LATVIJAS UNIVERSITĀTE

DATORIKAS FAKULTĀTE

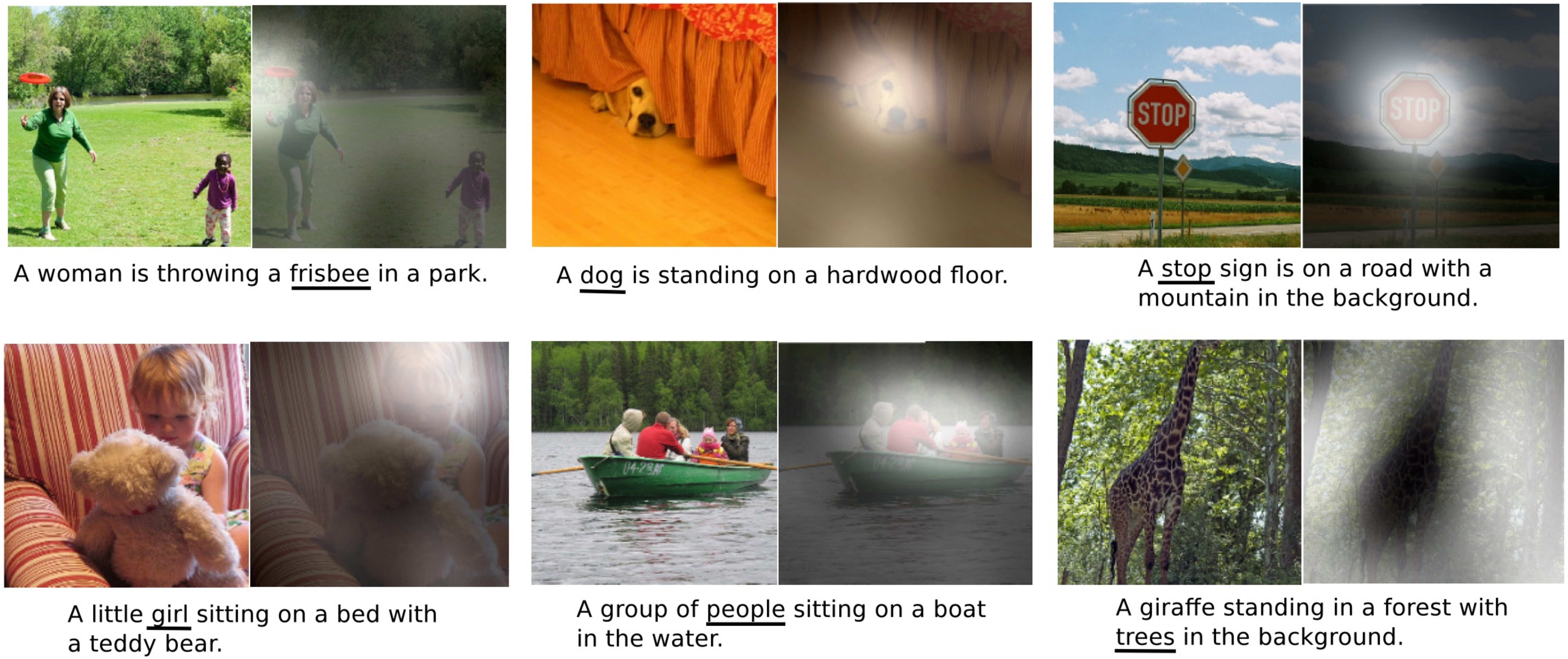
TRANSFORMERS

And its applications

Darba autors: **Eduards Mukāns**
Stud. apl. Nr. em18044
Darba vadītājs: Dr. Sc. Comp. Guntis Bārzdiņš

Rīga, 2021

Attention is all you need



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

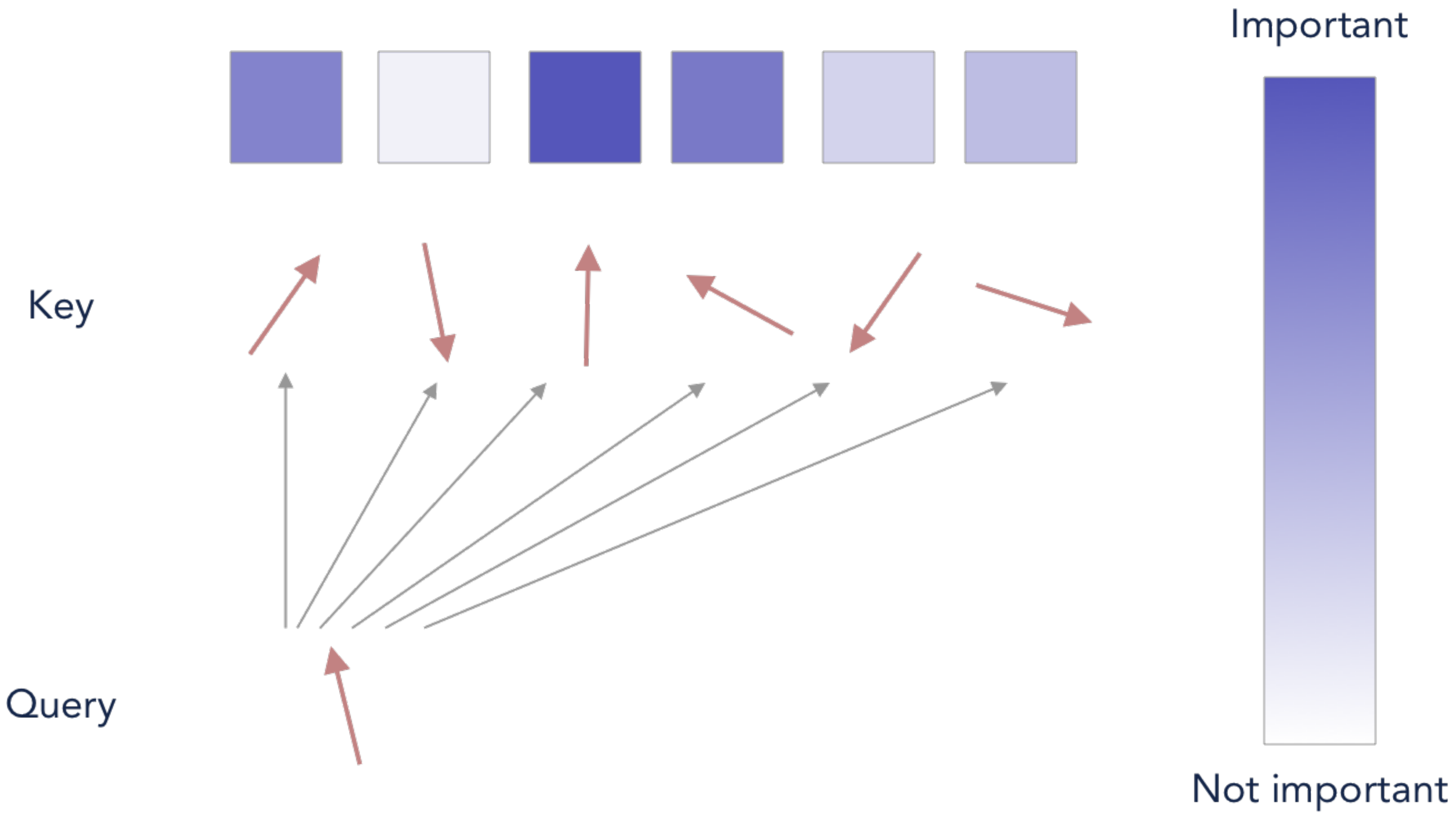
A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

How attention looks (Image source: Fig. 3 in Xu et al. 2015)

How attention works

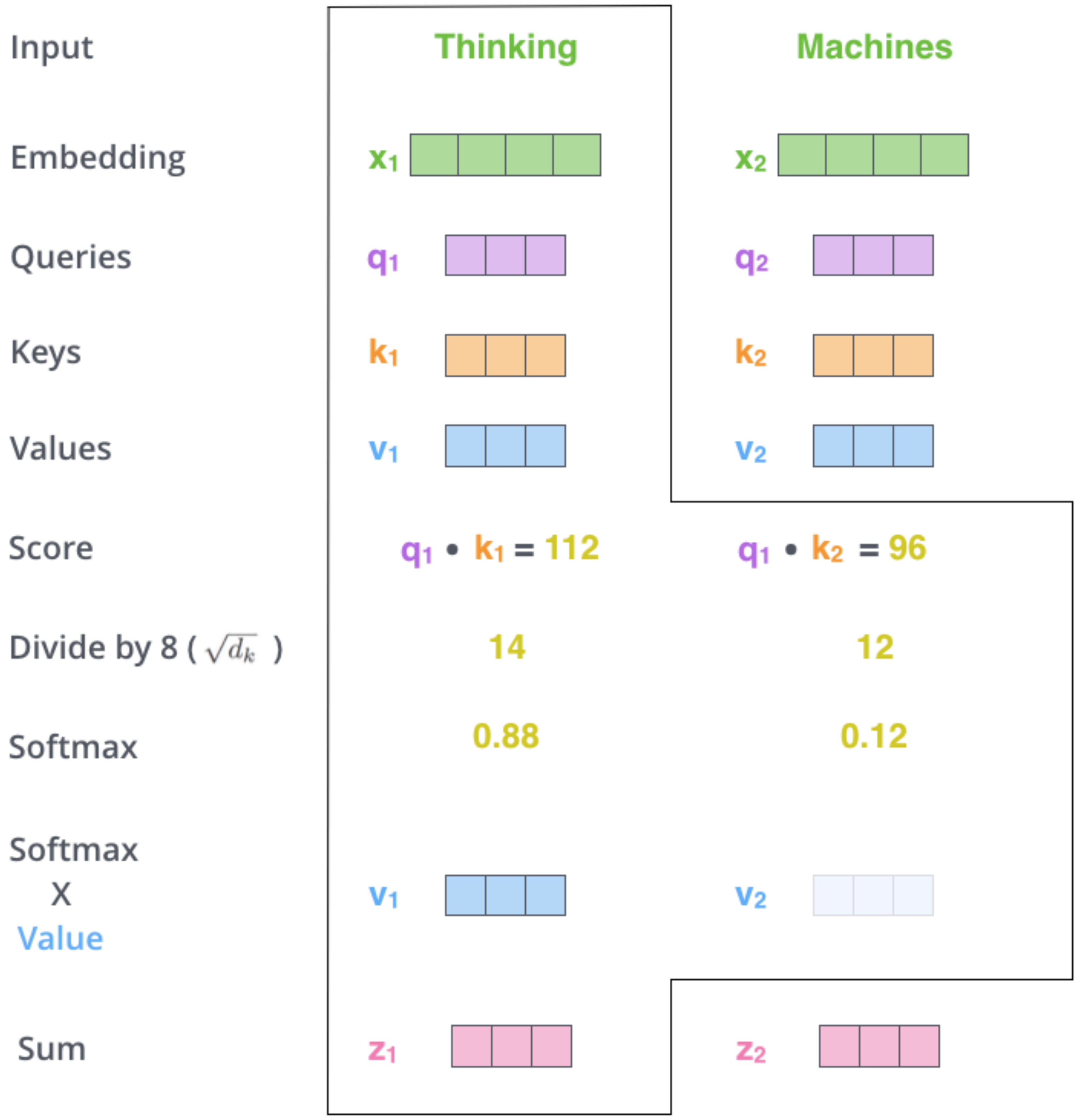


Attention mechanism core idea

LATVIJAS UNIVERSITĀTE
DATORIKAS
FAKULTĀTE

How attention works (2)

- Each entity have: key, query and value matrices;
- These matrices are calculated by multiplying the initial embedding to a trainable matrix;
- In the real models multiple attention heads are used.

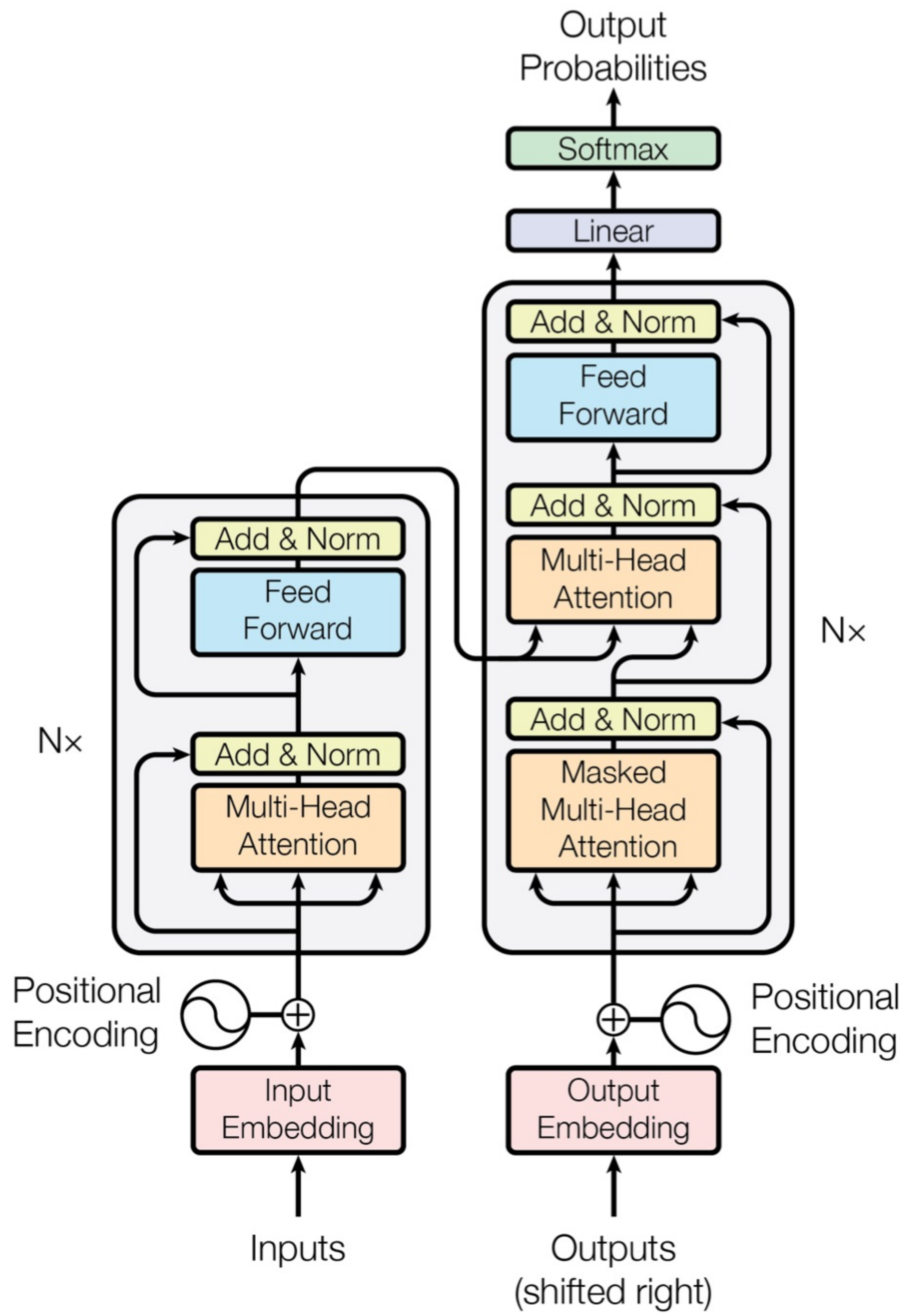


Input	Thinking	Machines
Embedding	x_1	x_2
Queries	q_1	q_2
Keys	k_1	k_2
Values	v_1	v_2
Score	$q_1 \cdot k_1 = 112$	$q_1 \cdot k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)	14	12
Softmax	0.88	0.12
Softmax X Value	v_1	v_2
Sum	z_1	z_2

(Image source: Jay Alammar, The Illustrated Transformer, 2018)

Transformer

- The main element in the architecture is the attention mechanism;
- The architecture consist of encode and decoder stacks;
- Encoders accept the input embeddings;
- Decoders accepts encoder output and previously decoded results;
- The most famous (the original) realisation is BERT.



Transformer architecture (Image source: Fig. 1 in Vaswani et al. 2017)

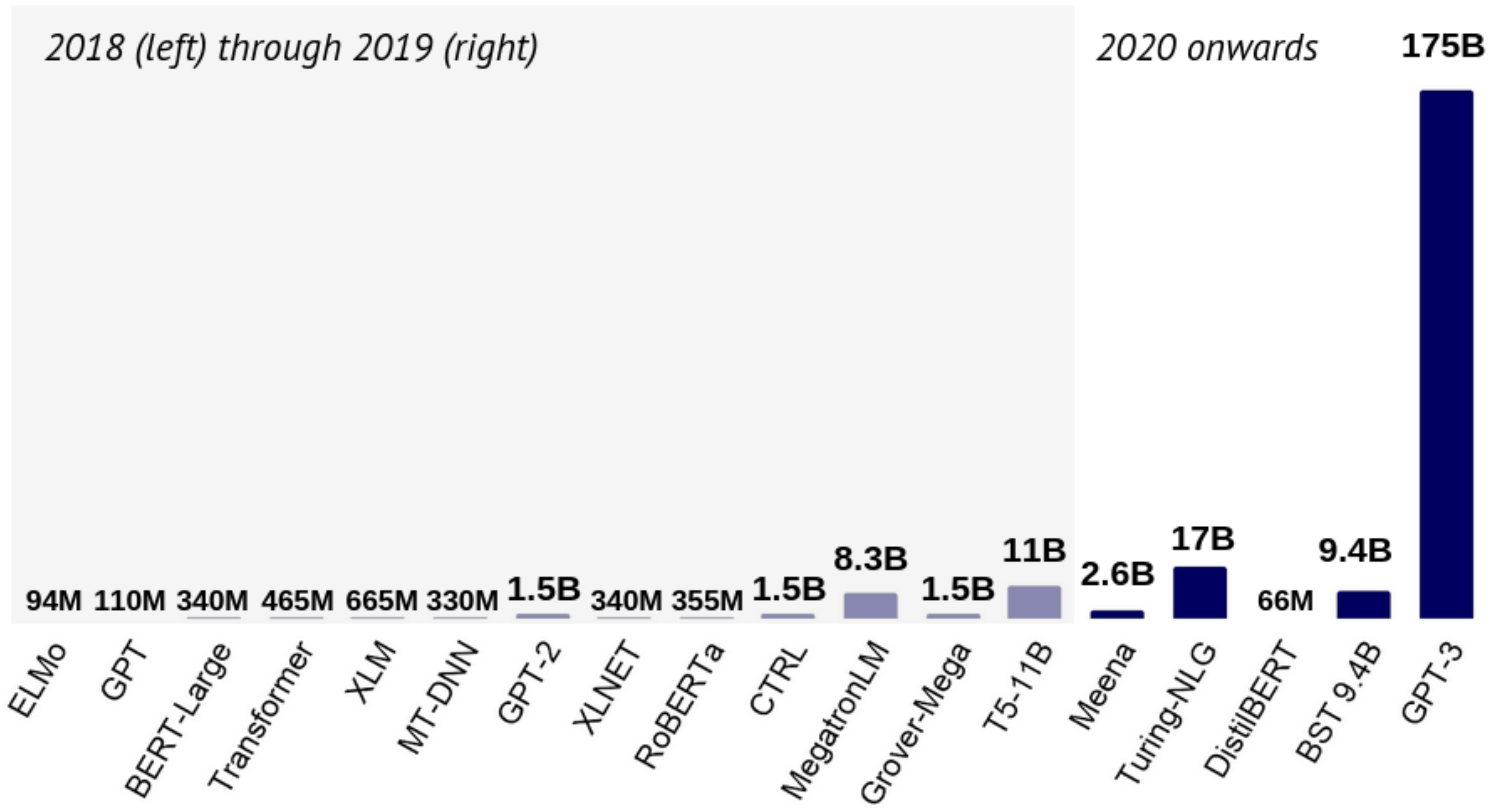
LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

GPT-2

- The model is based on the original Transformer architecture;
- The approach is to represent tasks as sequence of symbols. E.g. (translate to French, English text, French text);
- Dataset:
  - Custom built - WebText;
  - Motivated by building as large and diverse a dataset as possible;
  - The dataset contains the text subset of the 45 million links;
  - Curated by people.

GPT-3

- Uses the same architecture as GPT-2;
- Model has 175B trainable parameters.



Model size comparison (Image source: State of AI Report 2020)

minGPT

- Implementation of GPT model, used for educational purposes;
- Built on PyTorch (the original GPT-2 is built using TensorFlow);
- Available on GitHub: https://github.com/karpathy/minGPT;
- Have examples of models for text and image tasks.

Transformer biggest issues

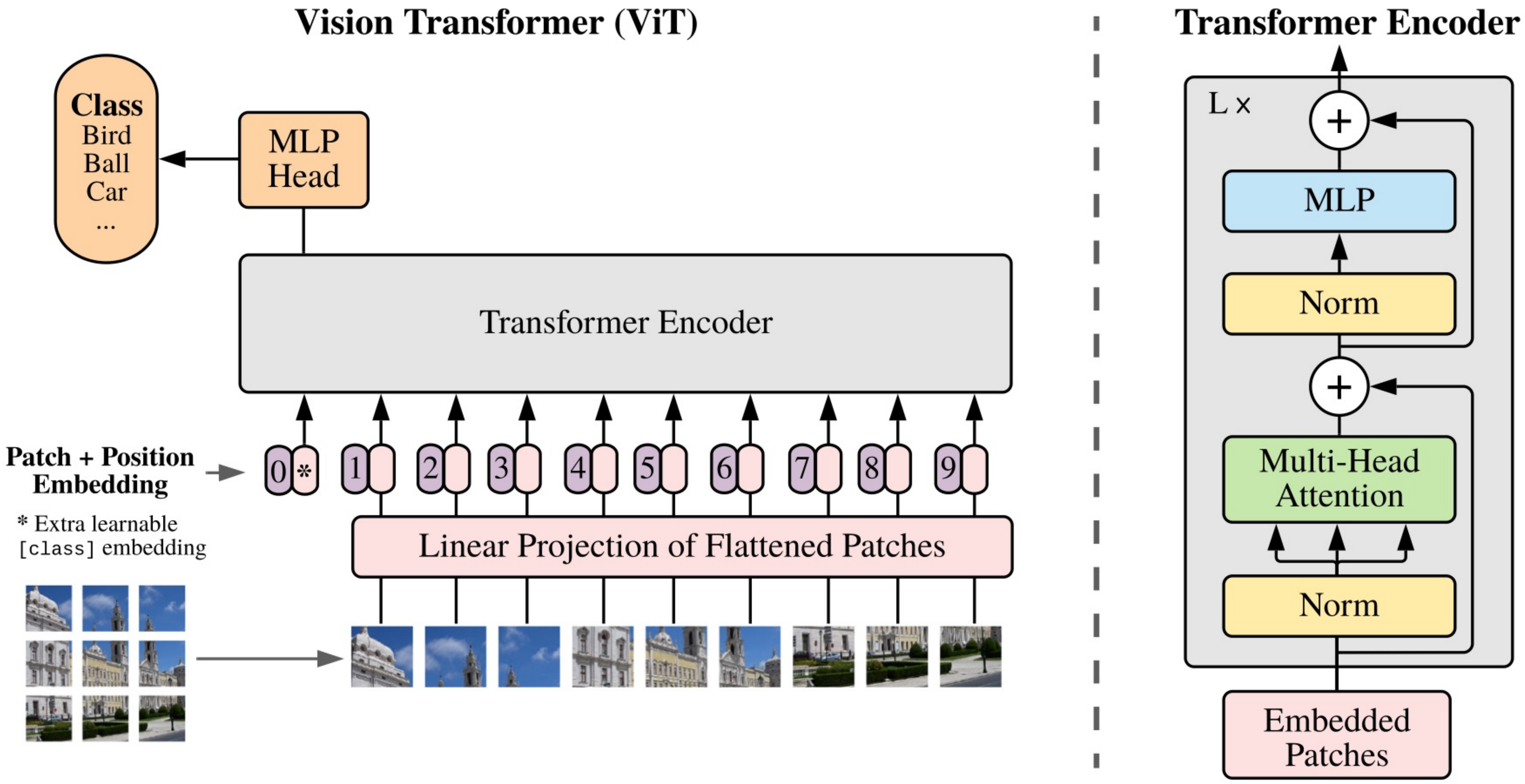
- Computation resources:
    - The original attention implementation has O(n^2) complexity;
    - The model layers take up a lot of memory;
- Attention is used to capture only temporal relations while processing input tokens in parallel. This helps to make computation more efficient, but it restricts the model from fully exploiting the sequential nature of the input. Especially if previous context matters.

Transformer applications on image tasks

ViT

- Using attention for solving visual tasks;
- The algorithm is:
    - Split an image into patches;
    - Map patches with a trainable linear projection;
    - Add positional information of every patch and prepend learnable [class] embedding;
    - Feed data into standard transformer encoder;
    - The output is mapped to multilayer perceptron which outputs class.

ViT (2)

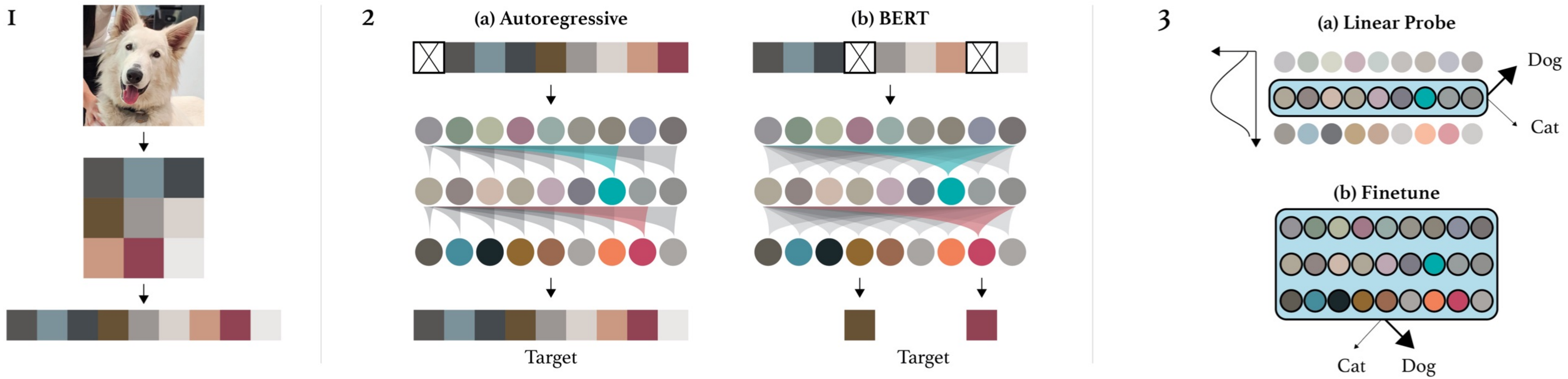


ViT architecture (Image source: Fig. 1 in Dosovitskiy et al 2020)

Image-GPT

- The architecture uses transformer decoder, the same as for GPT-2;
- The approach consist of a pre-training stage followed by a fine-tuning stage;
- In pre-training stage 2 types of tasks are used: auto-regressive and BERT;
- In fine-tuning stage image classification task is used.

Image-GPT (2)



An overview of Image-GPT approach (Image source: Fig. 1 in Chen et al 2020)

LATVIJAS UNIVERSITĀTE
DATORIKAS
FAKULTĀTE

CLIP

- A new approach to classify images;
- Instead of asking the model: "What class the image has?", researchers asking: "What probability is that the image has such class?";
- The first stage is pre-training on a huge dataset: 400 millions text-image pairs;
- Then train on a new datasets using zero-shot or a few-shot transfer method;
- The model performs very well on previously unseen datasets

DALL-E

- A new method of generating images from text;

- On stage 1, 256x256 images are compressed to 32x32 image tokens, each element of which can assume 8192 possible values;

- On stage 2, 256 BPE-encoded text is concatenated with 1024 image tokens and autoregressive transformer is trained trying to reconstruct the image.

DALL-E examples

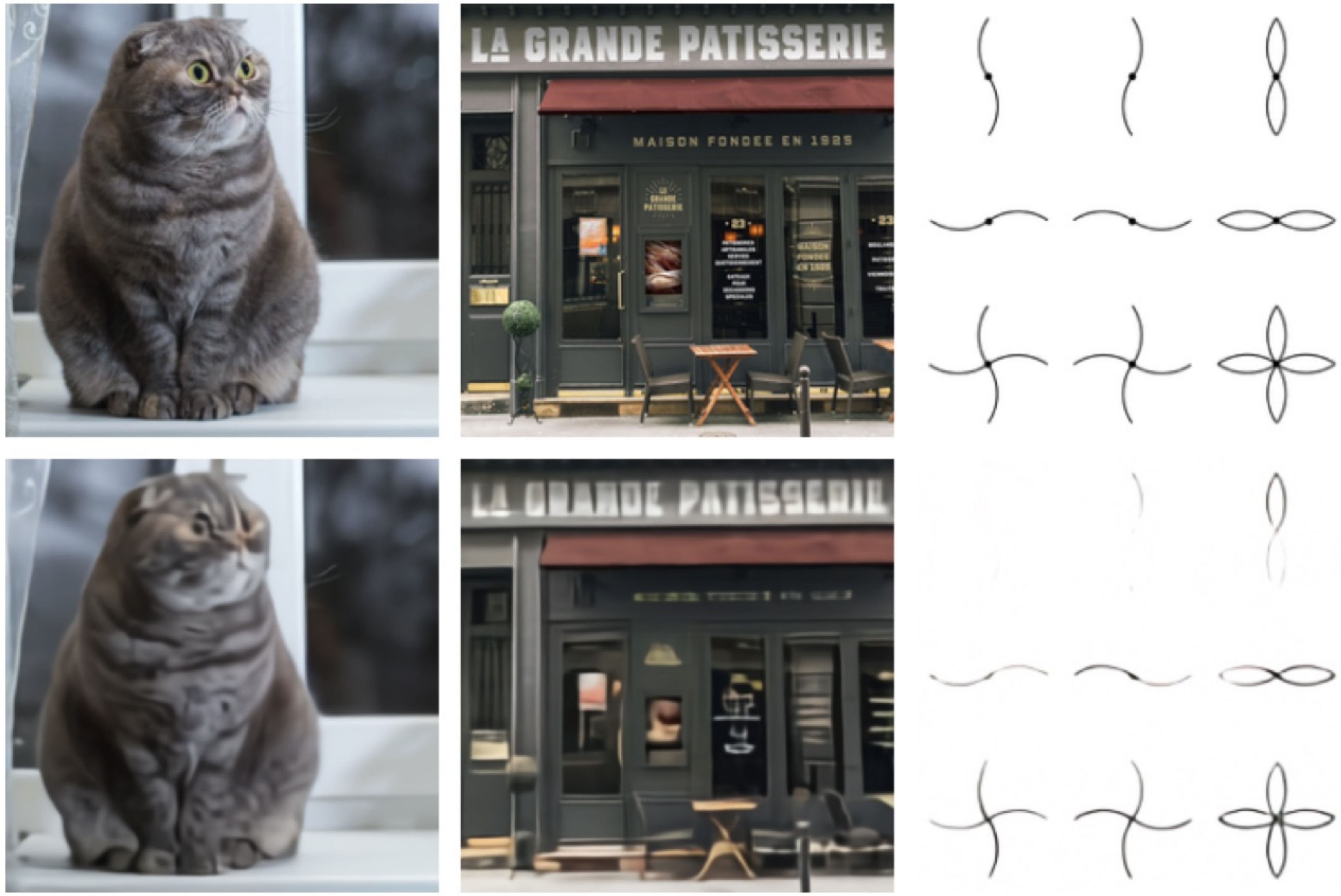
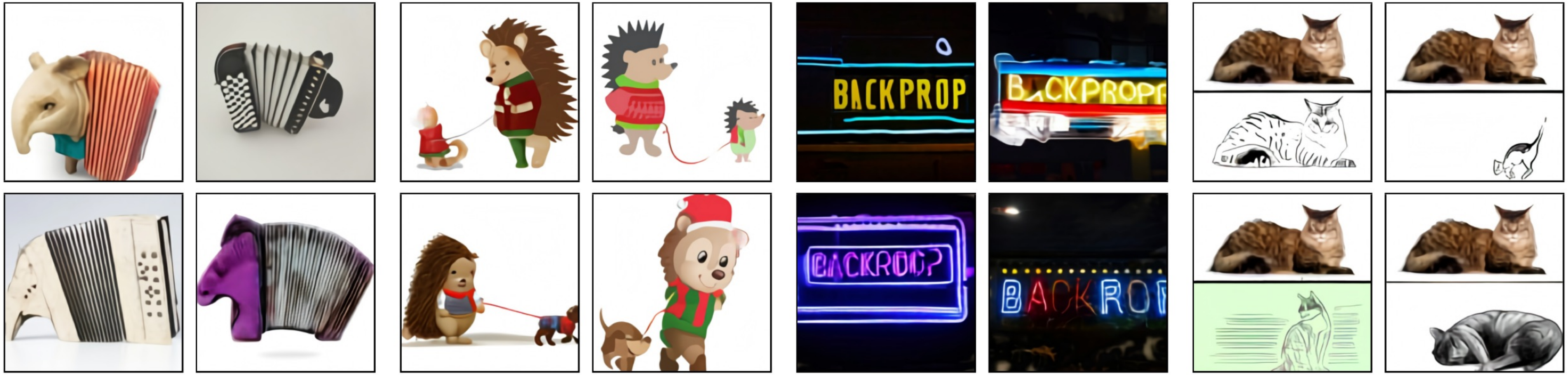


Image reconstruction from discrete VAE (Image source: Fig. 1 in Ramesh et al 2021)

DALL-E examples (2)



(a) a tapir made of accordion. a tapir with the texture of an accordion.

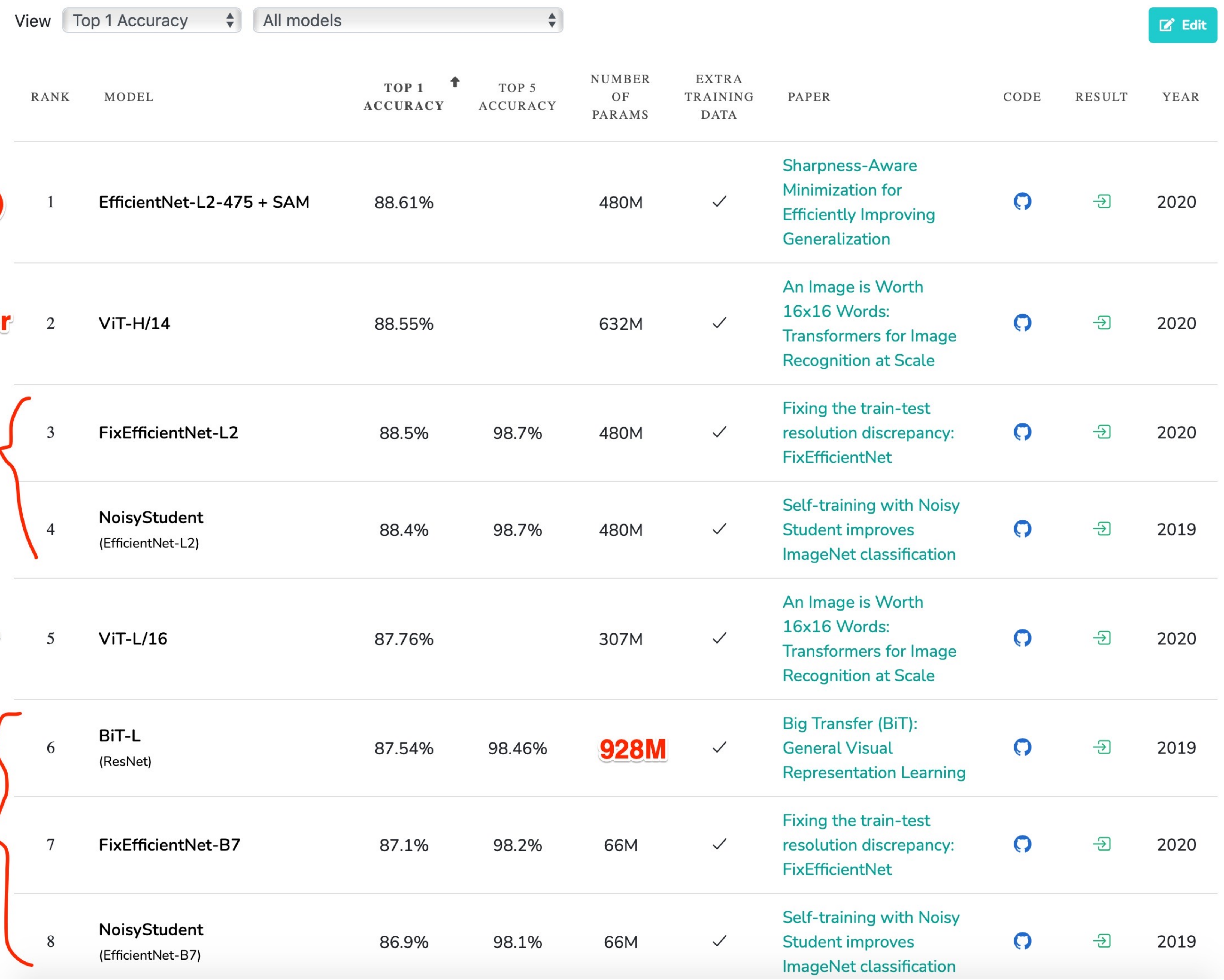
(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

(d) the exact same cat on the top as a sketch on the bottom

Generating images using annotation (Image source: Fig. 2 in Ramesh et al 2021)

Transformer vs ConvNets



View Top 1 Accuracy ▾ All models ▾ Edit

	RANK	MODEL	TOP 1 ACCURACY ↑	TOP 5 ACCURACY	NUMBER OF PARAMS	EXTRA TRAINING DATA	PAPER	CODE	RESULT	YEAR
ConvNet + New loss computing algorithm (not SGD)	1	EfficientNet-L2-475 + SAM	88.61%		480M	✓	Sharpness-Aware Minimization for Efficiently Improving Generalization			2020
Transformer	2	ViT-H/14	88.55%		632M	✓	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale			2020
ConvNet	3	FixEfficientNet-L2	88.5%	98.7%	480M	✓	Fixing the train-test resolution discrepancy: FixEfficientNet			2020
	4	NoisyStudent (EfficientNet-L2)	88.4%	98.7%	480M	✓	Self-training with Noisy Student improves ImageNet classification			2019
Transformer	5	ViT-L/16	87.76%		307M	✓	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale			2020
ConvNet	6	BiT-L (ResNet)	87.54%	98.46%	928M	✓	Big Transfer (BiT): General Visual Representation Learning			2019
	7	FixEfficientNet-B7	87.1%	98.2%	66M	✓	Fixing the train-test resolution discrepancy: FixEfficientNet			2020
	8	NoisyStudent (EfficientNet-B7)	86.9%	98.1%	66M	✓	Self-training with Noisy Student improves ImageNet classification			2019

ImageNet classification task ranking
(Image source: https://paperswithcode.com/sota/image-classification-on-imagenet
(screenshot actual for 13.12.2020))

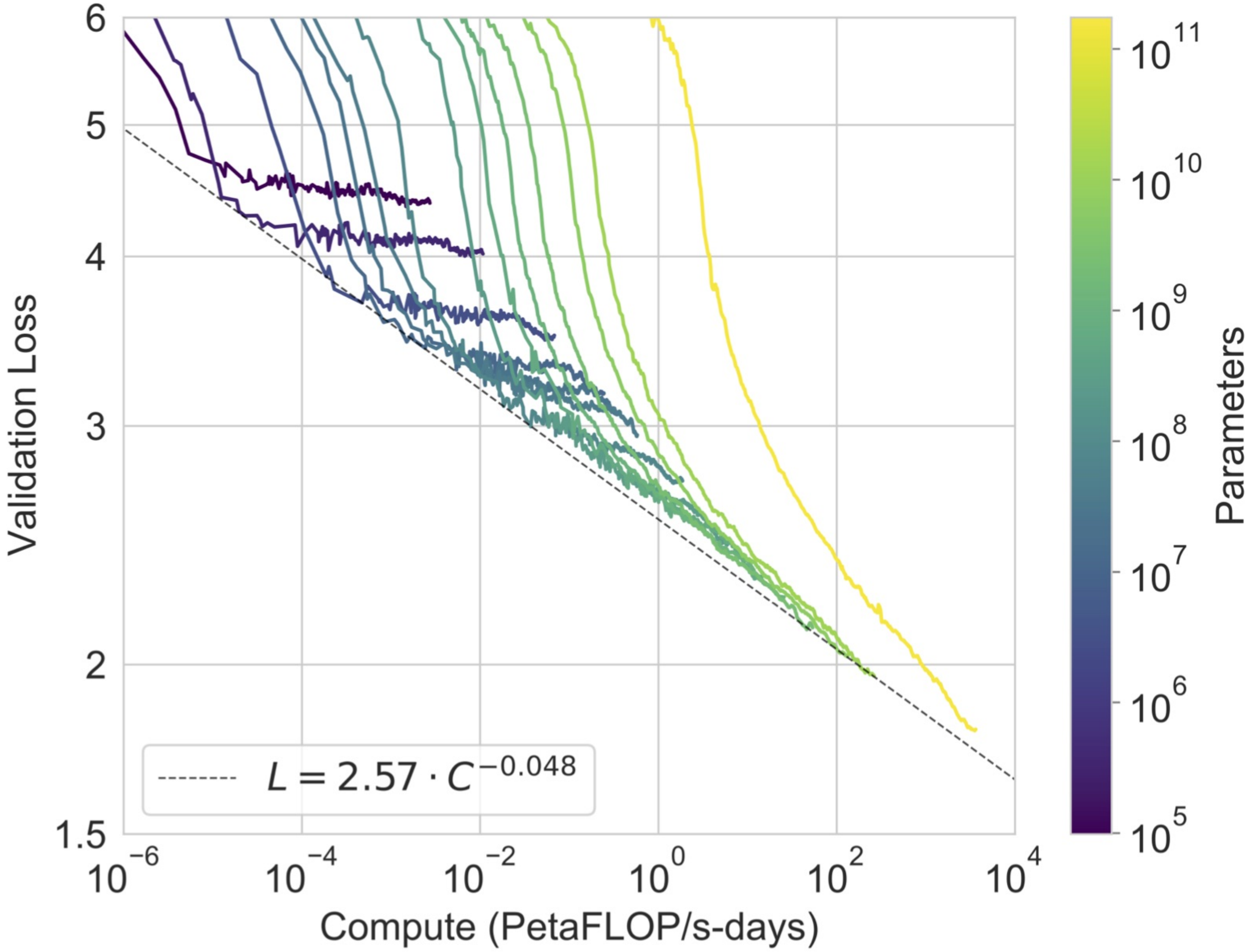
Transformer vs ConvNets (2)

- Investigating GPT-3, scientists found an interesting effect: increasing trainable parameters amount the loss function does not overfit;

- The effect leads to huge models, like GPT-3 with 1.75B parameters;

- The effect is very special to attention mechanism only and allows transformers to compete with other architectures.

Transformer vs ConvNets (3)



Loss function depending on parameters count
(Image source: Fig. 3.1 in Brown et al 2020)

LATVIJAS UNIVERSITĀTE
DATORIKAS
FAKULTĀTE

Transformer vs ConvNets (4)

Model	Parameters count	Accuracy	Epochs	Comment
Convolution network	62K	55 %	20	A default convnet from pytorch tutorial
ResNet18	11.7M	69 %	20	Smallest ResNet network
ResNet152	60.2M	33 %	10	Largest ResNet network (from the original paper)
ViT	1.3M	54 %	20	
ViT	11.6M	54 %	20	
ViT	89.8M	54 %	10	ViT Large from the original paper

Comparing ViT and ResNet with a different size and how it affects the accuracy

Current progress

- Working on BPE tokeniser and a new model for minGPT;
- Investigating new areas and approaches where transformers could be applied and comparing with existing solutions (TextWorld, image generation);

Thank you for your attention!