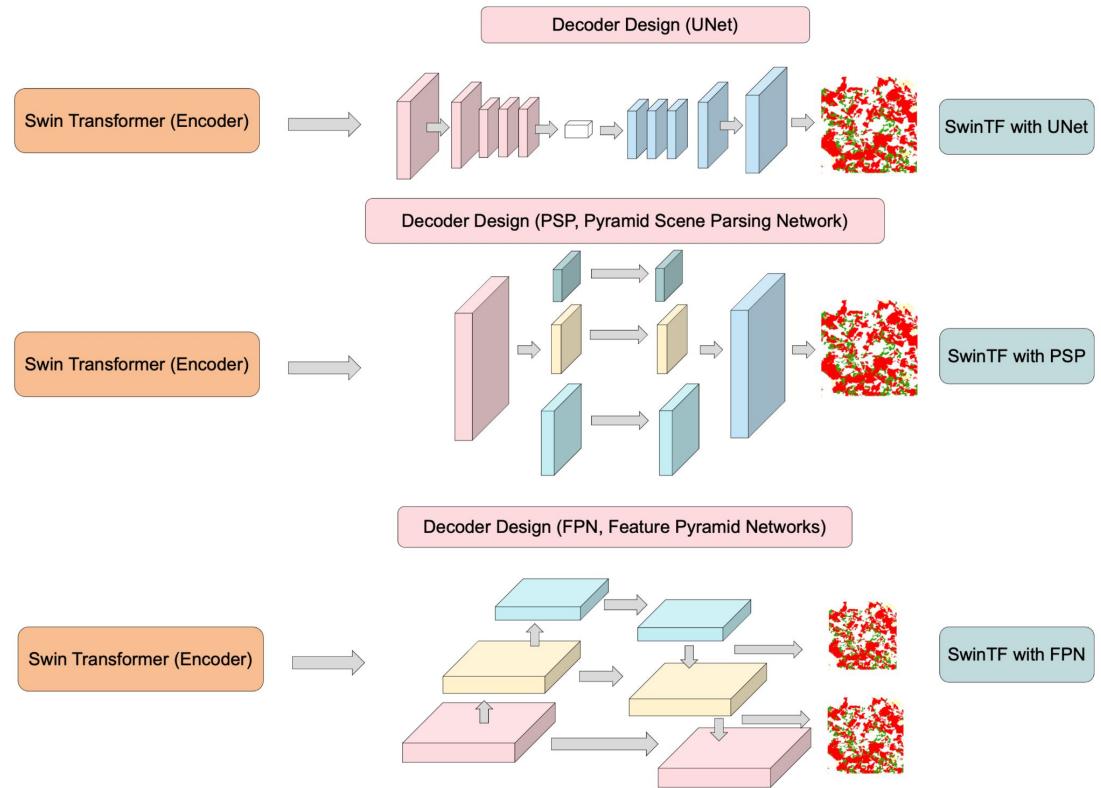
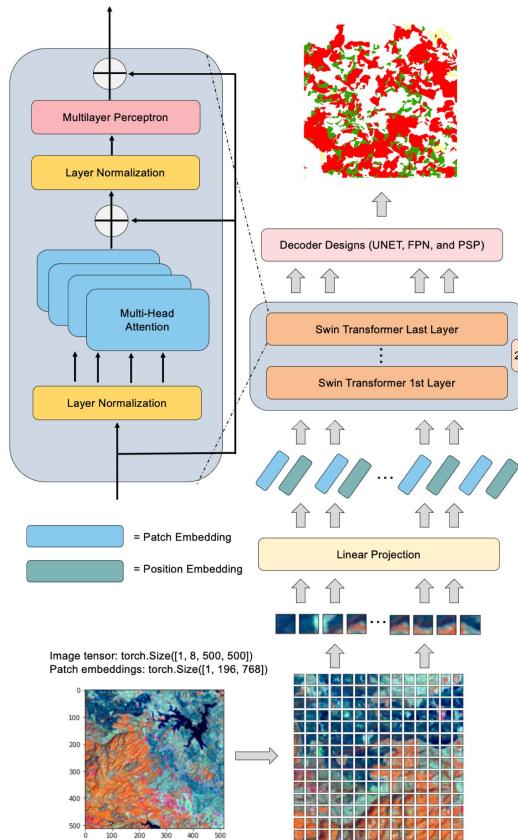


(d) Aleatoric Uncertainty



(e) Epistemic Uncertainty

[3] Panboonyuen, Teerapong, et al. "Transformer-based decoder designs for semantic segmentation on remotely sensed images." Remote Sensing 13.24 (2021): 5100.



[4] Wichakam, I., Panboonyuen, T., Udomcharoenchaikit, C., & Vateekul, P. (2018). **Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network**. In MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24 (pp. 393-404). Springer International Publishing.

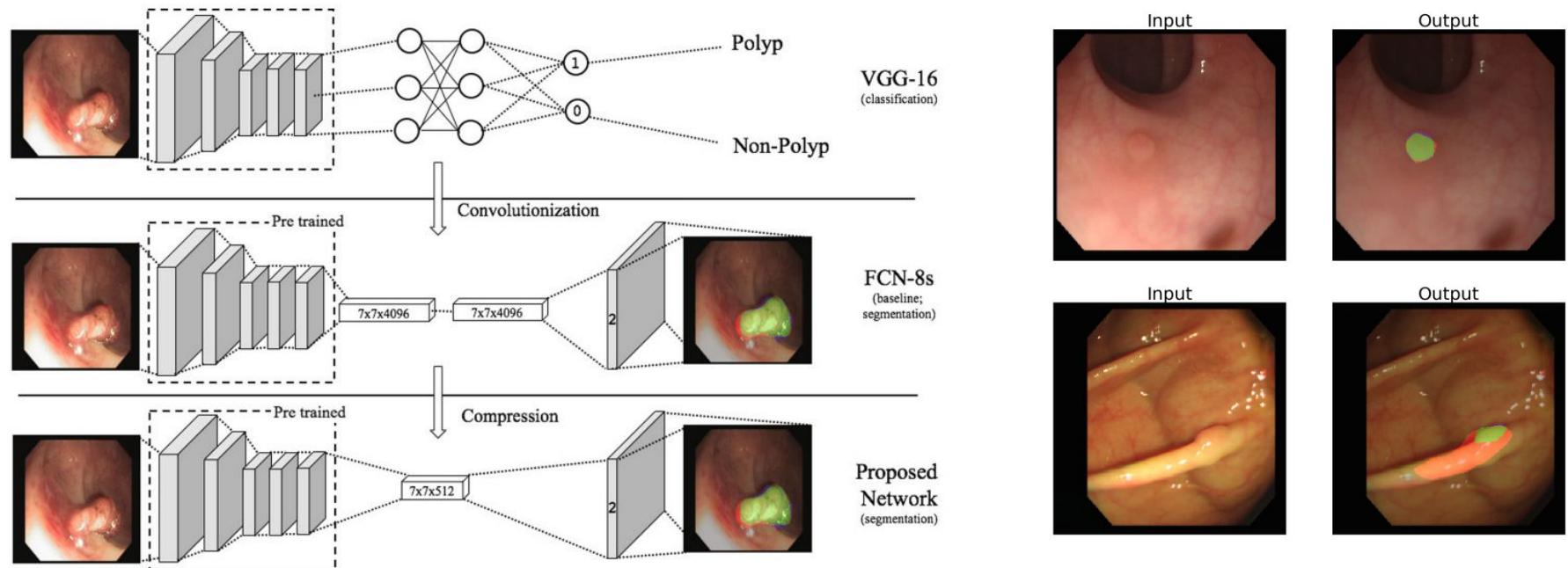
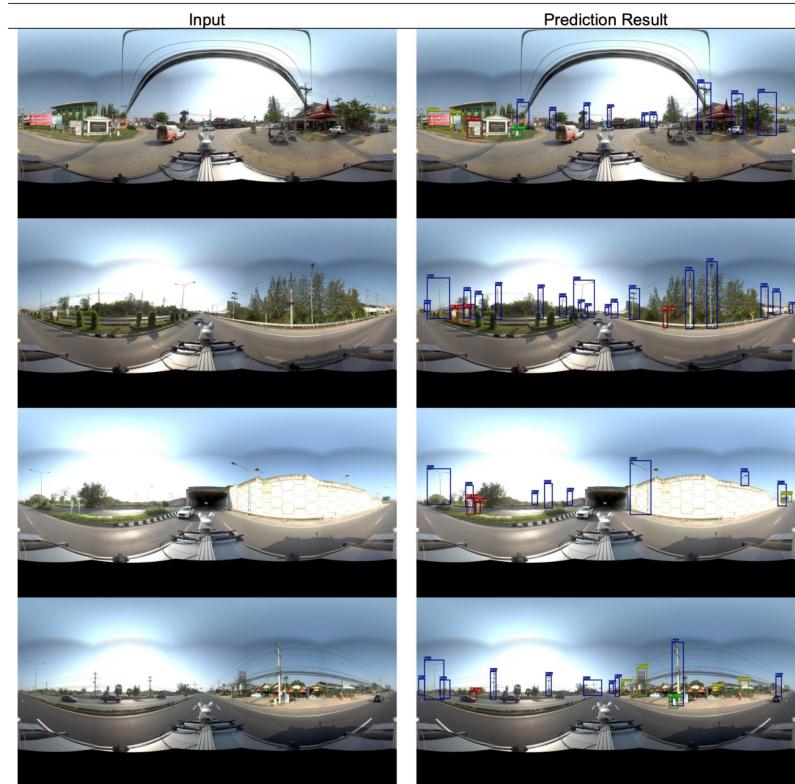
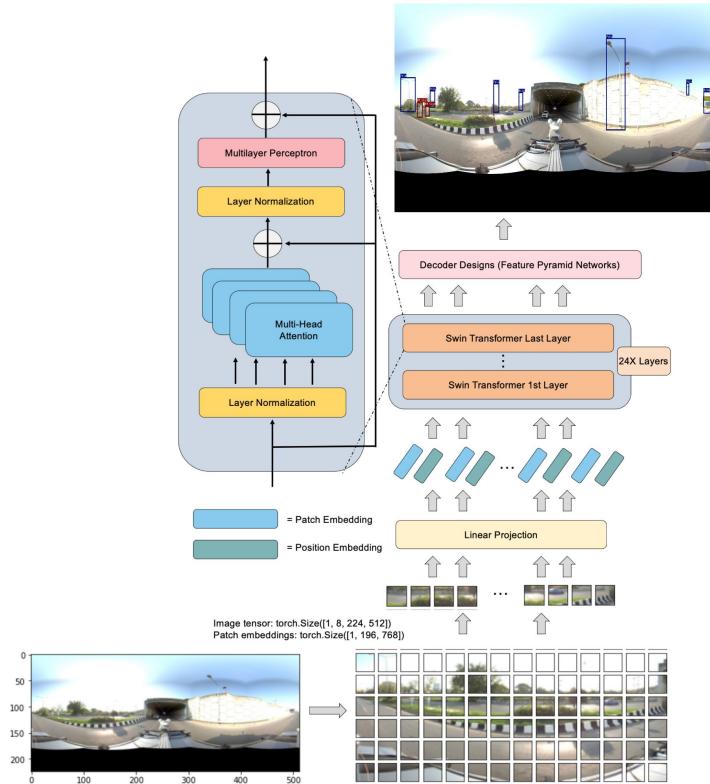


Fig. 1. Overview of our compressed network which is compressed from the original FCN-8s [8] based on VGG-16 [4] architecture.

[5] Panboonyuen, Teerapong, et al. "Object detection of road assets using transformer-based YOLOX with feature pyramid decoder on thai highway panorama." Information 13.1 (2022): 5.



[6] Thitisiriwech, K., Panboonyuen, T., Kantavat, P., Iwahori, Y., & Kijsirikul, B. (2022). The Bangkok Urbanscapes Dataset for Semantic Urban Scene Understanding Using Enhanced Encoder-Decoder With Atrous Depthwise Separable A1 Convolutional Neural Networks. *IEEE Access*, 10, 59327-59349.

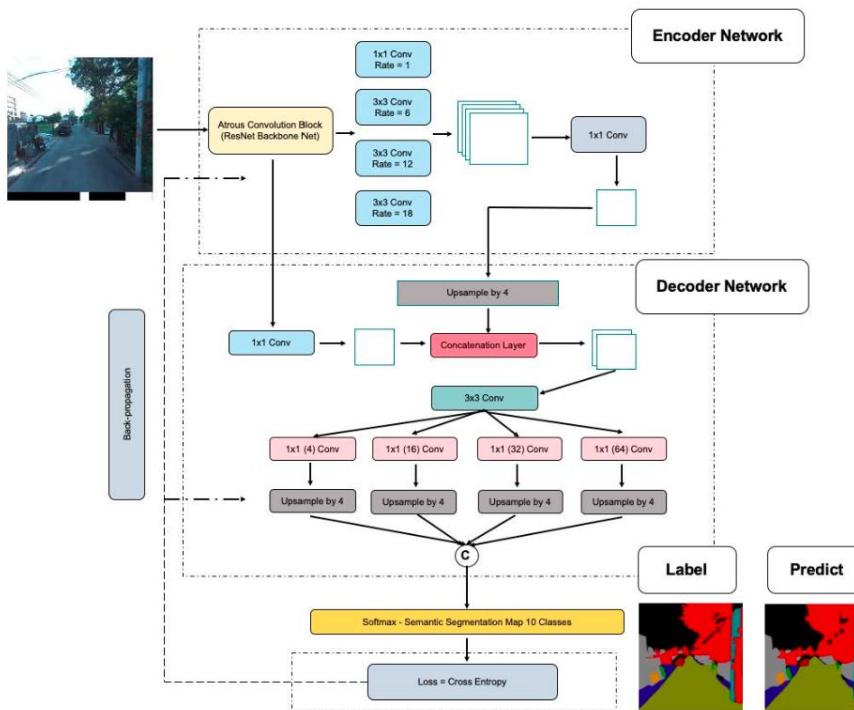


FIGURE 6. An overview of enhanced DeepLab-V3+ (Encoder-Decoder with atrous separable convolutional for semantic segmentation [15]) with ResNet-101 backbone [41] (DeepLab-V3-A1-ResNet-101).

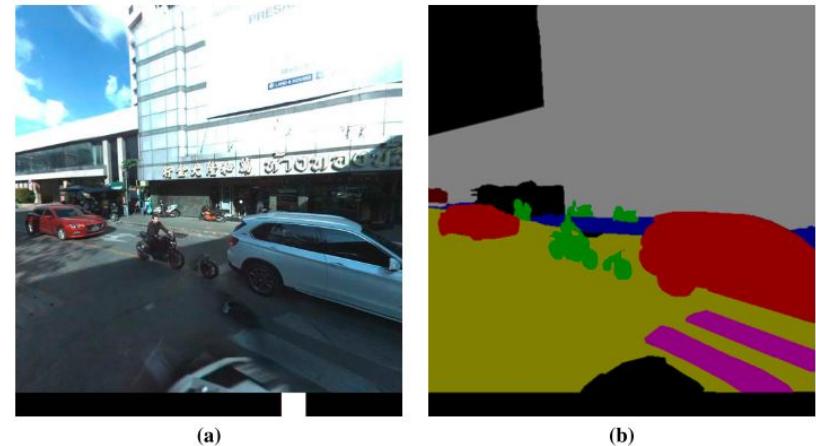


FIGURE 8. Sample 1: The example of Sukhumvit's large road from the training set of the Bangkok Urbanscapes dataset. The input image is shown in (a), and the ground truth is shown in (b).

Void	Building	Wall	Tree	VegetationMisc	Fence
Sidewalk	ParkingBlock	Column_Pole	TrafficConc	Bridge	SignSymbol
Misc_Text	TrafficLight	Sky	Tunnel	Archway	Road
RoadShoulder	LaneMkgsDriv	LaneMkgsNonDriv	Animal	Pedestrian	Child
CartLuggagePram	Bicyclist	MotorcycleScooter	Car	SUVPickupTruck	Truck_Bus
Train	OtherMoving				

FIGURE 3. The semantic color codes of the CamVid dataset. Each color is encoded with respect to the semantic class in the ground truth images.

[6] Vajeethaveesin, T., Panboonyuen, T., Lawawironjwong, S., Srestasathiern, P., Jaiyen, S., & Jitkajornwanich, K. (2022). **A performance comparison between GIS-based and neuron network methods for flood susceptibility assessment in ayutthaya province.** Trends in Sciences, 19(2), 2038-2038.

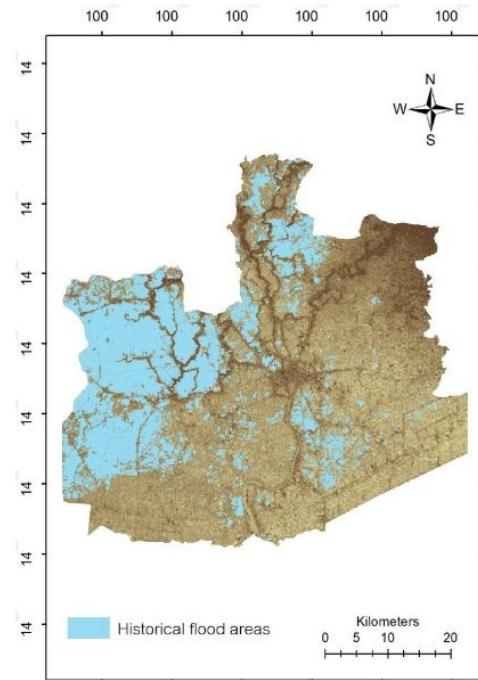


Figure 2 The historical flood zone shows flood events, which were captured in 2016 from GISTDA. Most of the flooding areas are located in the west of the study area.

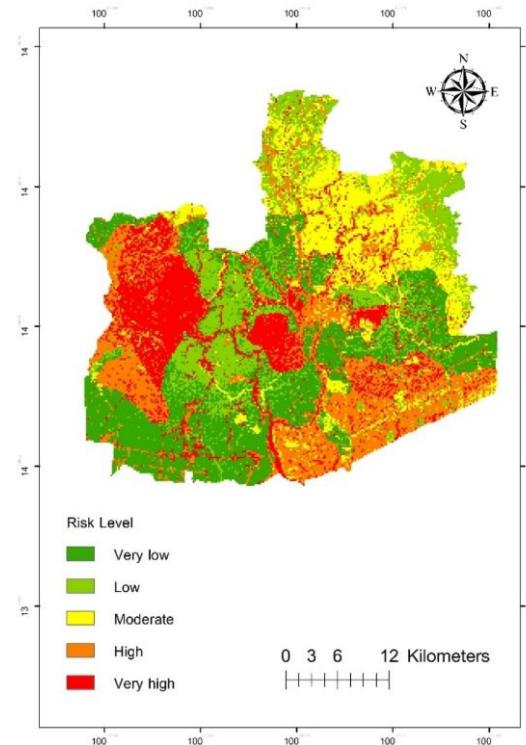


Figure 4 Flood susceptibility map from the flood risk assessment model (FRAM).

Classification of Primary Angle Closure Glaucoma (PACG) Using Deep Convolutional Neural Networks

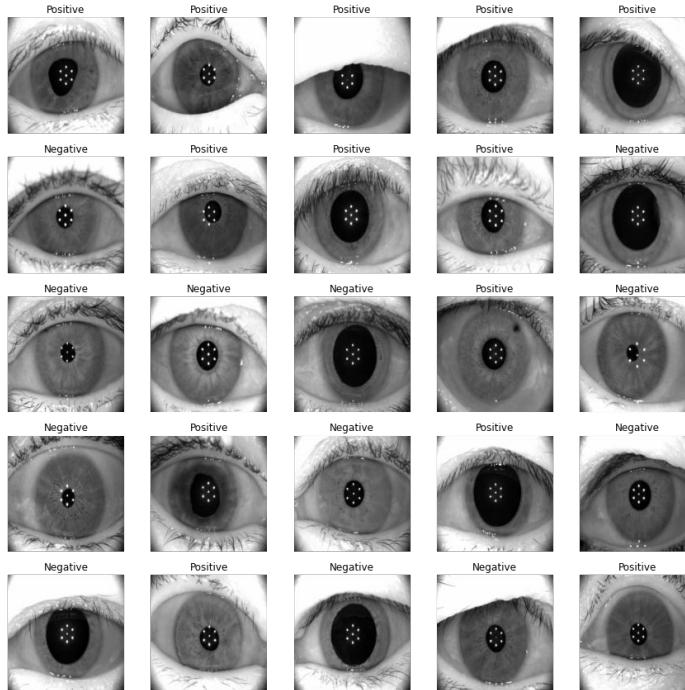


Table 4: Performance comparison on F1 score with existing models on the WTW-LAF23 dataset.

Model	F1 score			
	Mean	SD	Min	Max
VGG19 [Dutta et al., 2016]	61.34	4.12	54.67	65.76
DenseNet121 [Targ et al., 2016]	61.67	4.01	54.89	65.91
RestNet50 [Huang et al., 2017]	65.93	6.11	56.11	76.33
Transformer [Liu et al., 2021]	58.18	5.07	51.34	69.83
Our Proposed Method (SimCLR Chen et al. [2020])	67.82	N/A	N/A	N/A



DEPARTMENT OF MEDICAL SERVICES

Table 1: Details of used visual datasets.

Dataset	Type	Positive Size	Negative Size	Train Size	Val Size	# Classes
LAFx10_spss_Frequency2.3	Classification	612	612	1,101	123	2
AQD_Frequency2.4	Classification	3,338	3,153	5,841	650	2

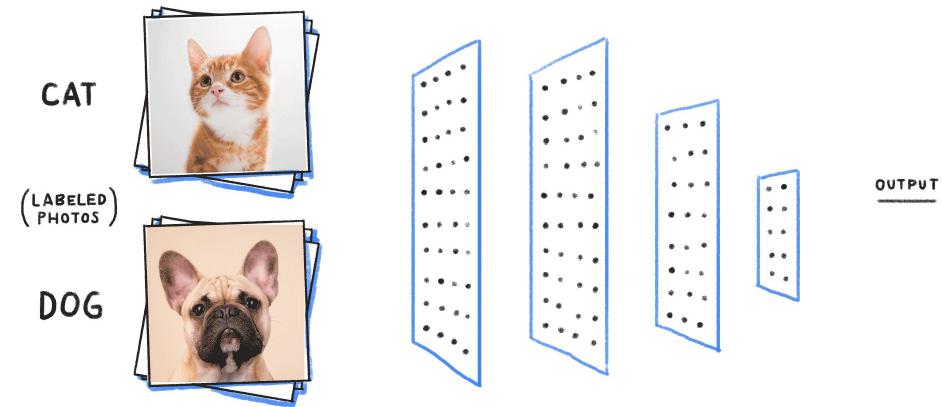
Table 2: Details of used visual datasets.

Dataset	Type	Female Size	Male Size	Unknown	Other	Train Size	Val Size	# Classes
GENDER	Classification	3,257	2,106	25	1	3,754	804	2

Table 3: Details of used visual datasets.

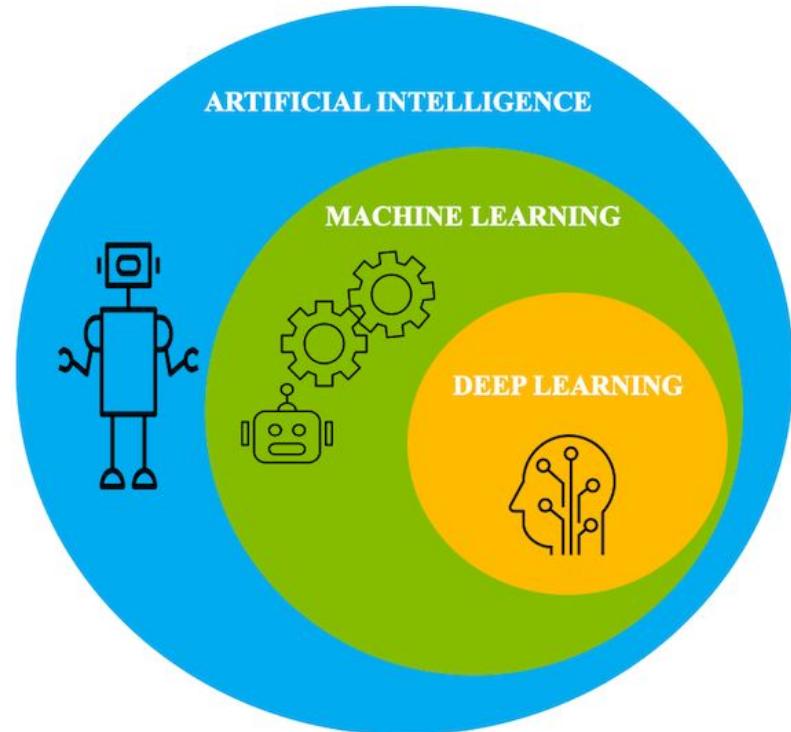
Dataset	Type	Young Adults	Middle-Aged	Old Adults	Train Size	Val Size	# Classes
AGE	Classification	224	3,885	1,271	3,765	807	3

Basic AI



AI vs Machine Learning vs Deep Learning

- **Artificial Intelligence (AI)** is the concept of creating smart intelligent machines.
- **Machine Learning (ML)** is a subset of artificial intelligence that helps you build AI-driven applications.
- **Deep Learning** is a subset of machine learning that uses vast volumes of data and complex algorithms to train a model.



(Recap): What Is Computer Vision? [Basic Tasks & Techniques]

Semantic
Segmentation



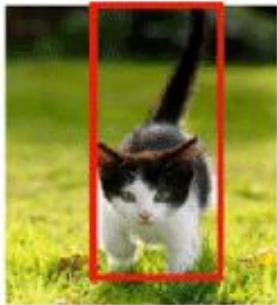
CAT GRASS
TREE

Classification



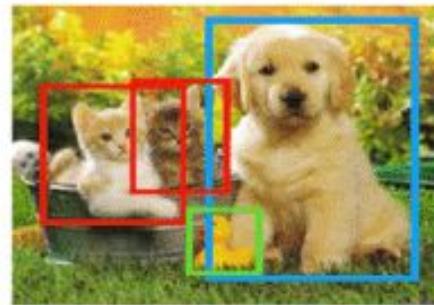
CAT

Classification
+ localization



CAT

Object detection



CAT DOG DUCK

Instance
segmentation



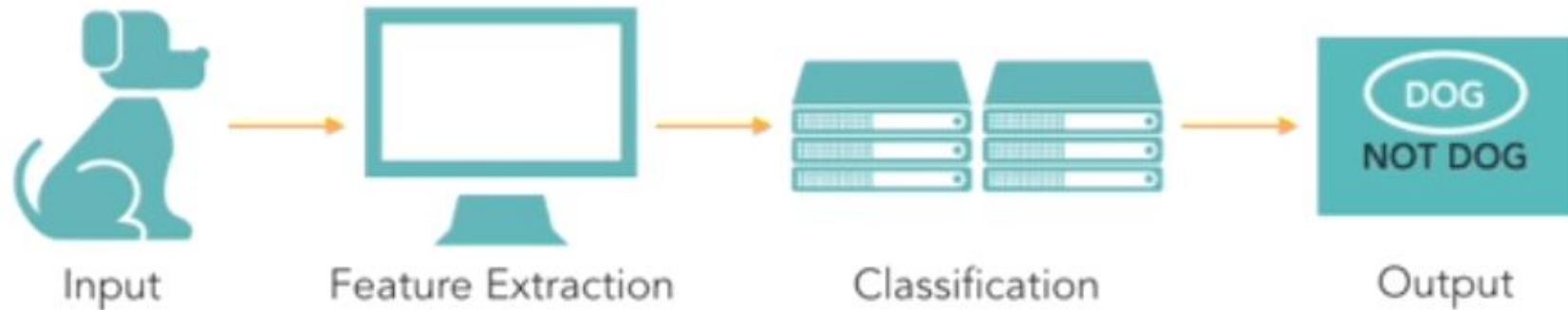
CAT CAT DOG DUCK

No object
Just pixels

Single object

Multiple objects

TRADITIONAL MACHINE LEARNING



DEEP LEARNING

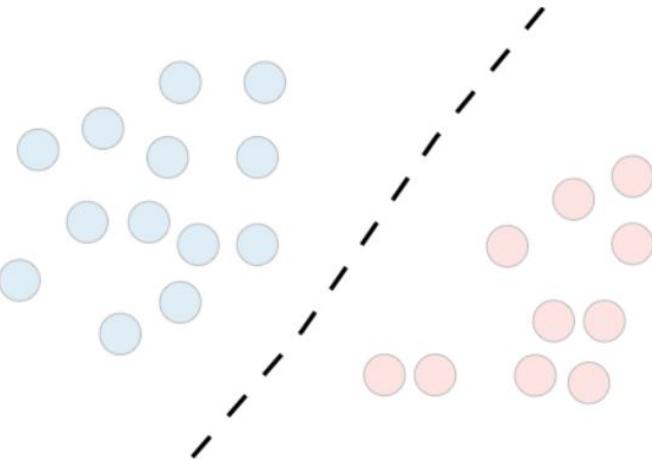
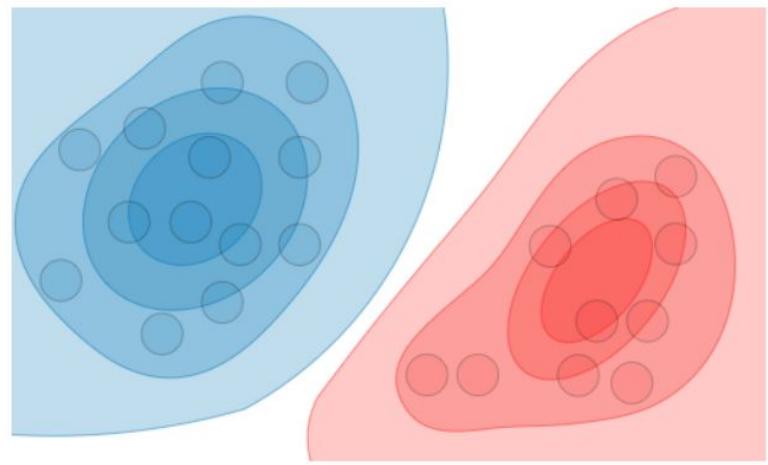


Introduction to Supervised Learning

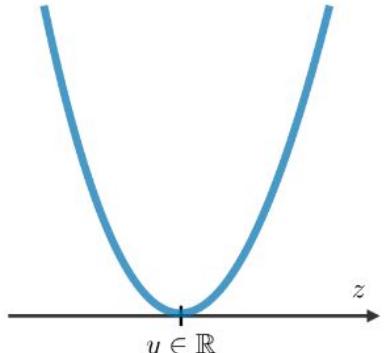
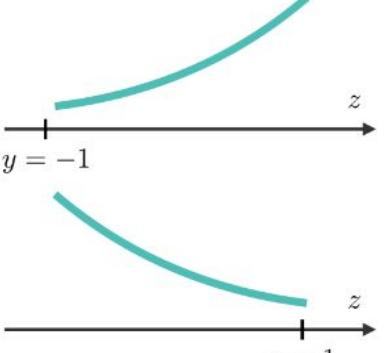
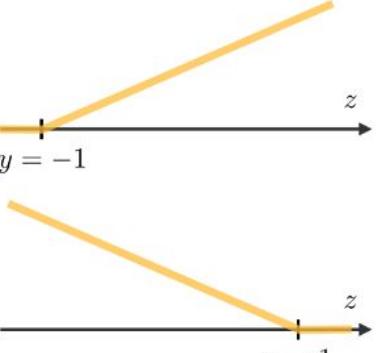
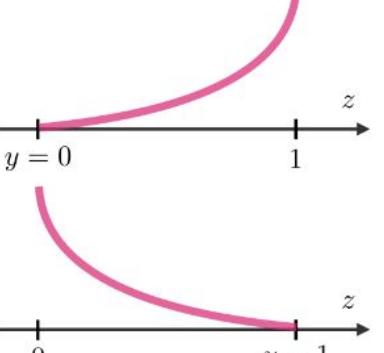
Given a set of data points $\{x^{(1)}, \dots, x^{(m)}\}$ associated to a set of outcomes $\{y^{(1)}, \dots, y^{(m)}\}$, we want to build a classifier that learns how to predict y from x .

□ **Type of prediction** — The different types of predictive models are summed up in the table below:

	Regression	Classification
Outcome	Continuous	Class
Examples	Linear regression	Logistic regression, SVM, Naive Bayes

	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

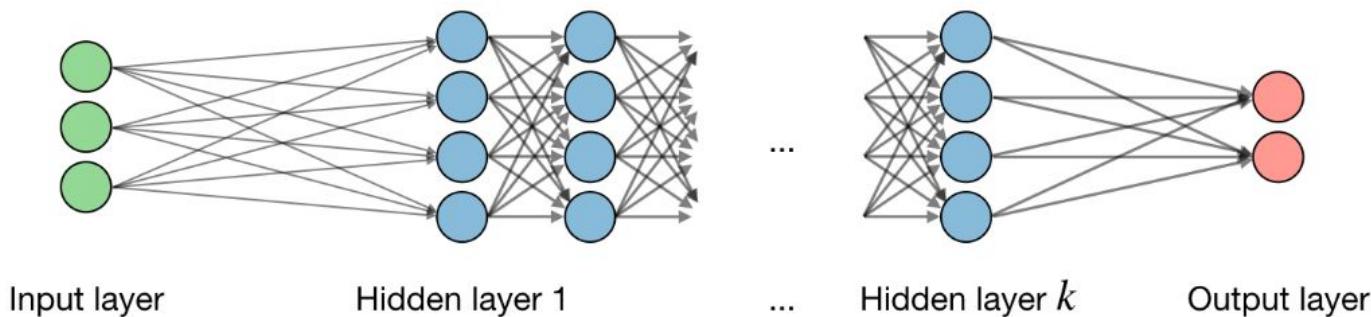
□ **Loss function** — A loss function is a function $L : (z, y) \in \mathbb{R} \times Y \mapsto L(z, y) \in \mathbb{R}$ that takes as inputs the predicted value z corresponding to the real data value y and outputs how different they are. The common loss functions are summed up in the table below:

Least squared error	Logistic loss	Hinge loss	Cross-entropy
$\frac{1}{2}(y - z)^2$	$\log(1 + \exp(-yz))$	$\max(0, 1 - yz)$	$-(y \log(z) + (1 - y) \log(1 - z))$
			
Linear regression	Logistic regression	SVM	Neural Network

Neural Networks

Neural networks are a class of models that are built with layers. Commonly used types of neural networks include convolutional and recurrent neural networks.

□ **Architecture** — The vocabulary around neural networks architectures is described in the figure below:



By noting i the i^{th} layer of the network and j the j^{th} hidden unit of the layer, we have:

$$z_j^{[i]} = w_j^{[i]T} x + b_j^{[i]}$$

where we note w , b , z the weight, bias and output respectively.

□ **Backpropagation** — Backpropagation is a method to update the weights in the neural network by taking into account the actual output and the desired output. The derivative with respect to weight w is computed using chain rule and is of the following form:

$$\frac{\partial L(z, y)}{\partial w} = \frac{\partial L(z, y)}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial w}$$

As a result, the weight is updated as follows:

$$w \leftarrow w - \alpha \frac{\partial L(z, y)}{\partial w}$$

□ **Updating weights** — In a neural network, weights are updated as follows:

- Step 1: Take a batch of training data.
- Step 2: Perform forward propagation to obtain the corresponding loss.
- Step 3: Backpropagate the loss to get the gradients.
- Step 4: Use the gradients to update the weights of the network.

Forward-Forward Algorithm: Will it replace Backpropagation?

The Forward-Forward Algorithm: Some Preliminary Investigations

Geoffrey Hinton
Google Brain
geoffhinton@google.com

Abstract

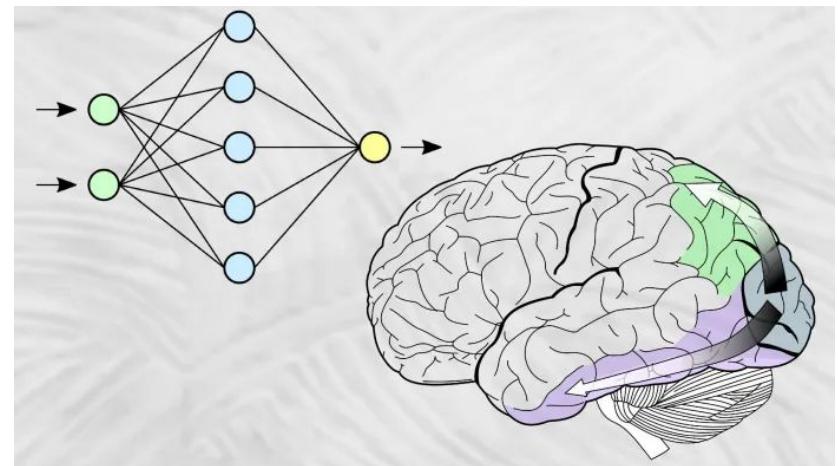
The aim of this paper is to introduce a new learning procedure for neural networks and to demonstrate that it works well enough on a few small problems to be worth serious investigation. The Forward-Forward algorithm replaces the forward and backward passes of backpropagation by two forward passes, one with positive (*i.e.* real) data and the other with negative data which could be generated by the network itself. Each layer has its own objective function which is simply to have high goodness for positive data and low goodness for negative data. The sum of the squared activities in a layer can be used as the goodness but there are many other possibilities, including minus the sum of the squared activities. If the positive and negative passes can be separated in time, the negative passes can be done offline, which makes the learning much simpler in the positive pass and allows video to be pipelined through the network without ever storing activities or stopping to propagate derivatives.



Forward-Forward Algorithm: Will it replace Backpropagation?

In the visual cortex (right), information moves in several directions. In neural networks (left), information moves in one direction.

- The idea behind the forward-forward algorithm is to replace the forward and backward passes of backpropagation with two forward passes.
- The two passes are similar, but they **work on different data and have opposite objectives**.
- The “**positive pass**” operates on real data and adjusts the weights of the network to increase the “**goodness**” of each layer.
- The “**negative pass**” operates on negative data and adjusts the weights to reduce goodness.

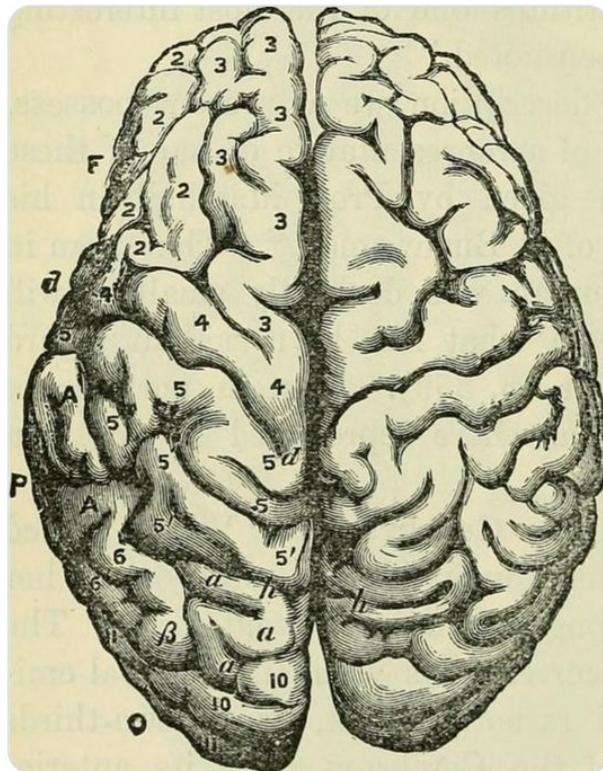




Martin Görner @martin_gorner · Dec 5, 2022

...

I seems very unlikely that the human brain uses back propagation to learn. There is little evidence of backprop mechanics in biological brains (no error derivatives propagating backwards, no storage of neuron activities to use in a packprop pass, ...).



Martin Görner

@martin_gorner

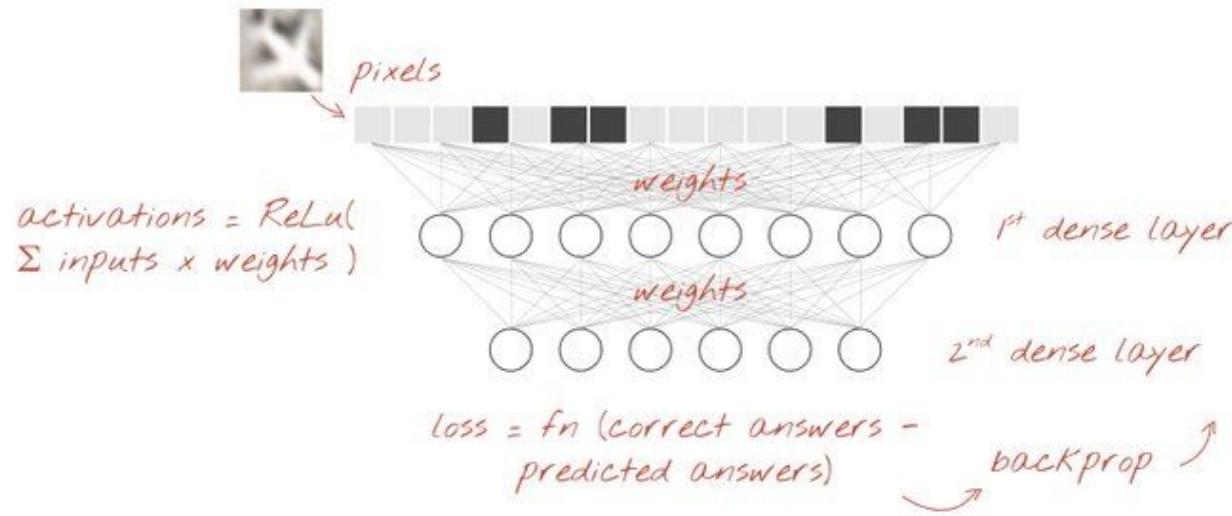
...

Also, the brain can learn from a continuous stream of incoming data and does not need to stop to run a backprop pass. Yes, sleep is beneficial for learning somehow, but we can learn awake too.



The Forward Forward (FF) algorithm

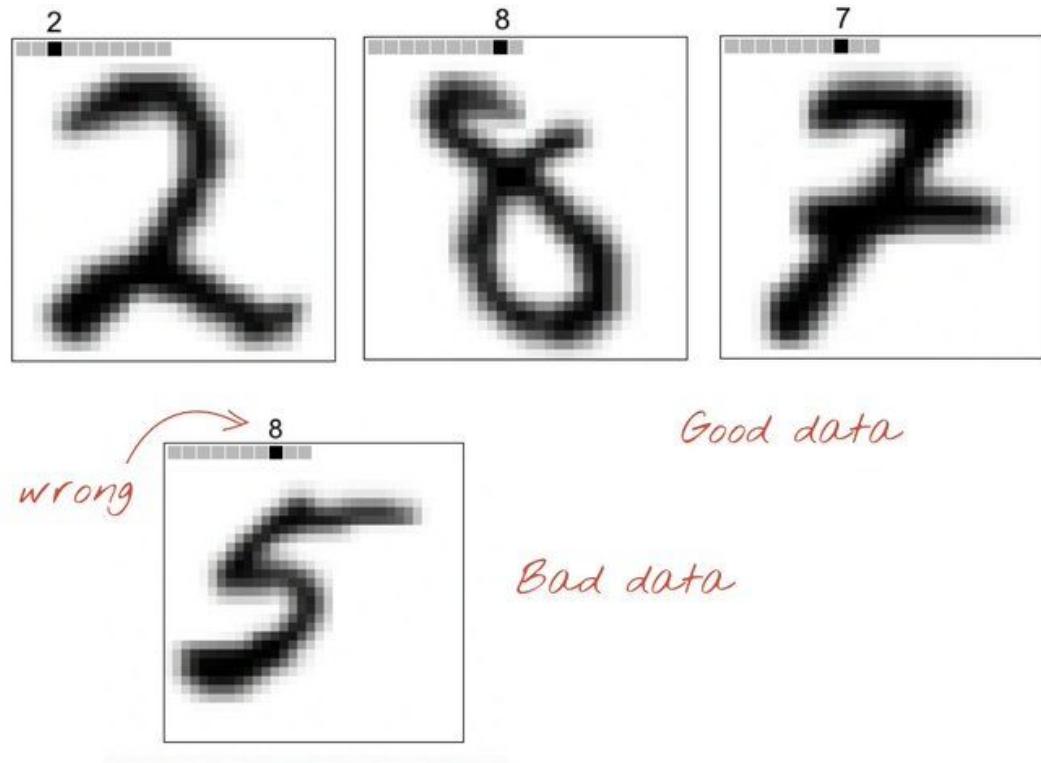
Let's take a 2 layer dense neural network and implement an image classifier. The regular, backprop way of training is to push data through the network, compare the outcome to labels, compute a loss fn, differentiate it and update weights through backprop.



The Forward Forward (FF) algorithm (Cont.)

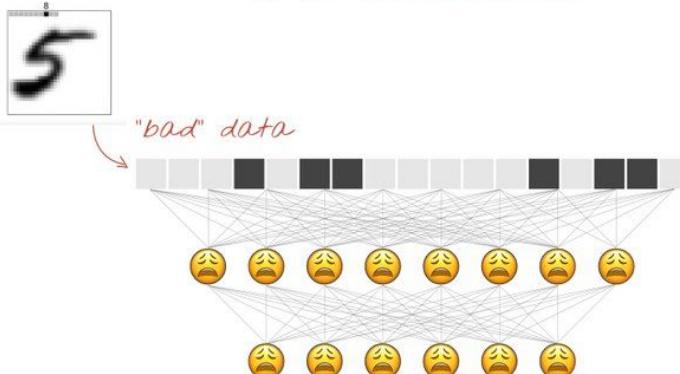
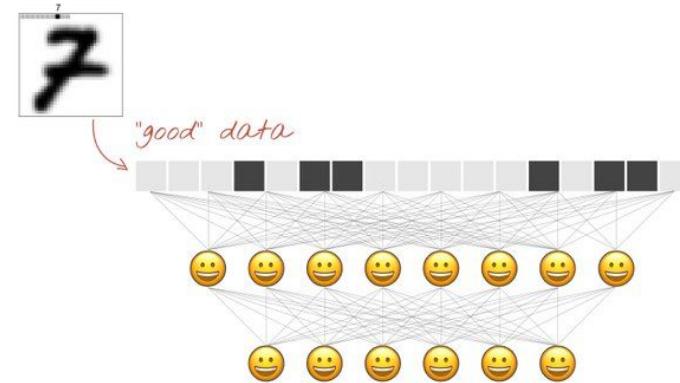
Instead in FF, we merge labels and data in one vector, the "**good data**", and then generate "**bad data**" by merging input images with the wrong labels.

Here is what this could look like on MNIST handwritten digits:



The Forward Forward (FF) algorithm (Cont.)

And now, we want each network layer to feel "excited" about the good data, i.e. exhibit high activations, and on the contrary have low activations on "bad data".



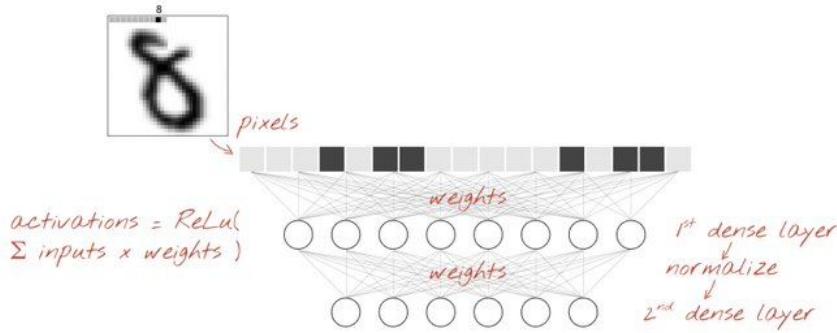
The Forward Forward (FF) algorithm (Cont.)

That's very easy to do, compute the sum of squared activations in a layer and **maximize it on "good data"** while **minimizing it on "bad data"**.

This is purely within one layer, so **no backprop is necessary**, just a derivative of the function performed by the layer itself (easy).

When chaining multiple layers, you want to avoid seeing height activations just because you pumped in high inputs from a very excited previous layer.

So layer activations are **normalized** (divided by the norm of the activation vector) between layers.



The Forward Forward (FF) algorithm (Cont.)

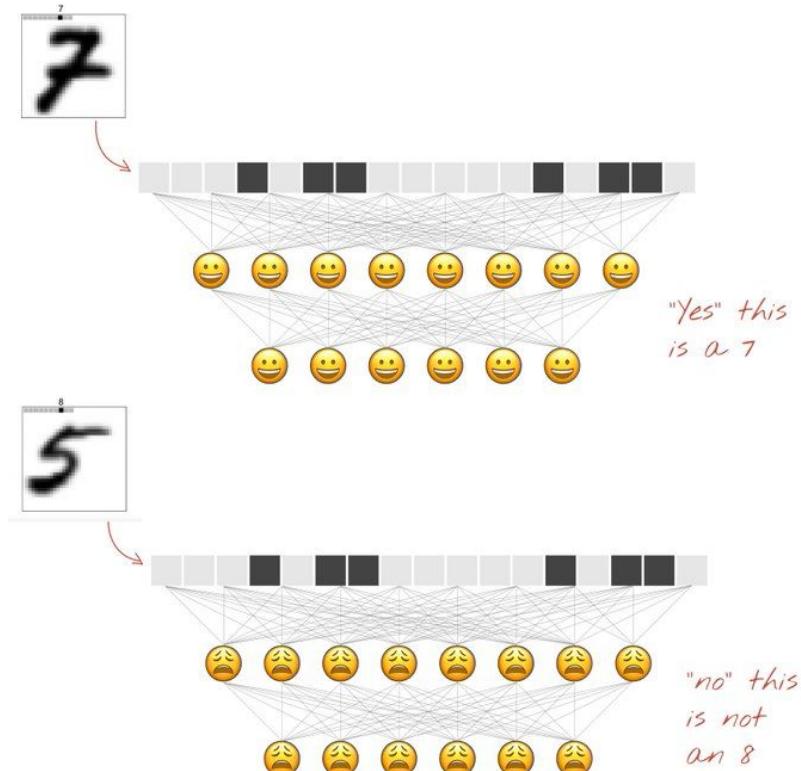
In short, the activation vector retains its direction between layers, but its norm is reset to 1 before passing it to the next layer.

The training then alternates good data and bad data, and performs weight updates in each layer accordingly.

How about inference ?

Again concatenate an MNIST digit with a label, any label, feed it through the network and see if the network as a whole is excited about it.

You will have to do it for all image-label combinations though.



The Forward Forward (FF) algorithm (Cont.)

Even better, the Forward-Forward algorithm lends itself quite naturally to self-supervised learning.

You can learn about "digits" by feeding non-digits as "bad data" (example non-digit generation from paper). Such a network can then be fine-tuned on a downstream task.

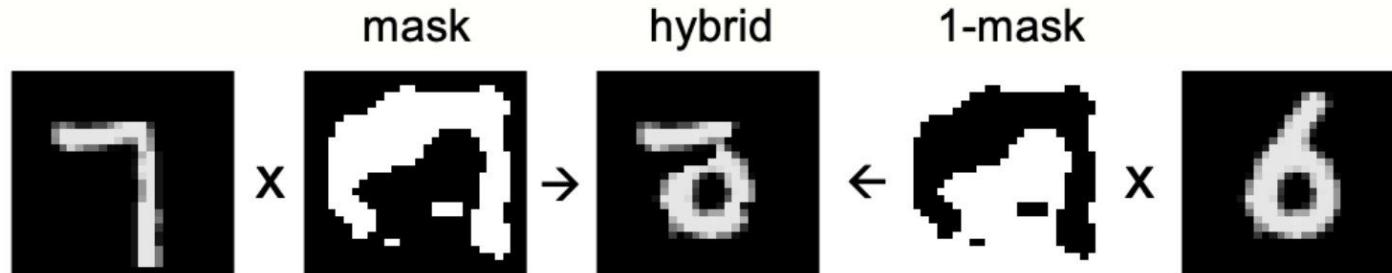
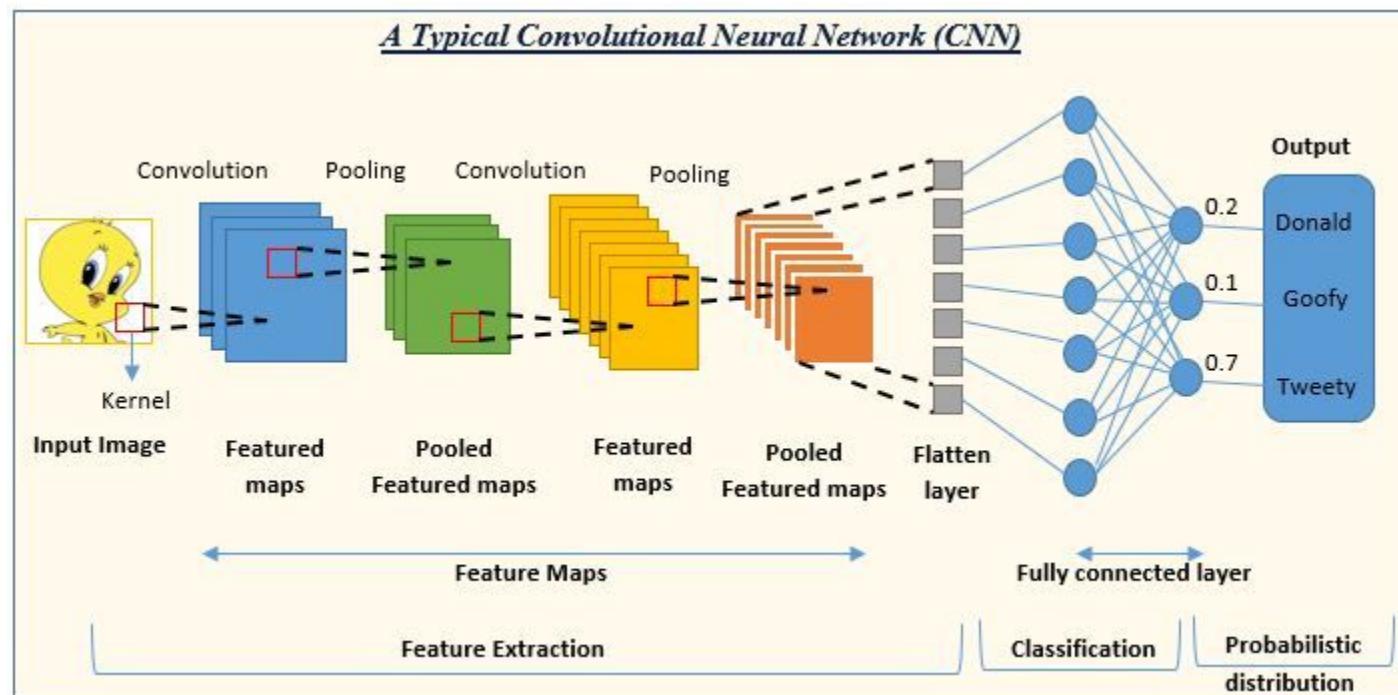


Figure 1: A hybrid image used as negative data

Convolutional Neural Networks



Convolutional Neural Networks (cont.)

□ **Convolutional layer requirement** — By noting W the input volume size, F the size of the convolutional layer neurons, P the amount of zero padding, then the number of neurons N that fit in a given volume is such that:

$$N = \frac{W - F + 2P}{S} + 1$$

□ **Batch normalization** — It is a step of hyperparameter γ, β that normalizes the batch $\{x_i\}$. By noting μ_B, σ_B^2 the mean and variance of that we want to correct to the batch, it is done as follows:

$$x_i \leftarrow \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

It is usually done after a fully connected/convolutional layer and before a non-linearity layer and aims at allowing higher learning rates and reducing the strong dependence on initialization.

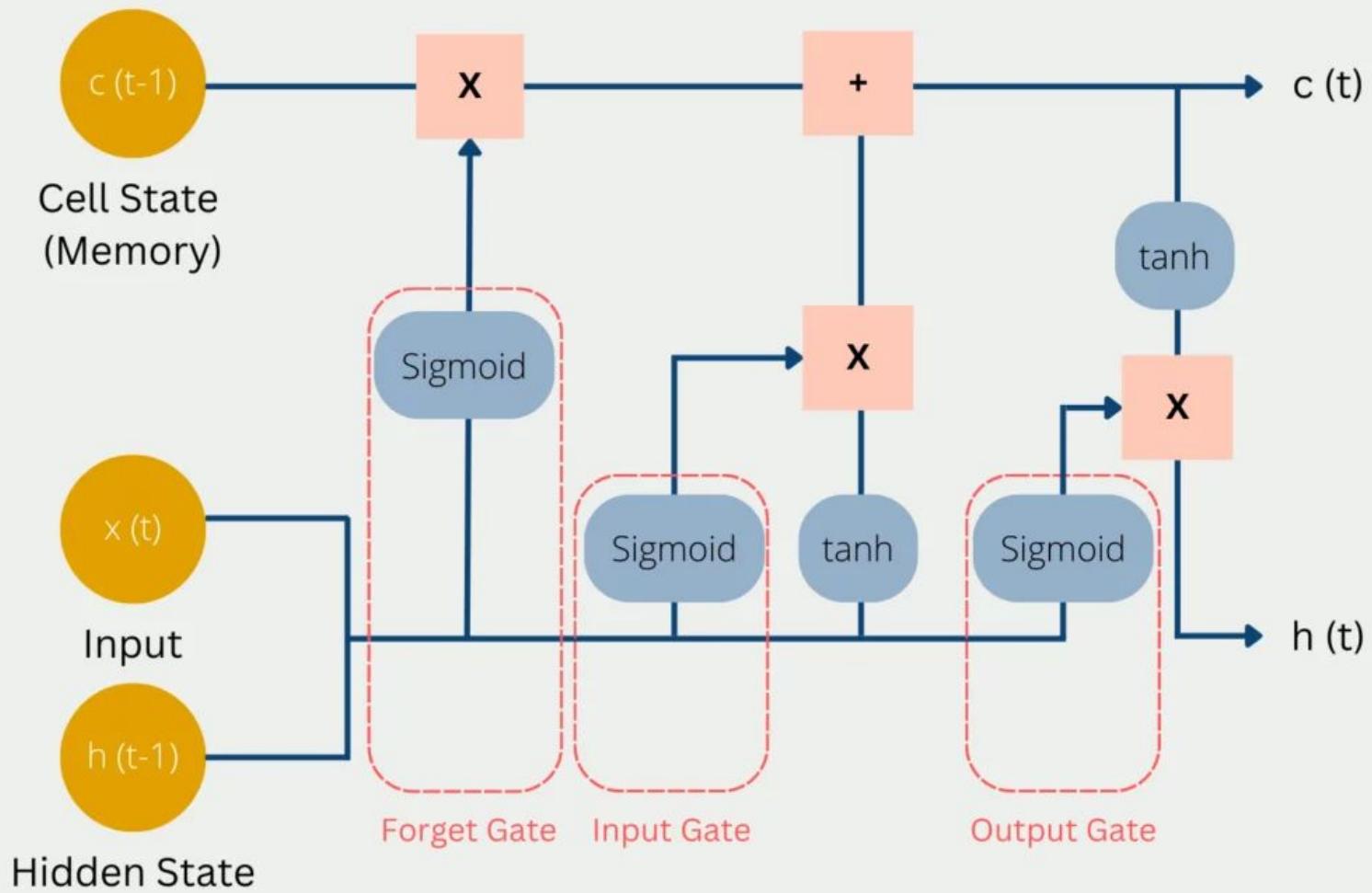
Recurrent Neural Networks

Recurrent Neural Networks

□ **Types of gates** — Here are the different types of gates that we encounter in a typical recurrent neural network:

Input gate	Forget gate	Gate	Output gate
Write to cell or not?	Erase a cell or not?	How much to write to cell?	How much to reveal cell?

□ **LSTM** — A long short-term memory (LSTM) network is a type of RNN model that avoids the vanishing gradient problem by adding 'forget' gates.



Reinforcement Learning and Control

The goal of reinforcement learning is for an agent to learn how to evolve in an environment.

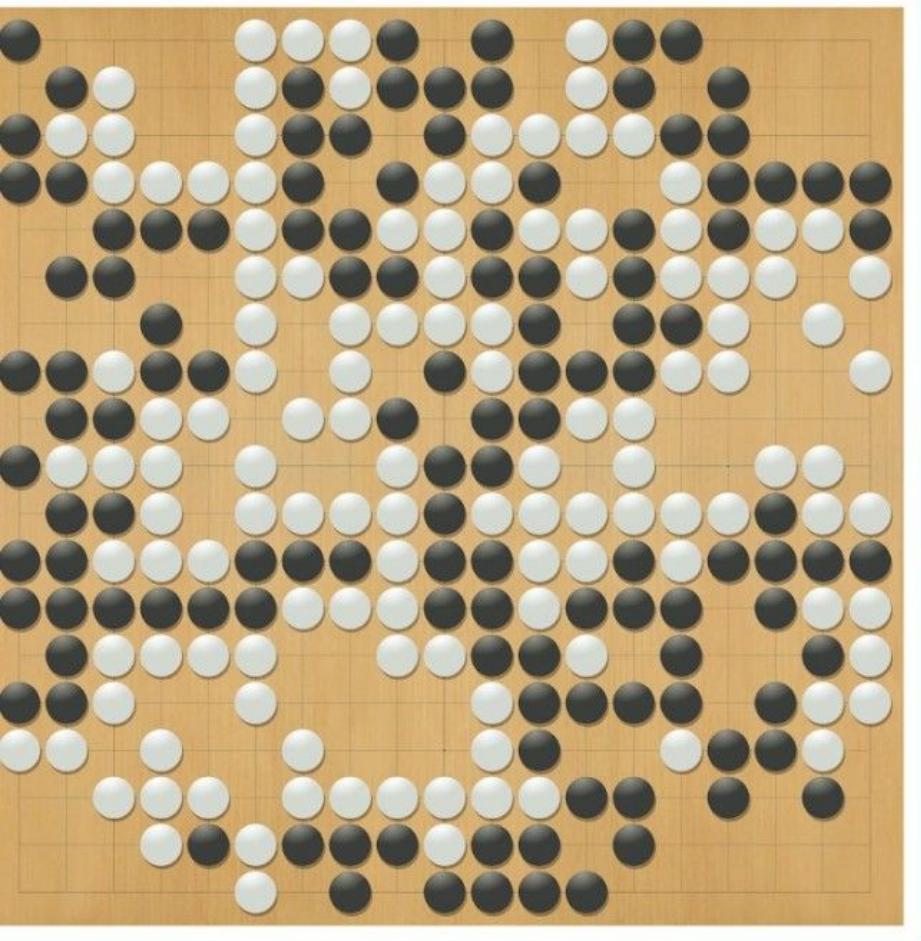
Definitions

□ **Markov decision processes** — A Markov decision process (MDP) is a 5-tuple $(\mathcal{S}, \mathcal{A}, \{P_{sa}\}, \gamma, R)$ where:

- \mathcal{S} is the set of states
- \mathcal{A} is the set of actions
- $\{P_{sa}\}$ are the state transition probabilities for $s \in \mathcal{S}$ and $a \in \mathcal{A}$
- $\gamma \in [0, 1[$ is the discount factor
- $R : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$ or $R : \mathcal{S} \longrightarrow \mathbb{R}$ is the reward function that the algorithm wants to maximize

□ **Policy** — A policy π is a function $\pi : \mathcal{S} \longrightarrow \mathcal{A}$ that maps states to actions.

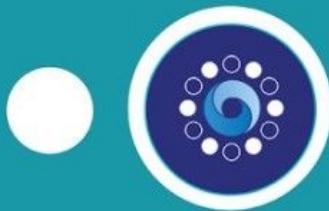
Remark: we say that we execute a given policy π if given a state s we take the action $a = \pi(s)$.



THE ULTIMATE GO CHALLENGE

GAME 5 OF 5

15 MARCH 2016



AlphaGo
Won 4 of 5



Lee Sedol
Won 1 of 5

RESULT

W+
Res

NUMBER
OF MOVES

280

TIME
WHITE

2h+

TIME
BLACK

2h+

The Illustrated Transformer



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

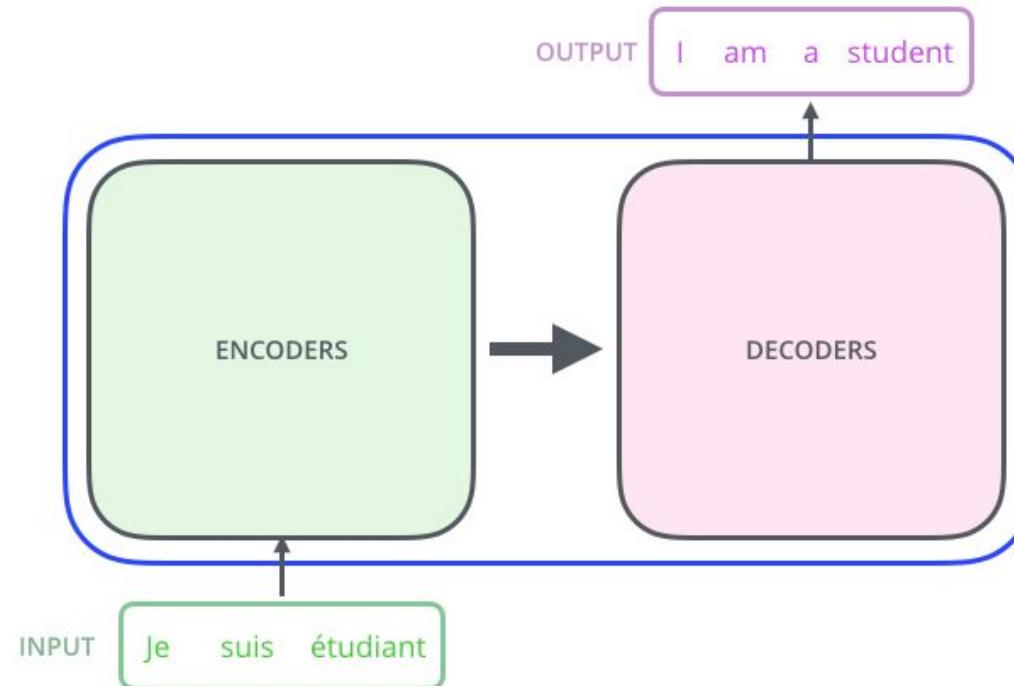
Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

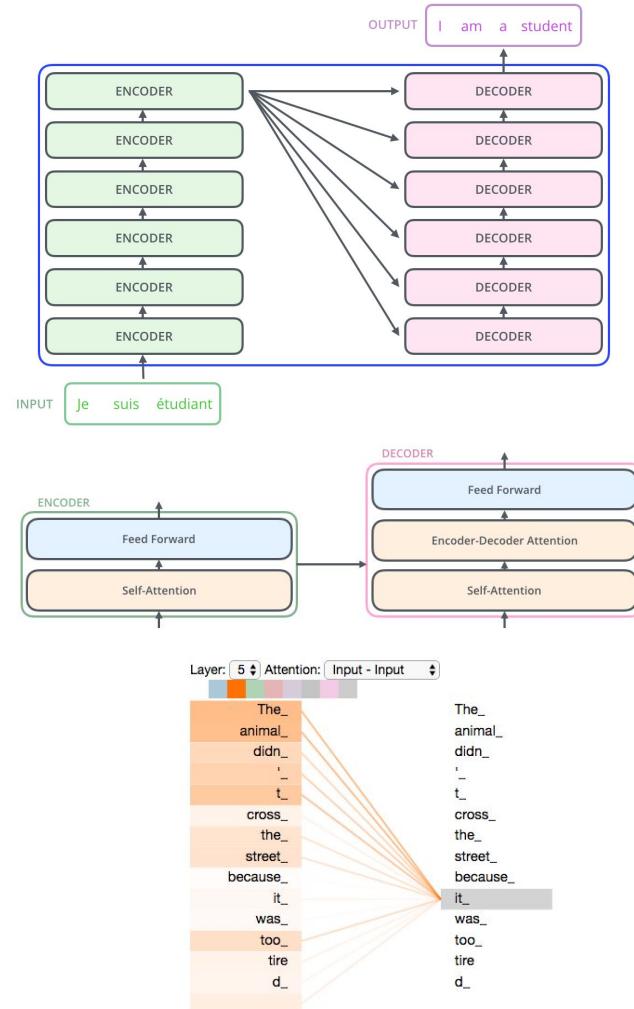
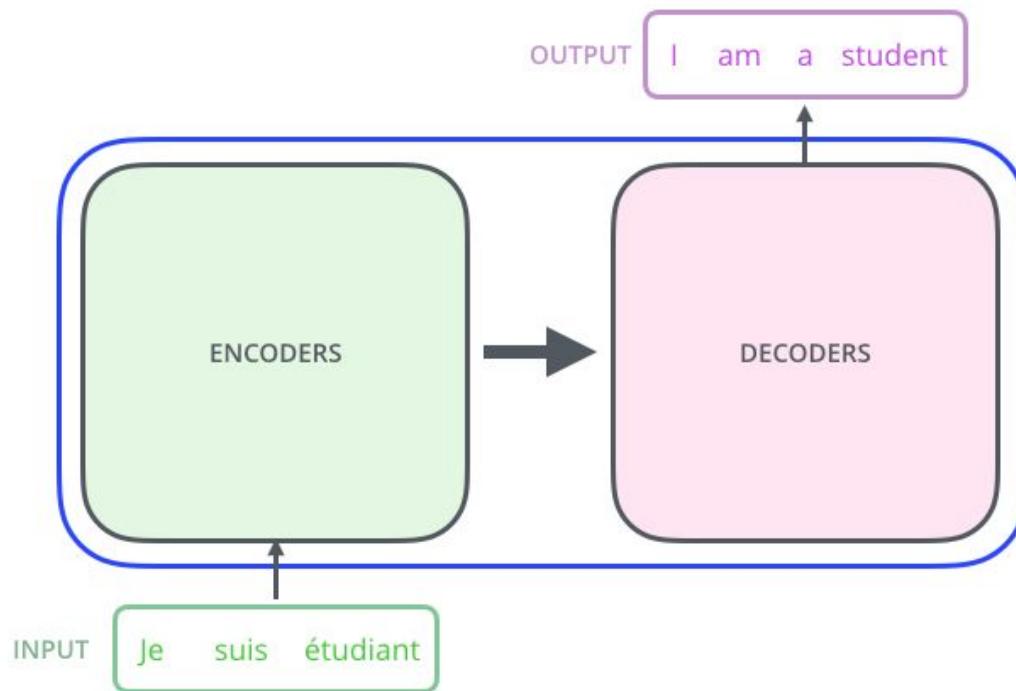
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

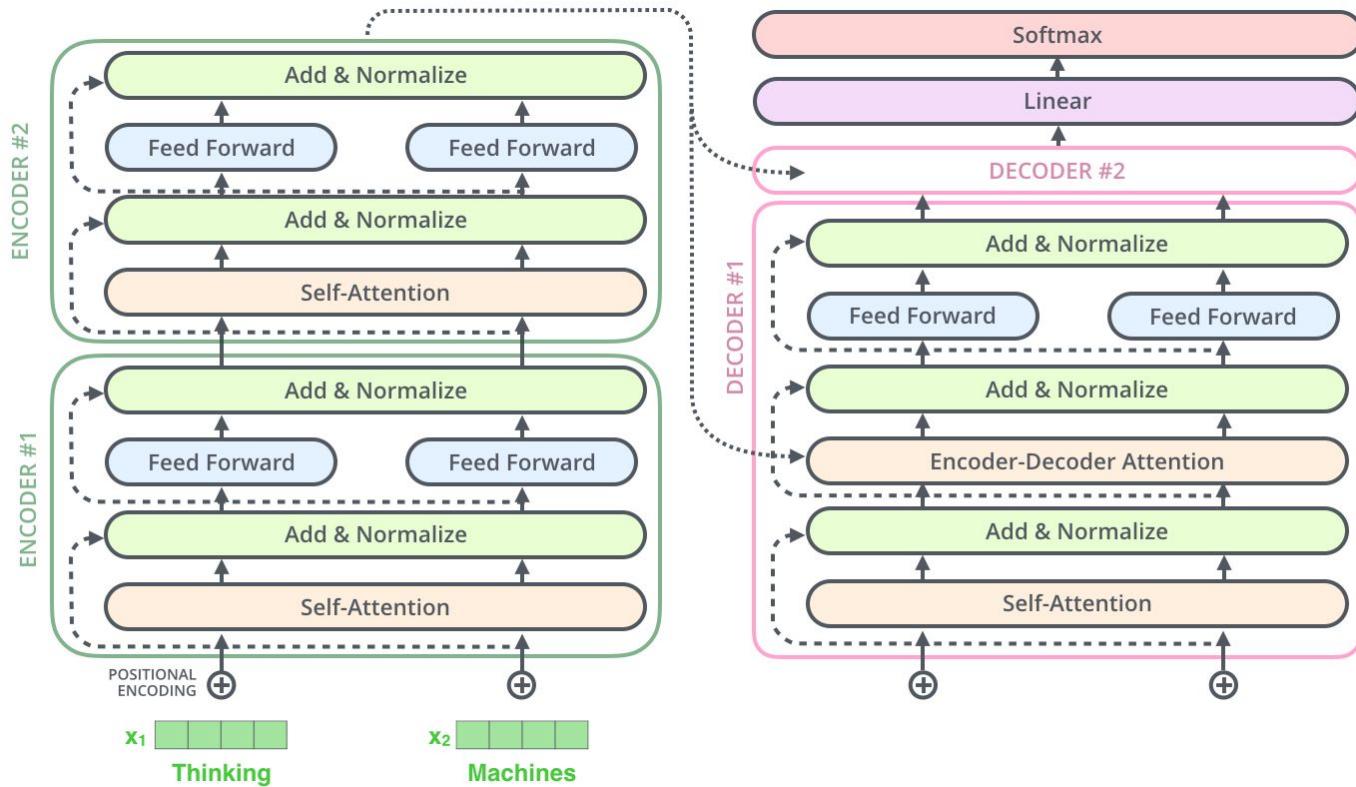
The Illustrated Transformer



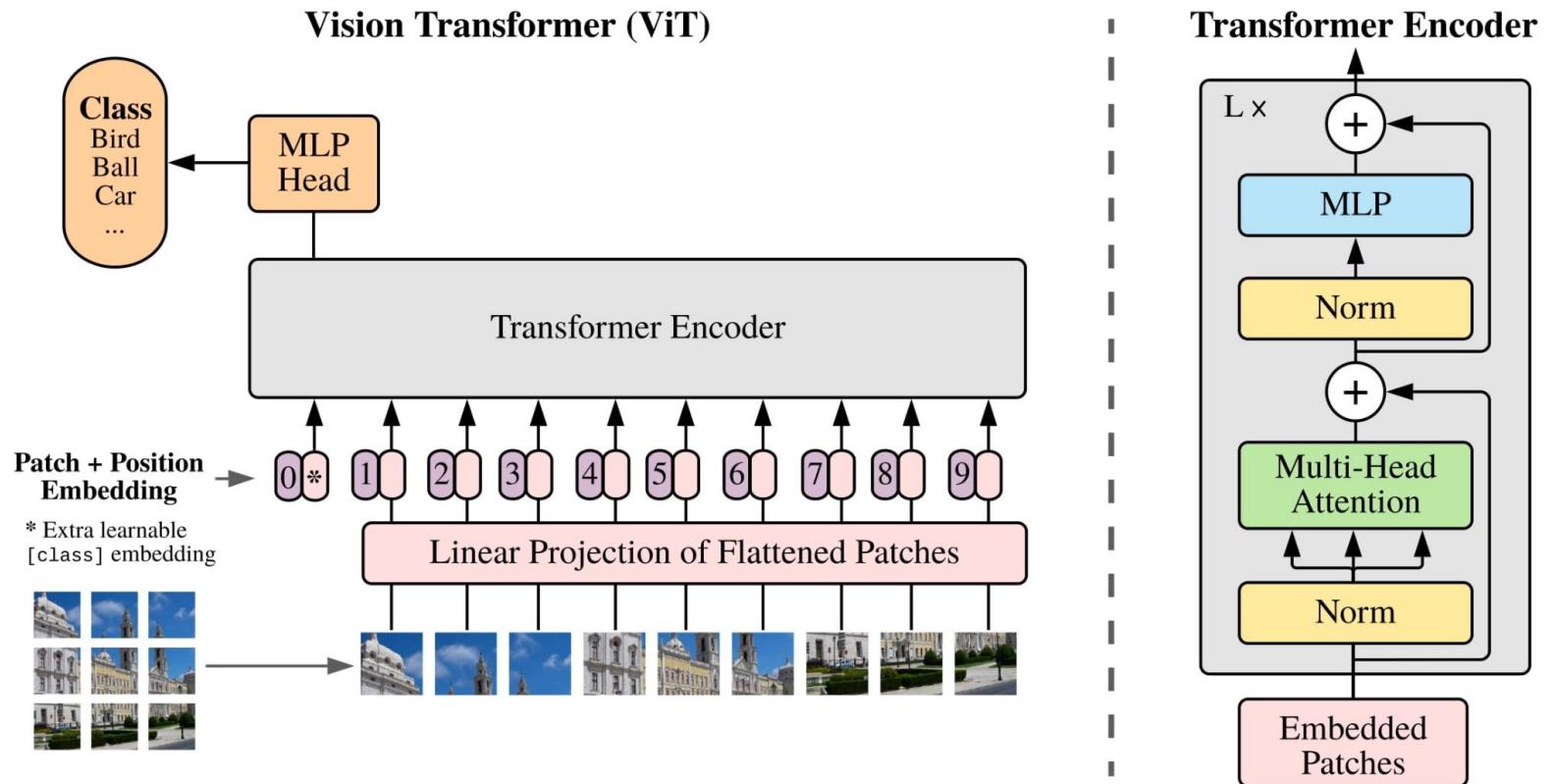
The Illustrated Transformer



The Illustrated Transformer



Vision Transformer (ViT)



Advancing the state of the art

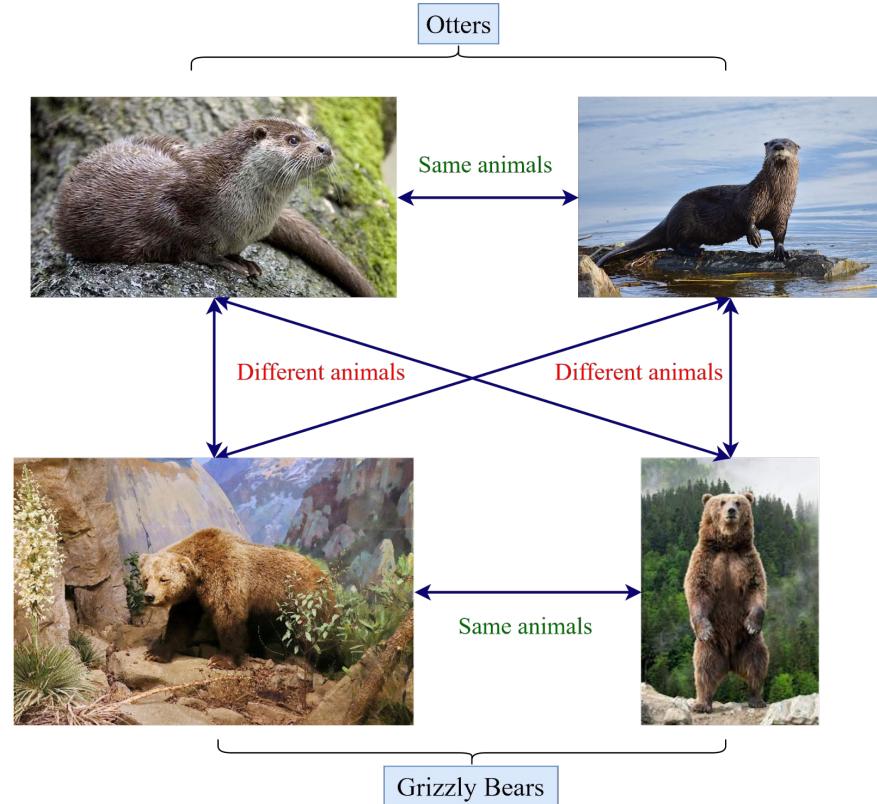
Google Research



Meta

Contrastive Learning

Machine Learning paradigm where unlabeled data points are juxtaposed against each other to teach a model which points are similar and which are different.



<https://ai.googleblog.com/2021/06/extending-contrastive-learning-to.html>



<https://ai.googleblog.com/2021/06/extending-contrastive-learning-to.html>



Anchor



Positive

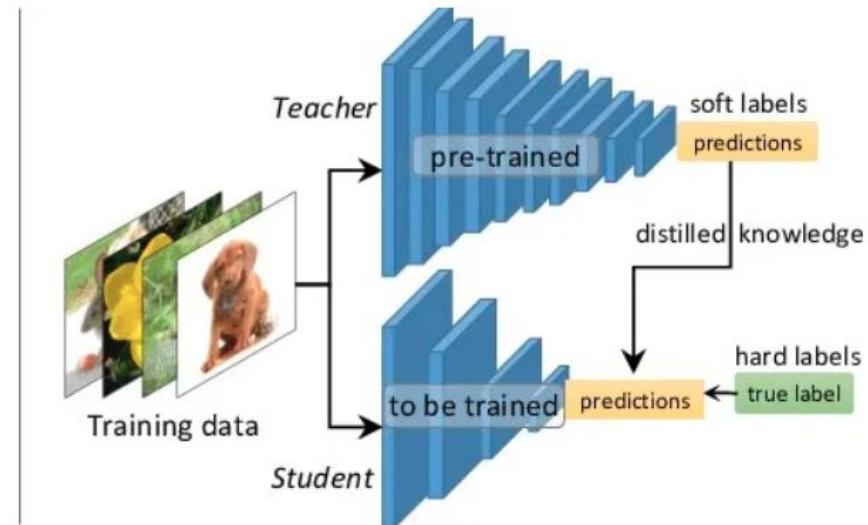
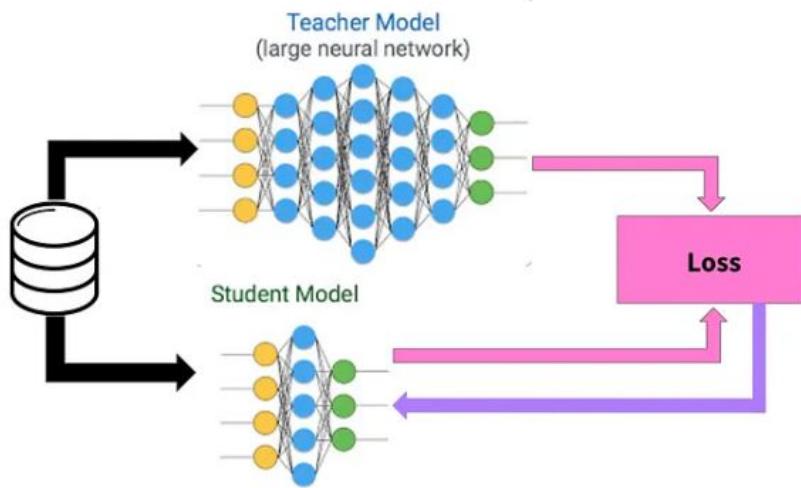
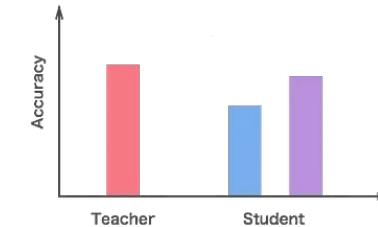
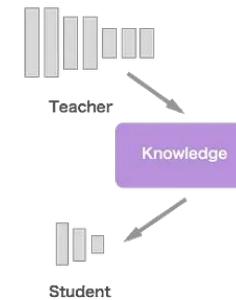


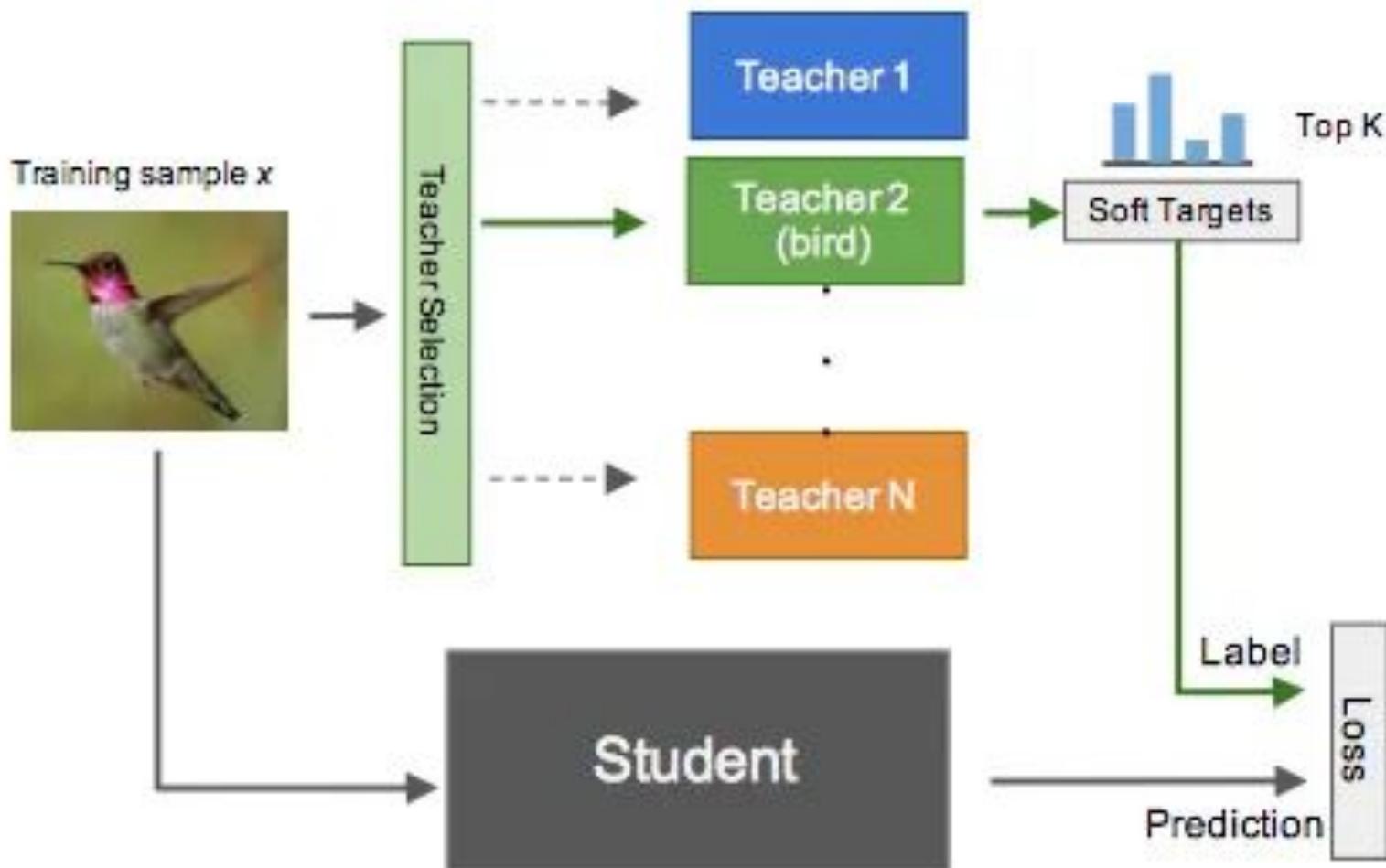
Anchor



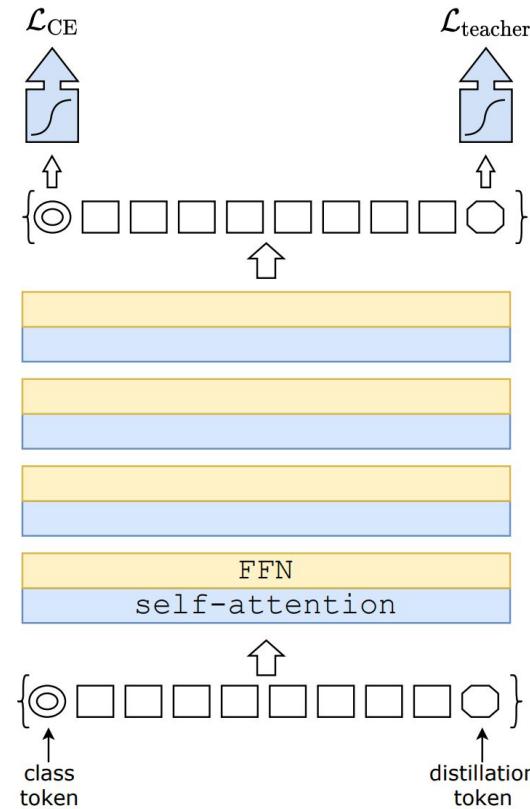
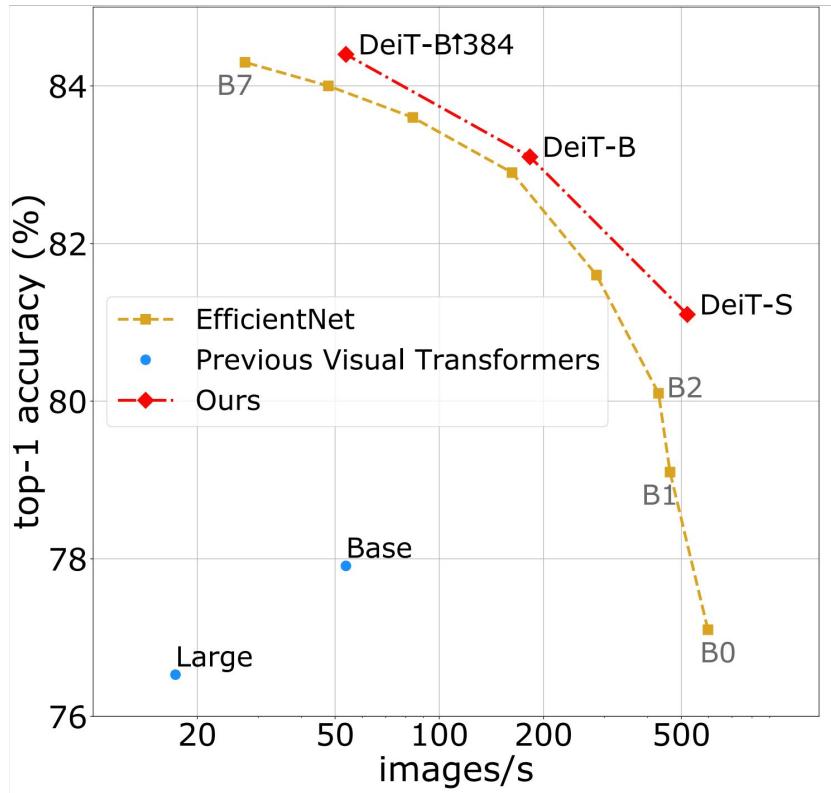
Negative

Knowledge Distillation: Simplified





Data-efficient image Transformers (DeiT)



Panoptic Segmentation

Alexander Kirillov^{1,2} Kaiming He¹ Ross Girshick¹ Carsten Rother² Piotr Dollár¹

¹Facebook AI Research (FAIR)

²HCI/IWR, Heidelberg University, Germany



(a) Image



(b) Semantic Segmentation



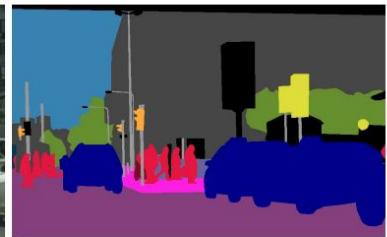
(c) Instance Segmentation



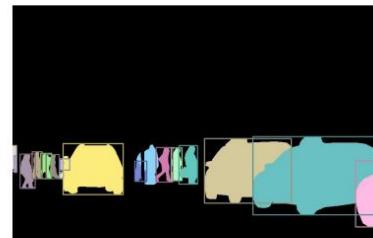
(d) Panoptic Segmentation



(a) image



(b) semantic segmentation



(c) instance segmentation



(d) panoptic segmentation

Stuff & things in panoptic segmentation

Things: In the realm of computer vision, "things" typically refer to objects that

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g) \in TP} \text{IoU}(p, g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}}.$$

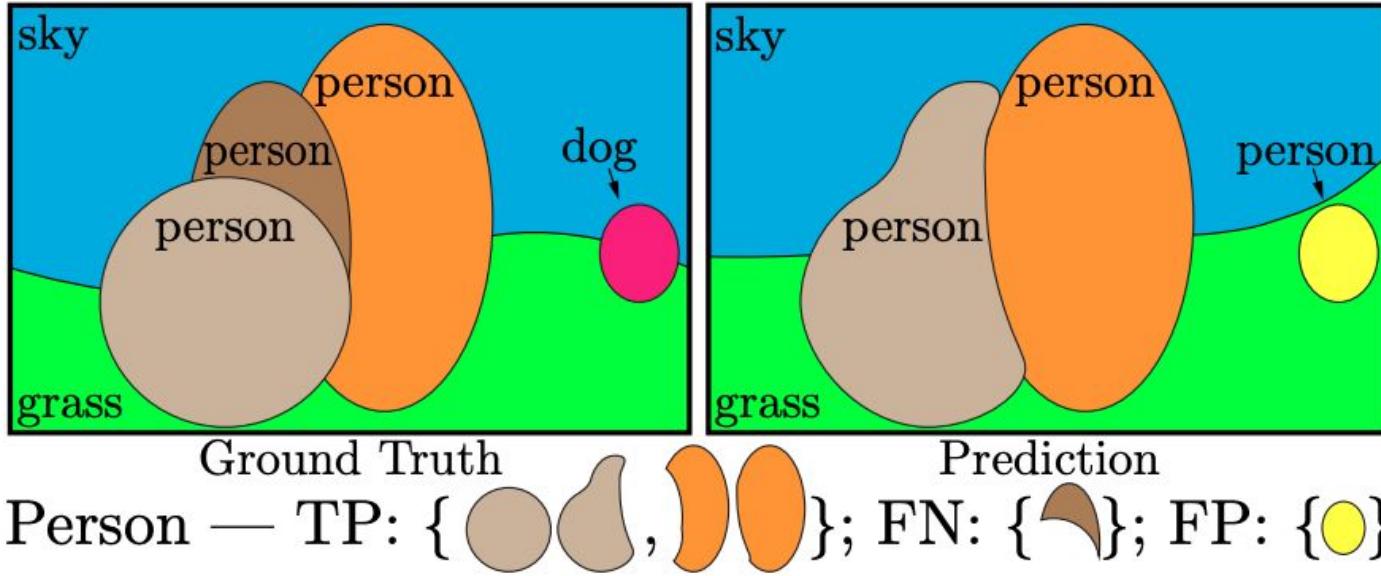


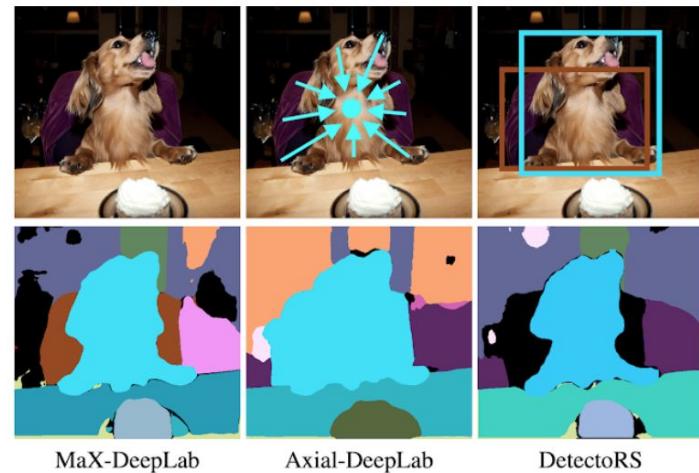
Figure 2: Toy illustration of ground truth and predicted panoptic segmentations of an image. Pairs of segments of the same color have IoU larger than 0.5 and are therefore matched. We show how the segments for the *person* class are partitioned into true positives *TP*, false negatives *FN*, and false positives *FP*.

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g) \in TP} \text{IoU}(p, g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}}.$$

MaX-DeepLab: Dual-Path Transformers for End-to-End Panoptic Segmentation

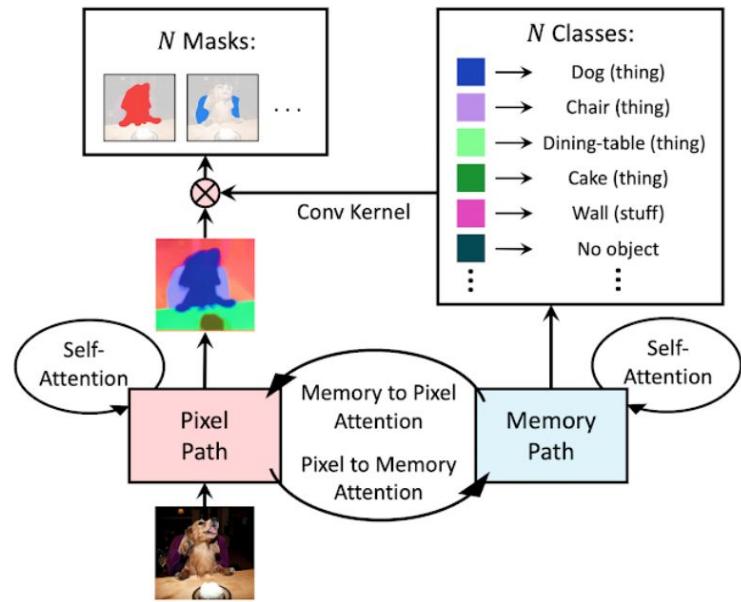
Method		PQ
Box-based	DETR	46.0 (-5.3)
	DetectoRS	49.6 (-1.7)
Box-free	Panoptic-DeepLab	41.4 (-9.9)
	Axial-DeepLab	44.2 (-7.1)
	MaX-DeepLab	51.3

Comparison on COCO test-dev set.

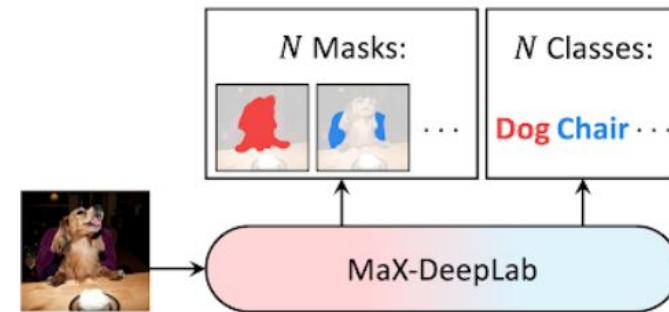


A case study for MaX-DeepLab and state-of-the-art box-free and box-based methods.

MaX-DeepLab: Dual-Path Transformers for End-to-End Panoptic Segmentation



An overview of the dual-path transformer architecture.



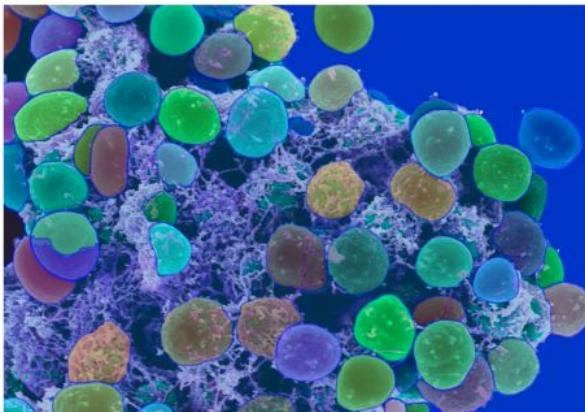
MaX-DeepLab directly predicts *N* masks and *N* classes with a CNN and a mask transformer.



When the centers of the dog and the chair are close to each other, Axial-DeepLab merges them into one object.

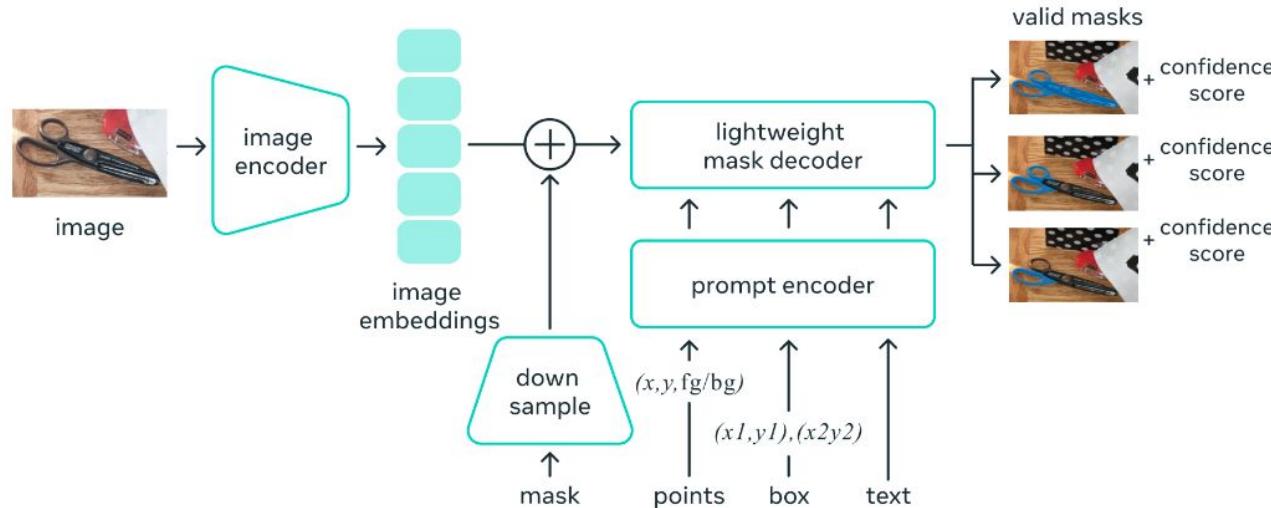
Segment Anything Model (SAM)

- a new AI model from Meta AI that can "cut out" any object, in any image, with a single click



Segment Anything Model (SAM)

Universal segmentation model



In a web browser, SAM efficiently maps the image features and a set of prompt embeddings to produce a segmentation mask.

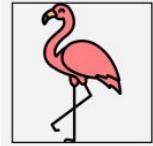
DeepMind Flamingo: A Visual Language Model for Few-Shot Learning

Flamingo is a new visual language model (VLM) capable of multimodal tasks like captioning, visual dialogue, classification, and visual question answering. As you can see, it works rather well:

Single Image Samples			
Input Prompt	Completion	Completion	Completion
	<p>Question: What do you think the capacities of these are? Answer:</p> <p>The floppy disk is 1.44MB and the CD is 700MB.</p>	<p>Question: What is odd about this image? Explain why it is unusual. Answer:</p> <p>The image is odd because the elephant is in the back of the truck. It is unusual because elephants are not usually transported in the back of a truck.</p>	<p>Question: What country is this? Why do you think so? Answer:</p> <p>It is Canada. I think so because the flag is the Canadian flag.</p>

Examples of inputs and outputs obtained from 80B parameter Flamingo model

Input Prompt				Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.	 This is a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvre Museum, Paris.	 What is the name of the city where this was painted? Answer: Arles.
	Output: "Underground"		Output: "Congress"	 Output: "Soulomes"
	2+1=3		5+6	 3x6=18
	Output: A propaganda poster depicting a cat dressed as French emperor Napoleon holding a piece of cheese.		Output: A pink room with a flamingo pool float.	 Output: A portrait of Salvador Dali with a robot head.
	Les sanglots longs des violons de l'automne blessent mon cœur d'une langueur monotone.		Pour qui sont ces serpents qui sifflent sur vos têtes?	 Je suis un cœur qui bat pour vous.
	pandas: 3		dogs: 2	 giraffes: 4
I like reading		, my favourite play is Hamlet. I also like		 , my favorite book is Dreams from my Father.
				What happens to the man after hitting the ball? Answer: he falls down.



What is the common thing about these three images?

They are all flamingos.

What is the difference between these three images?

The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.

Key Ideas of Flamingo

- A novel architecture for accepting arbitrarily interleaved **visual and text data as input** and generating **output text in an open-ended manner**.
- Architectural innovations and training strategies that effectively leverage large pre-trained vision-only and language-only models, saving tons of compute and preserving the benefits of these initial models while efficiently fusing the modalities. Specifically, the authors used **Chinchilla**, a 70B state-of-the-art LM (which is frozen in Flamingo) and trained Flamingo, an 80B parameter VLM.
- Efficient ways to adapt to **visual inputs of varying sizes**, making Flamingo applicable to **images and videos**.

A New AI Trend: Chinchilla (70B) Greatly Outperforms GPT-3 (175B) and Gopher (280B)

DeepMind has found the secret to cheaply scale large language models.



Alberto Romero · [Follow](#)

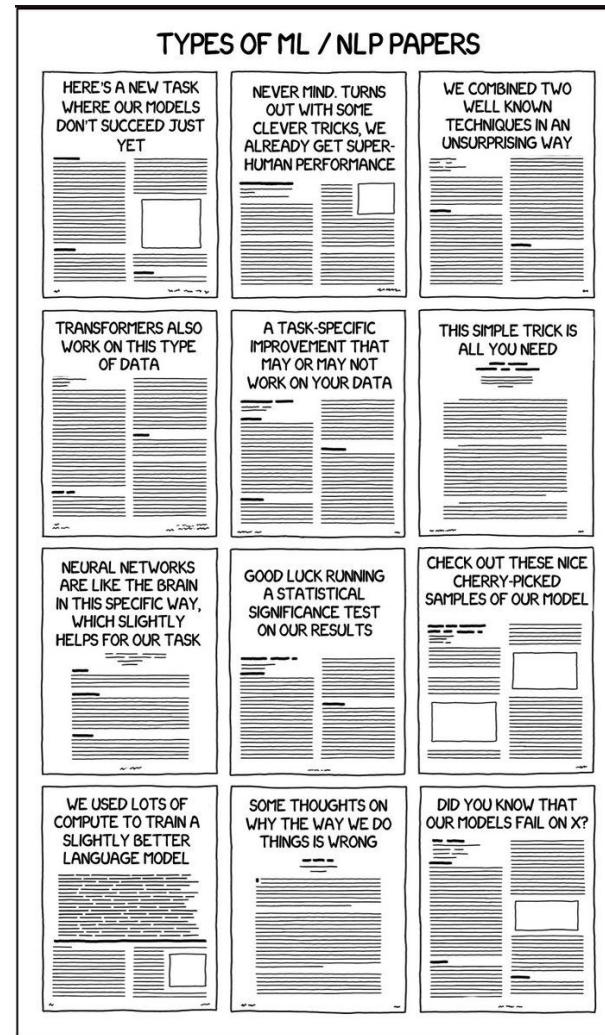
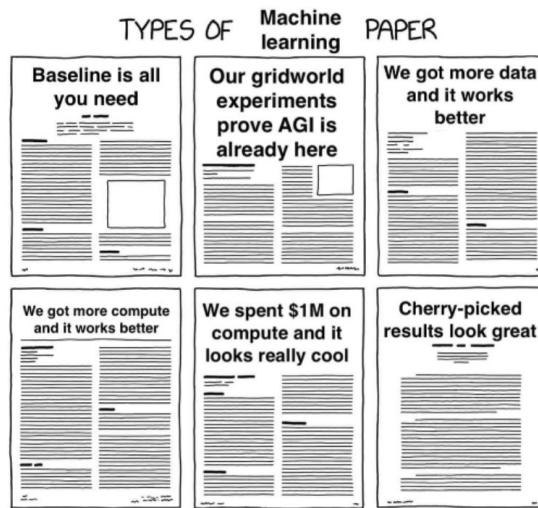
Published in [Towards Data Science](#) · 11 min read · Apr 11, 2022



Types of ML / NLP Papers



Sebastian Ruder ✅
@seb_ruder

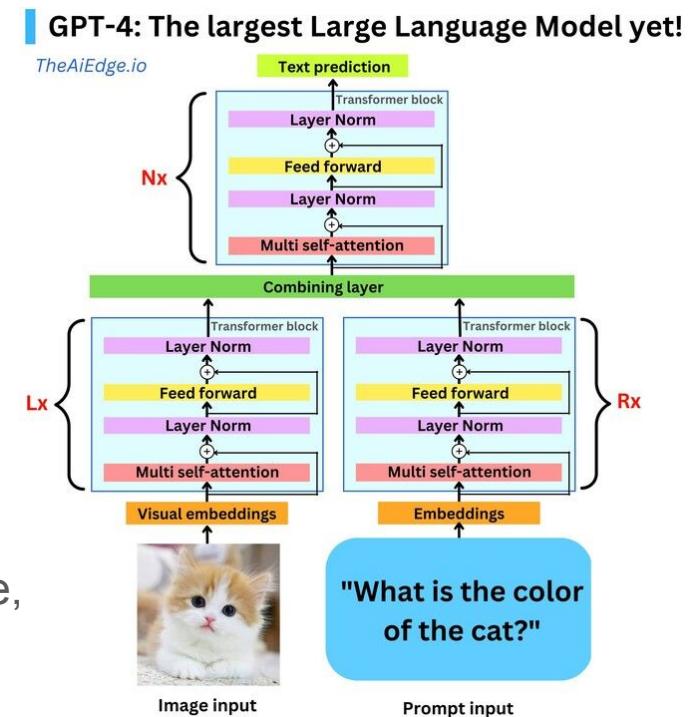


An Overview of the GPT-4 Architecture and Capabilities of Next-Generation AI

GPT-4 is a new language model created by OpenAI that is a large multimodal that can accept image and text inputs and emit outputs. It exhibits human-level performance on various professional and academic benchmarks.

** Generative Pre-trained Transformers

- **Multimodal technology** refers to systems that can process and integrate multiple types of inputs and outputs, such as text, speech, image, video, gesture, etc. Multimodal systems can enable more natural and efficient human-computer interactions



<https://arxiv.org/pdf/2302.14045.pdf>

An Overview of the GPT-4 Architecture and Capabilities of Next-Generation AI

Language Is Not All You Need: Aligning Perception with Language Models

Shaohan Huang*, Li Dong*, Wenhui Wang*, Yaru Hao*, Saksham Singhal*, Shuming Ma*
Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal
Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhajit Som, Xia Song, Furu Wei†
Microsoft

<https://github.com/microsoft/unilm>