NORTHEASTERN UNIVERSITY 18 98 LVX VERITAS VIRTVS

IMAGE CAPTIONING WITH GIT MODEL

DATA ANALYTICS FOR GOOD CASE COMPETITION

TEAM TOMATO JUICE:

ZIWEI DUAN, ROBIN CHEN, JINCHUAN HE;

KEY BUSINESS PROBLEMS

BACKGROUND

Language is essential to human life.

The languages we speak or sign are at the very core of our human identity and integral to our ability to flourish in life.

New text-to-image and image-to-text models are taking the AI world by storm. One related task, which has direct business impact, is called "**image captioning**." This task takes an image as input and generates a text caption as output. Businesses can utilize image captioning models to:

- Create HTML header and alt text content for images on their website, which boosts search engine scoring and user acquisition
- Tag user-submitted images with relevant text descriptions for improved filtering and search
- Ensure that images submitted to websites match tagged categories and/or do not violate terms of use (e.g., containing content in protected categories or content used to spread misinformation).

BUSINESS PROBLEM

Problem

How to make the image-to-text for non-dominant language?

Challenge

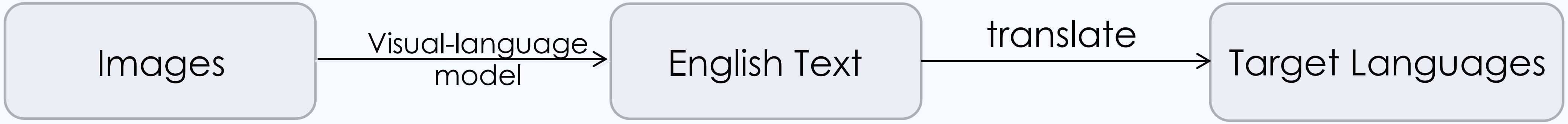
Most image captioning models only built for a handful of the popular language

Impact

If a company just captures the image in one language,
the company cannot connect more target markets.

BUSINESS PROBLEM ANALYSIS

- Existing Image-to-text models have thrived in limited languages, including English.



Images

Visual-language model

English Text

translate

Target Languages

METHODOLOGY

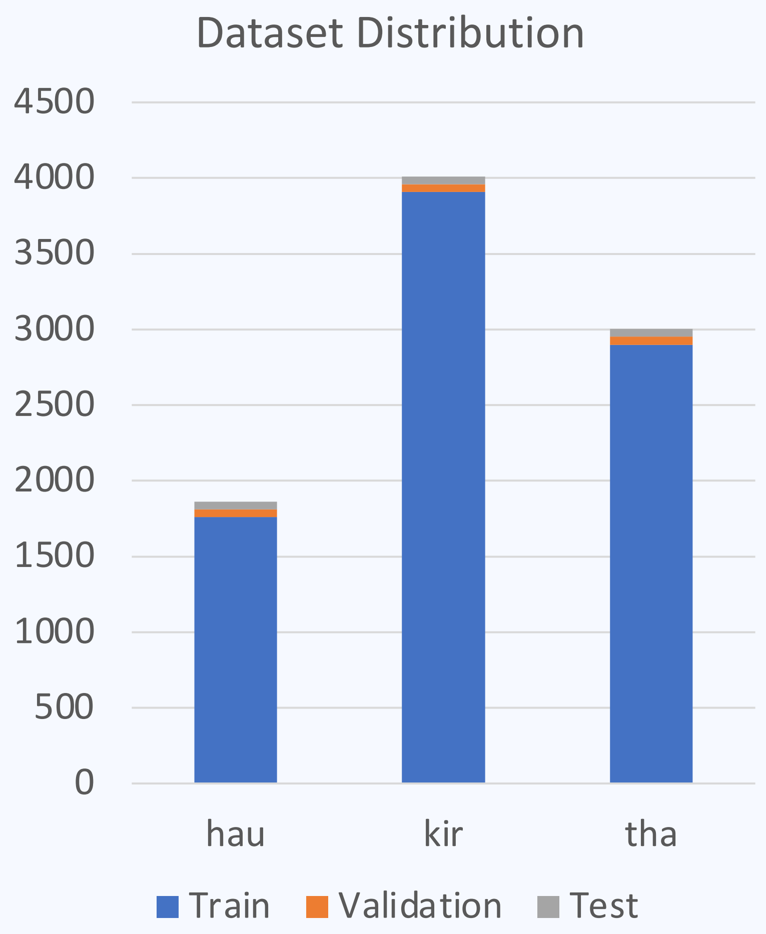
DATA RESOURCE

- SIL International is a global, faith-based nonprofit institution that provide reading material and education opportunity around the world.
- One of their project is **Bloom library,** it is the only open-source book writing software that can be used offline and in any language--including sign languages.
- The purpose of this tool is to solve the lack of appropriate and engaging children's reading materials for non-dominant languages.

https://bloomlibrary.org/landing

DATASET

- Hausa Language: 1866 images
- Kyrgyz Language: 4027 images
- Thai Language: 3024 images
- Test Datasets: 200 images



Generative Image-to-Text Transformers[1]

GIT'S MULTI-TASK CAPABILITY

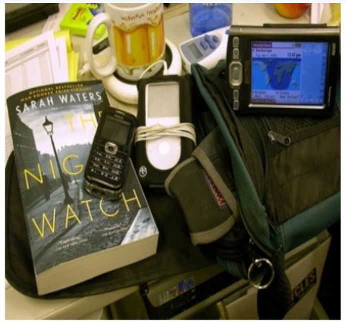
Image captioning:

A microsoft store in the mall.

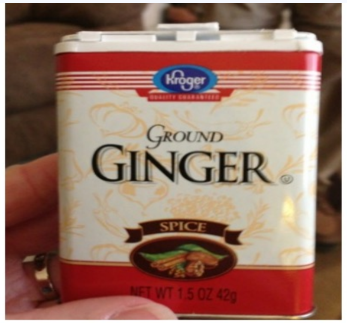
Bart simpson is shown in a scene from the simpsons.

The colosseum is lit up at night with a fence in the background.

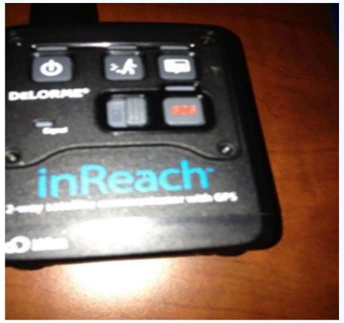
Question answering:



Q: What is the person's first name at the top of the book?
Pred: sarah



Q: WHICH BRAND IS IT
Pred: kroger



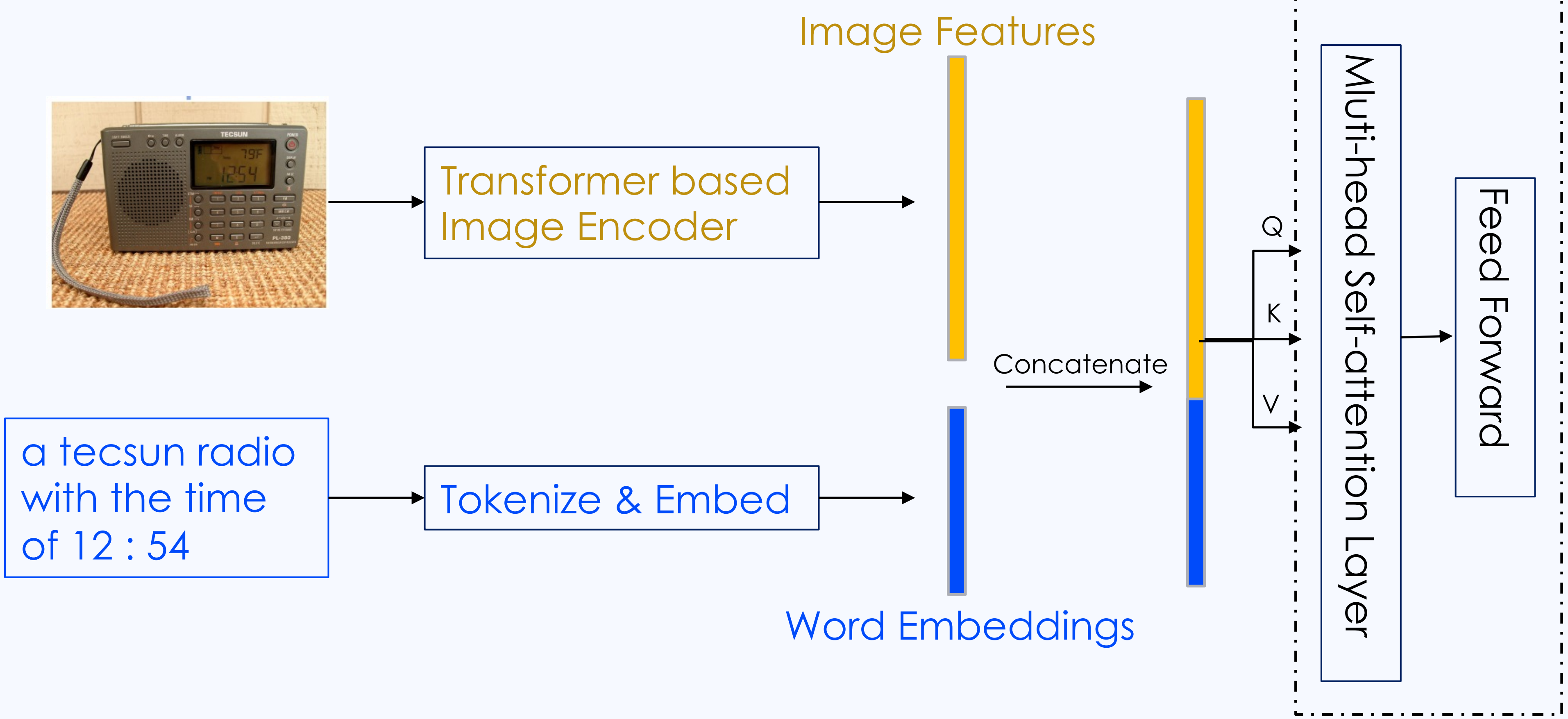
Q: what is written in blue color?
Pred: inreach

Jianfeng Wang, et al. GIT: A Generative Image-to-text Transformer for Vision and Language. arXiv preprint arXiv:2205.14100v4, 2022.

Introduction to deep learning, read more.

Introduction to Transformers, learn more.

MODEL ARCHITECTURE / PRETRAINING



Text Decoder

Image Features

Transformer based Image Encoder

a tecsun radio with the time of 12 : 54

Tokenize & Embed

Word Embeddings

Concatenate

Q

K

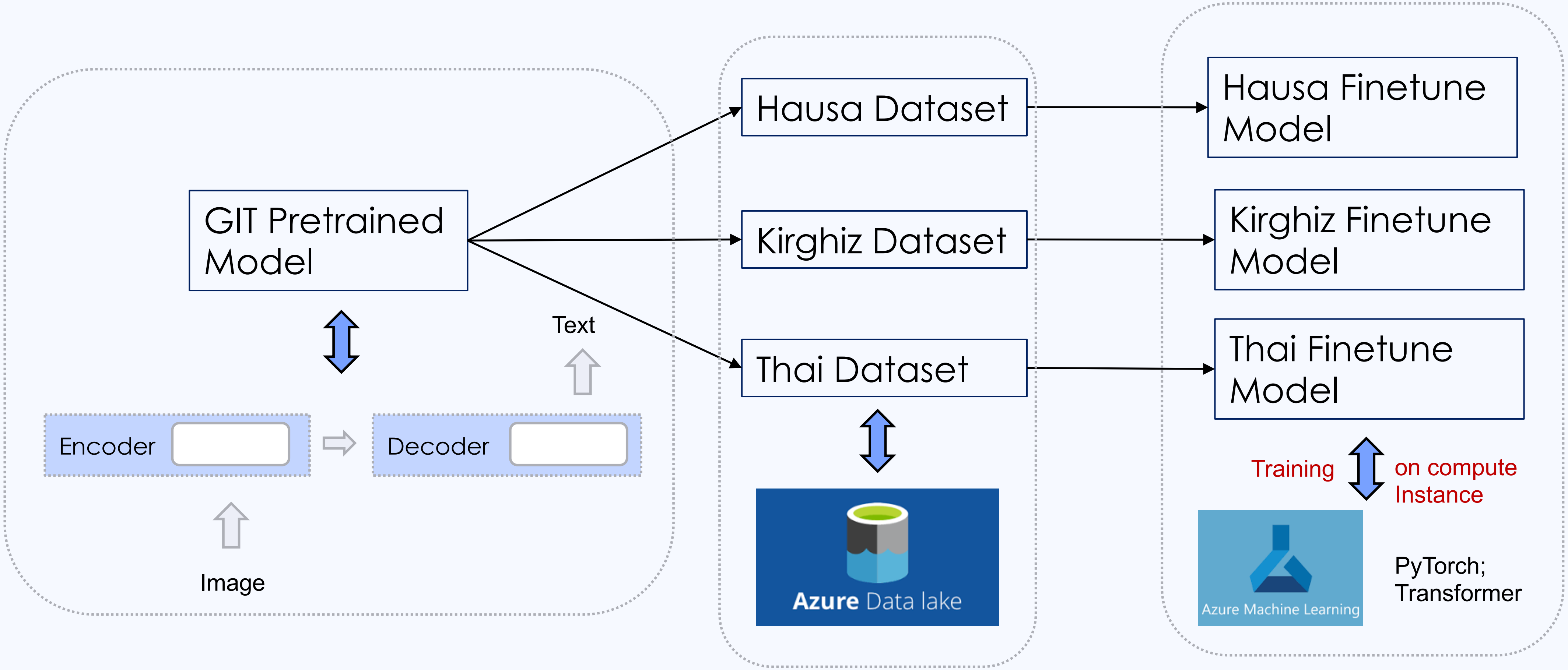
V

Mluti-head Self-attention Layer

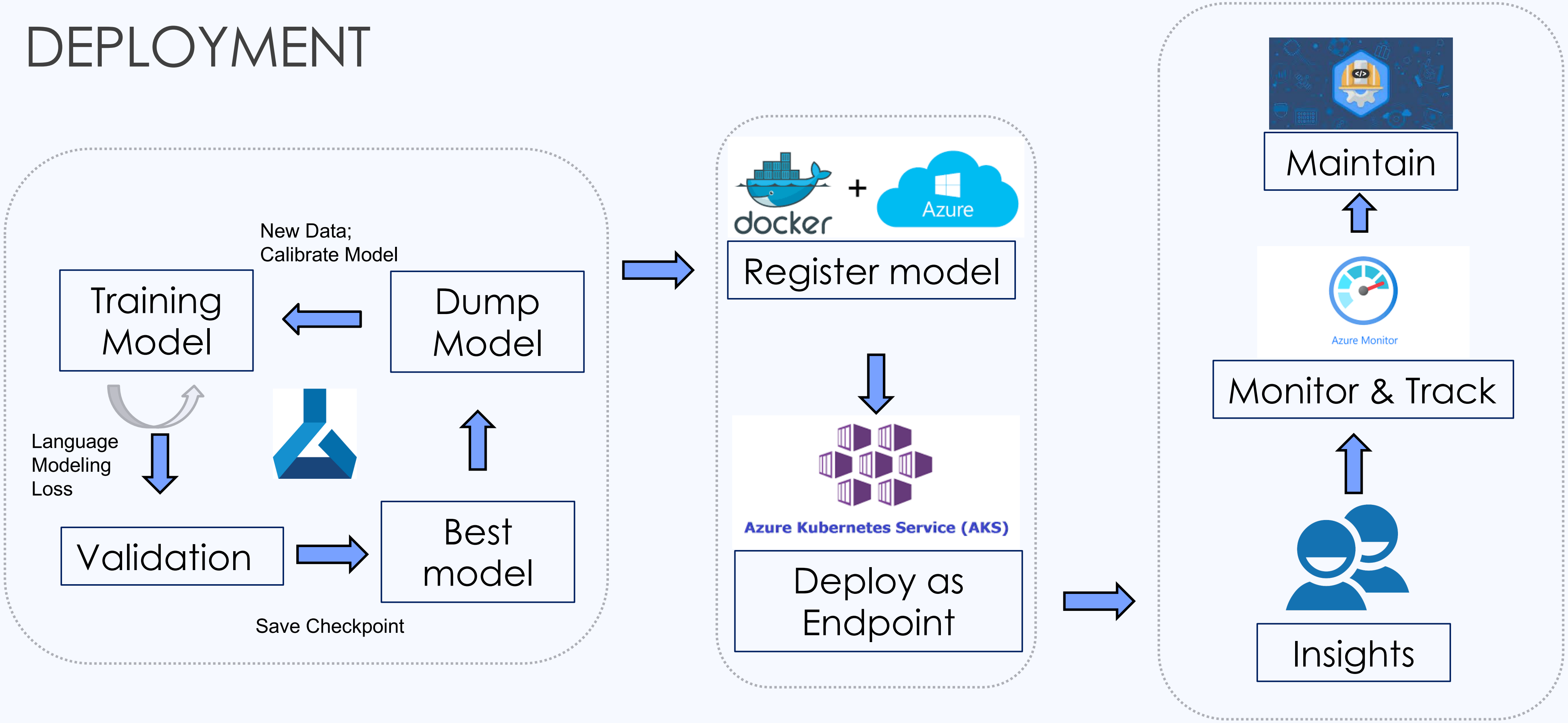
Feed Forward

MODELING & DEPLOYMENT

MODEL FINE-TUNNING



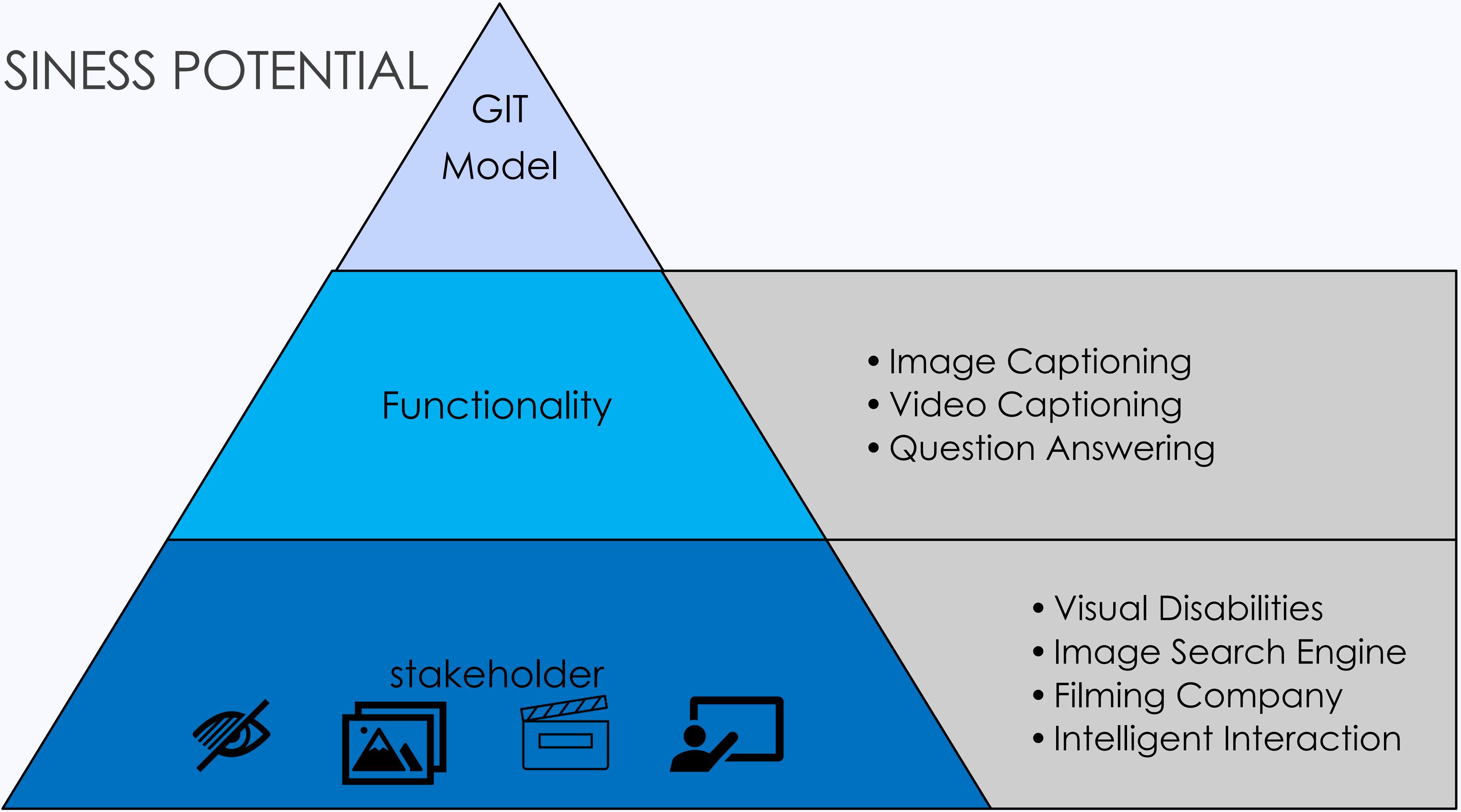
DEPLOYMENT



New Data;
Calibrate Model
Training Model
Dump Model
Language Modeling Loss
Validation
Best model
Save Checkpoint
docker
+
Azure
Register model
Azure Kubernetes Service (AKS)
Deploy as Endpoint
Maintain
Azure Monitor
Monitor & Track
Insights

TAKEAWAYS

BUSINESS POTENTIAL



GIT Model

Functionality

- Image Captioning
- Video Captioning
- Question Answering

stakeholder

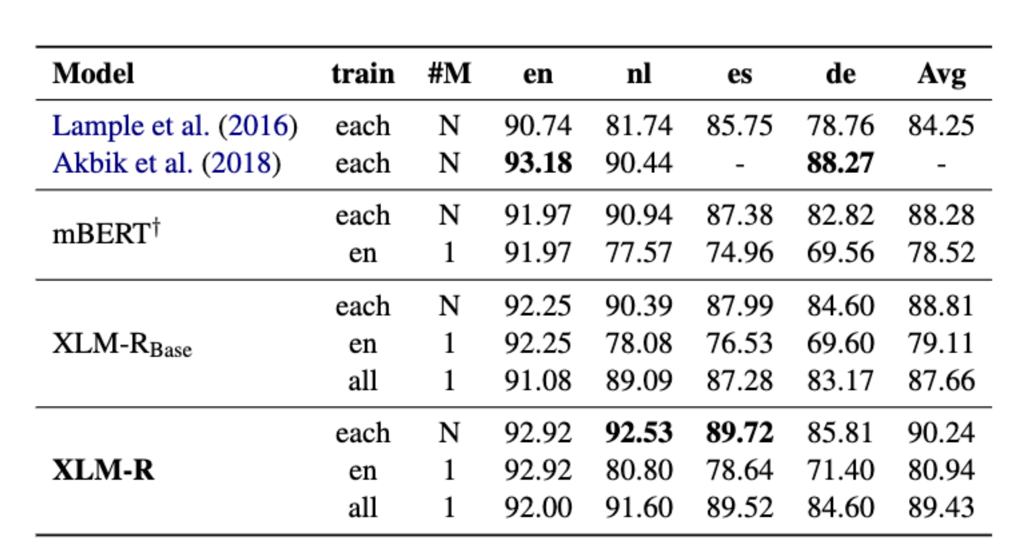
- Visual Disabilities
- Image Search Engine
- Filming Company
- Intelligent Interaction

FUTURE DEVELOPMENTS

- Create a universal(monolingual) sentence encoder with a shared embedding space across languages.

- Currently the most of art of state model is XLM-R[3], which is capable to solve about hundreds languages.

- With this universal encoder, we can training model on dominant language and predict on minor languages.



Model	train	#M	en	nl	es	de	Avg
Lample et al. (2016)	each	N	90.74	81.74	85.75	78.76	84.25
Akbik et al. (2018)	each	N	**93.18**	90.44	-	**88.27**	-
mBERT†	each	N	91.97	90.94	87.38	82.82	88.28
	en	1	91.97	77.57	74.96	69.56	78.52
XLM-R$_{\text{Base}}$	each	N	92.25	90.39	87.99	84.60	88.81
	en	1	92.25	78.08	76.53	69.60	79.11
	all	1	91.08	89.09	87.28	83.17	87.66
XLM-R	each	N	92.92	**92.53**	**89.72**	85.81	90.24
	en	1	92.92	80.80	78.64	71.40	80.94
	all	1	92.00	91.60	89.52	84.60	89.43

REFERENCE

[1]Jianfeng Wang, et al. GIT: A Generative Image-to-text Transformer for Vision and Language. arXiv preprint arXiv:2205.14100v4, 2022.

[2] Lample, Guillaume, and Alexis Conneau. “Cross-lingual language model pretraining.” arXiv preprint arXiv:1901.07291 (2019).

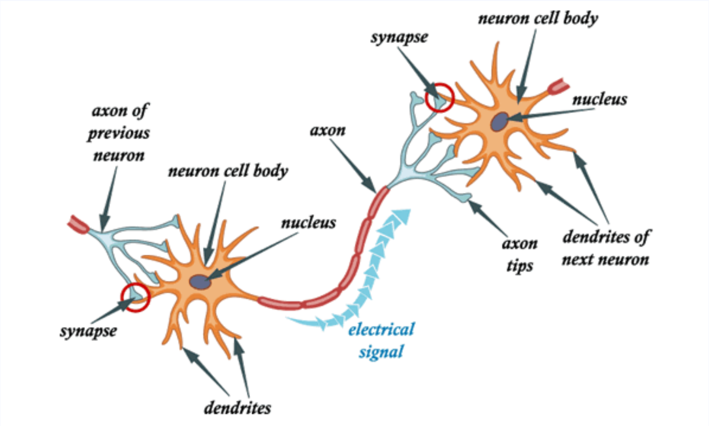
[3] Conneau et al. Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020)

Q&A

Ziwei; Jinchuan; Chushi

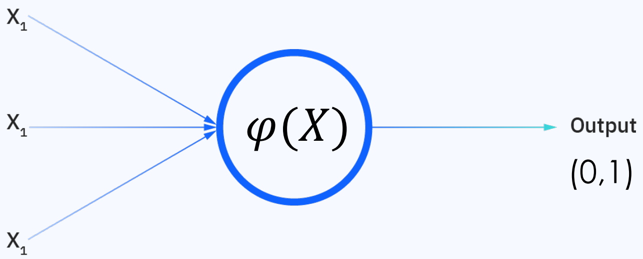
Appendix

Artificial neurons are inspired by the biological neurons that are found in our brains. In fact, the artificial neurons simulate some basic functionalities of the neurons in our brains, but in a very simplified way.



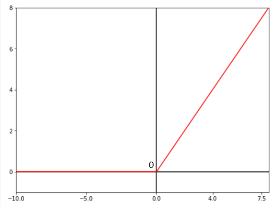
Biological Neuron

Simplify

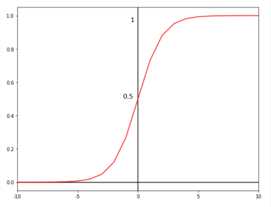


Artificial Neuron

$\varphi(X)$

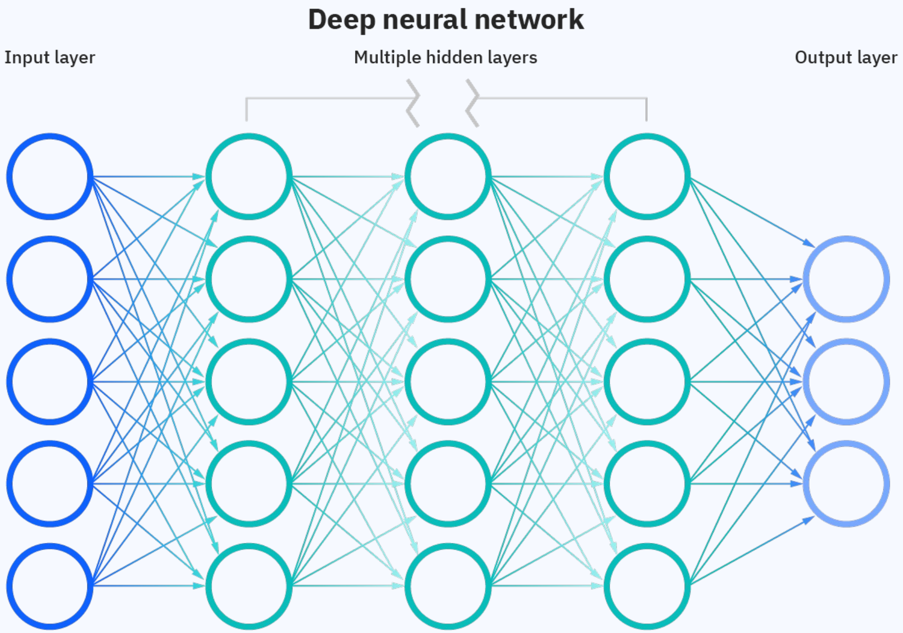


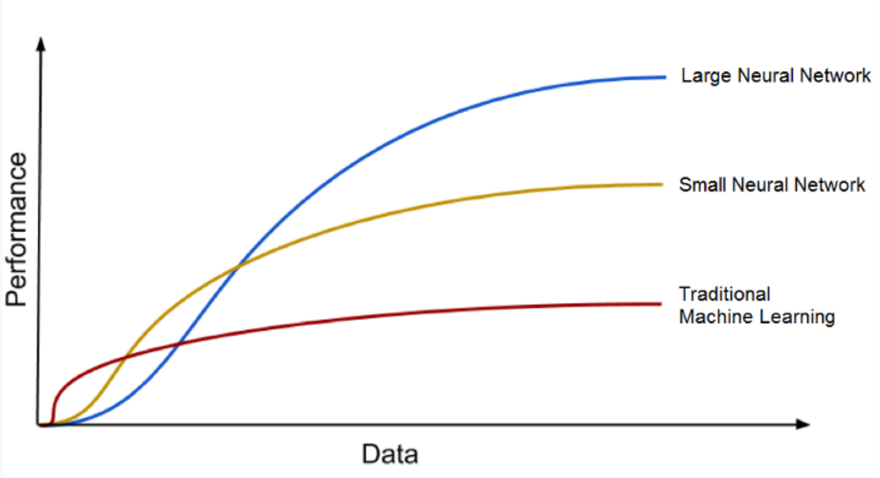
ReLU



Sigmoid

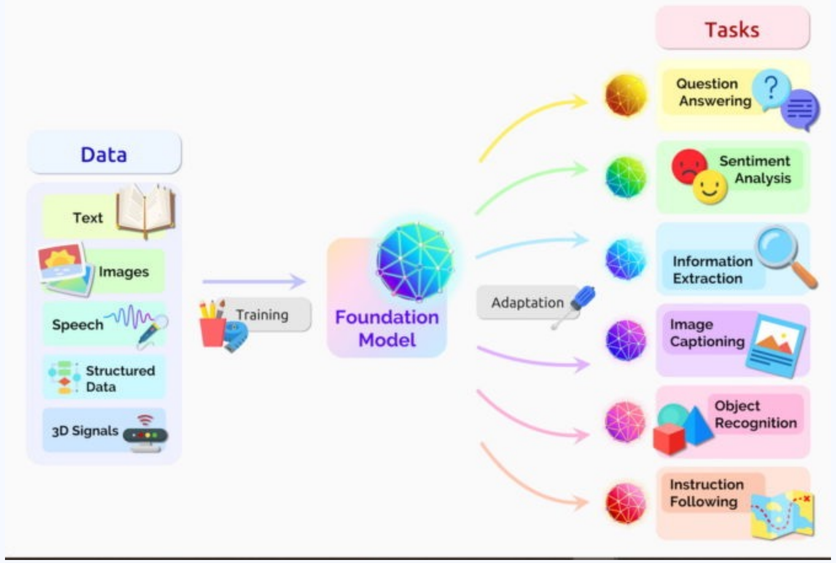
- A neural network generally consists of a collection of connected neurons. These artificial neurons loosely model the biological neurons of our brain.
- As the layers go deeper and deeper, the model getting more powerful. With different connection structure, the model can solve various problem.
- As network keep growing and developing into many different variants, it independent from **Machine Learning** and people named is as **Deep Learning**.





Back

- A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of natural language processing (NLP)and computer vision (CV).



Transformers, sometimes called foundation models

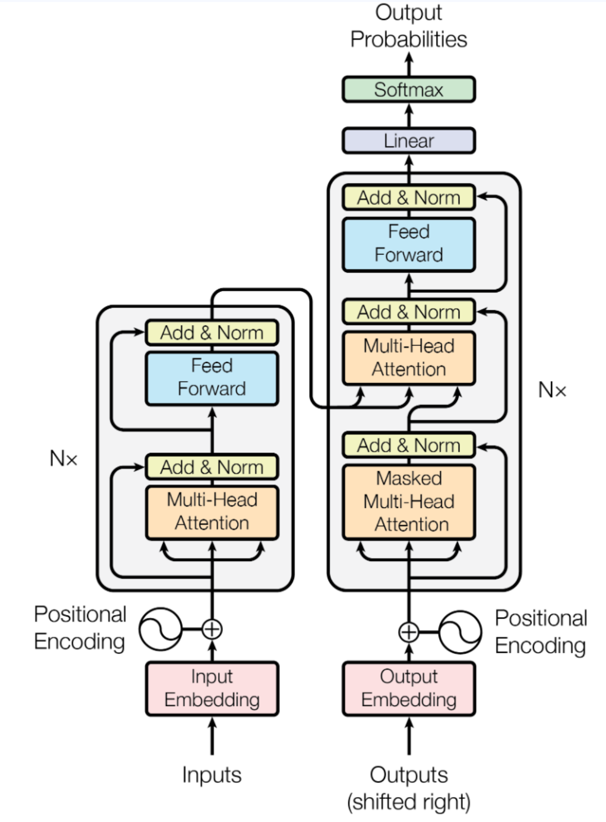


Figure 1: The Transformer - model architecture.

Back

Sufeng Duan and Hai Zhao. 2020. **Attention Is All You Need** for Chinese Word Segmentation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3862–3872, Online.