北京航空航天大學
BEIHANG UNIVERSITY

BIGSCITY
Beihang Interest Group on SmartCity

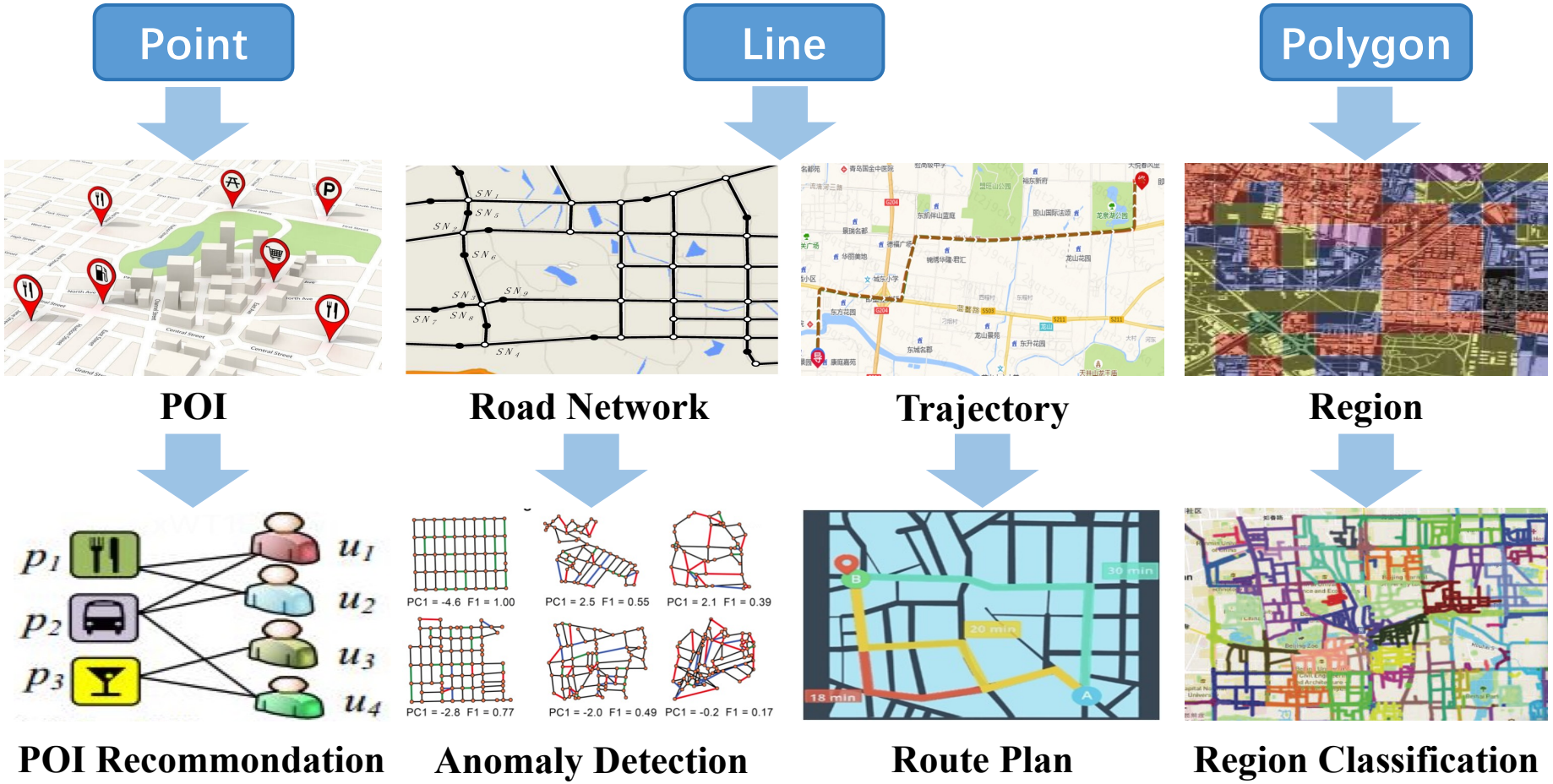
Spatio-Temporal Data and Foundation Model

Jingtian Ma

Background

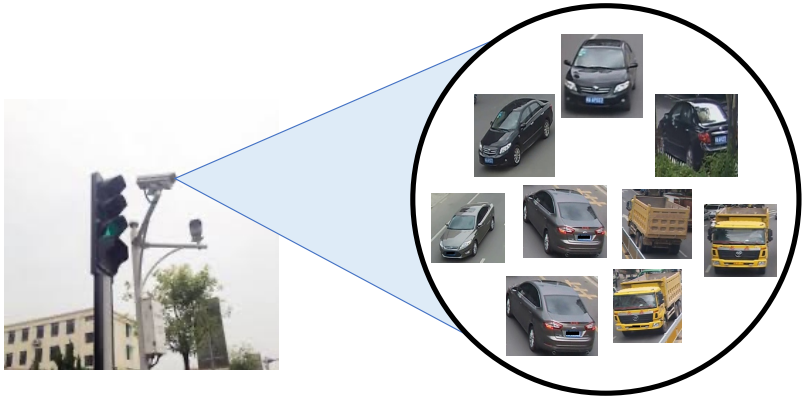
Background

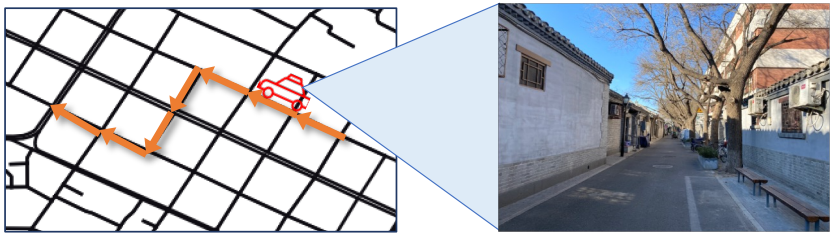
- **Spatio-Temporal Data in Traffic Domain**



Background

- **Visual-Textual Data in Traffic Domain**



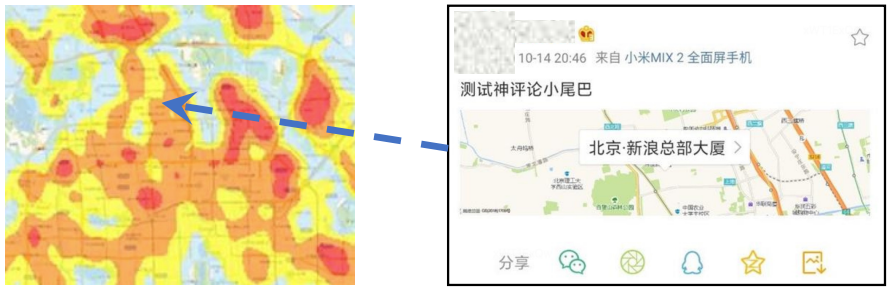


Captured Image Data

天安门（*Tian'anmen*），坐落在中华人民共和国首都北京市的中心、故宫的南端，与天安门广场以及人民英雄纪念碑、毛主席纪念堂、人民大会堂、中国国家博物馆隔长安街相望，占地面积4800平方米，以杰出的建筑艺术和特殊的政治地位为世人所瞩目。

天安门是明清两代北京皇城的正门，始建于明朝永乐十五年（*1417年*），最初名"承天门"，寓"承天启运、受命于天"之意。设计者为明代御用建筑匠师蒯祥。清朝顺治八年（*1651年*）更名为天安门。由城台和城楼两部分组成，有汉白玉石的须弥座，总高34.7米。天安门城楼长66米、宽37米。城台下有券门五阙，中间的券门最大，位于北京市皇城中轴线上，过去只有皇帝才可以由此出入。正中门洞上方悬挂着毛泽东画像，两边分别是"中华人民共和国万岁"和"世界人民大团结万岁"的大幅标语。[1]

1925年10月10日，故宫博物院成立，天安门开始对民众开放。1949年10月1日，在这里举行了中华人民共和国开国大典，由此被设计入国徽，并成为中华人民共和国的象征。1961年，中华人民共和国国务院公布为第一批全国重点文物保护单位之一。[2]



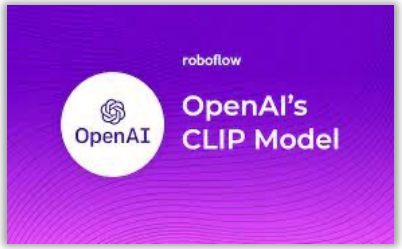
Text Description Data

Background

Foundation Models

A **Foundation Model** is "any model that is trained on **broad data** (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to **a wide range of downstream tasks**", such as BERT [Devlin et al. 2019], GPT-3 [Brown et al. 2020], and CLIP [Radford et al. 2021].

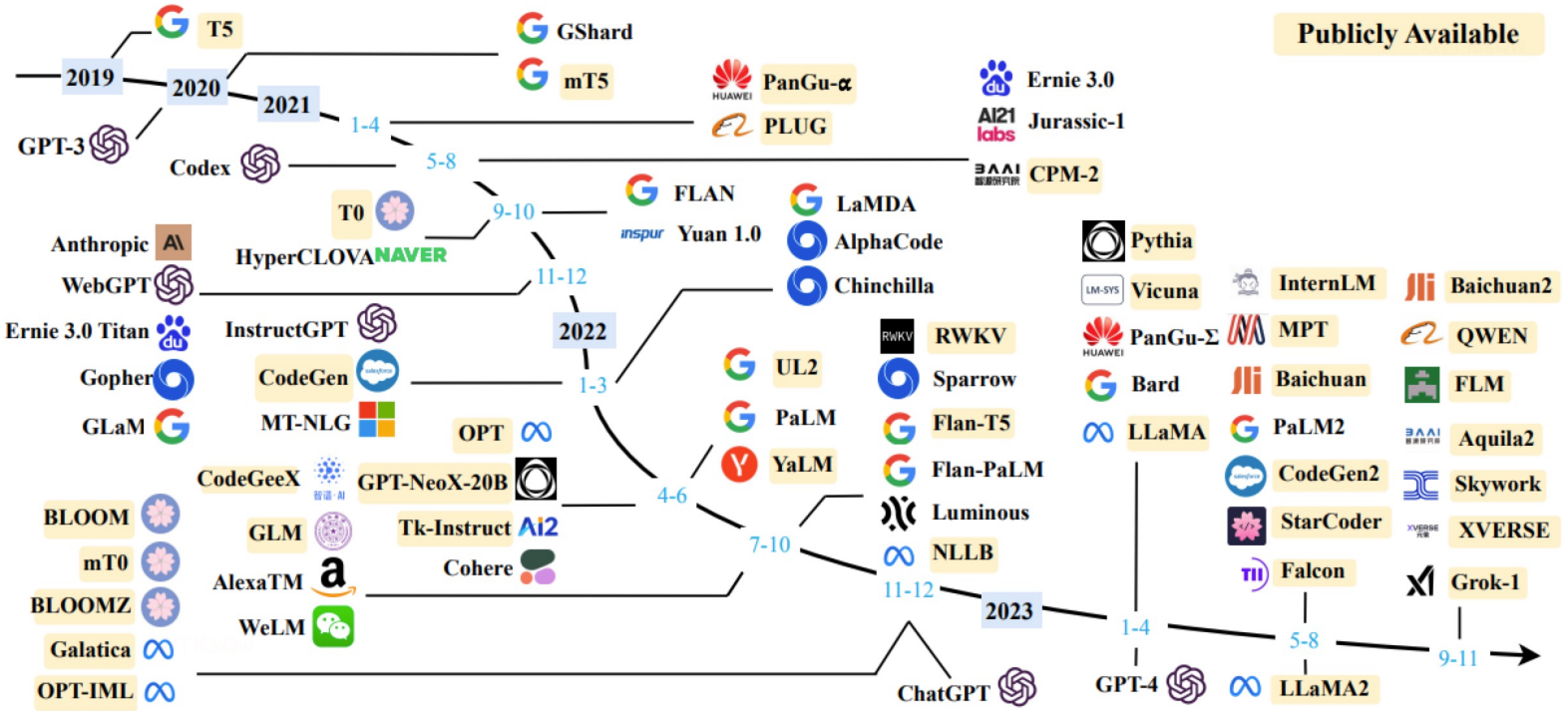
——[On the Opportunities and Risks of Foundation Models 2021 Li Fei-Fei etc.]



Development of Foundation Models

Development of Foundation Models

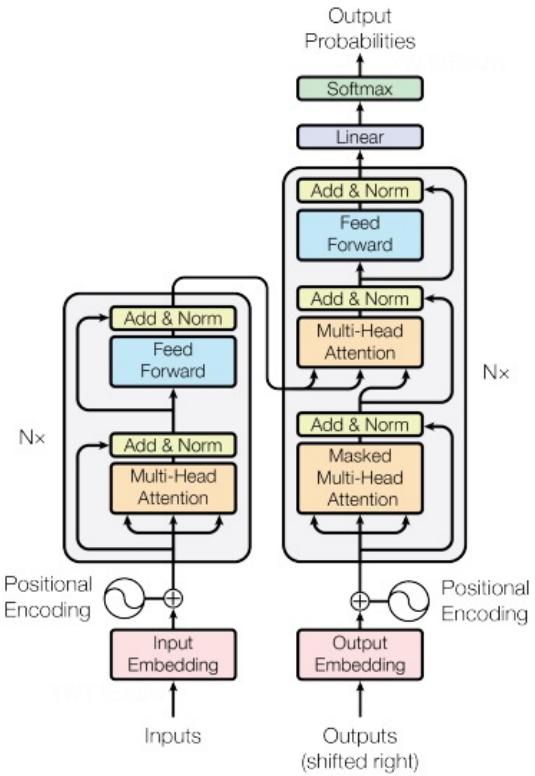
- **Large Language Models (LLMs)**



[A Survey of Large Language Models, 2023]

Development of Foundation Models

- **Large Language Models (LLMs)**



Transformer

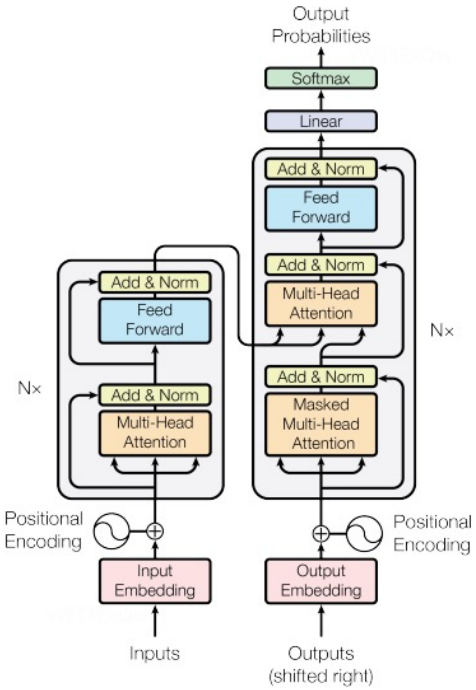
Feature	BERT	GPT	T5
Model Type	Bidirectional encoder	Unidirectional decoder	Encoder-decoder structure
Task Domain	Mainly used for downstream tasks such as sentiment analysis, named entity recognition, etc.	Mainly used for generative tasks such as text generation, dialogue systems, etc.	Universal sequence-to-sequence tasks, supports various tasks like translation, summarization, etc.
Pretraining Task	Masked Language Model (MLM)	Autoregressive Language Model	Sequence-to-Sequence (seq2seq)
Context Understanding	Bidirectional encoding, understands the entire context information	Unidirectional decoding, understanding of the current token relies on previous tokens	Encoder-decoder structure, capable of processing global context information
Advantages	Strong context understanding, applicable to a variety of tasks	Generates coherent, context-aware text	Universal sequence-to-sequence structure, supports multiple tasks, easy to fine-tune
Disadvantages	High computational resource requirements, sensitive to sequence length	Limited understanding of context, inability to capture global information	Sensitive to sequence length, high computational resource requirements

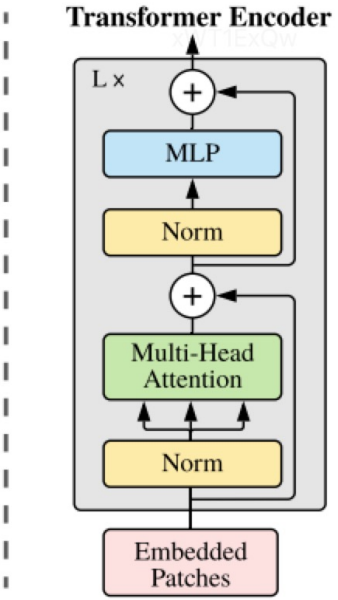
[Attention Is All You Need, 2017]

Development of Foundation Models

Large Vision Models (LVMs)

Vision Transformer (ViT): Addressing tasks in the field of computer vision using the standard Transformer.



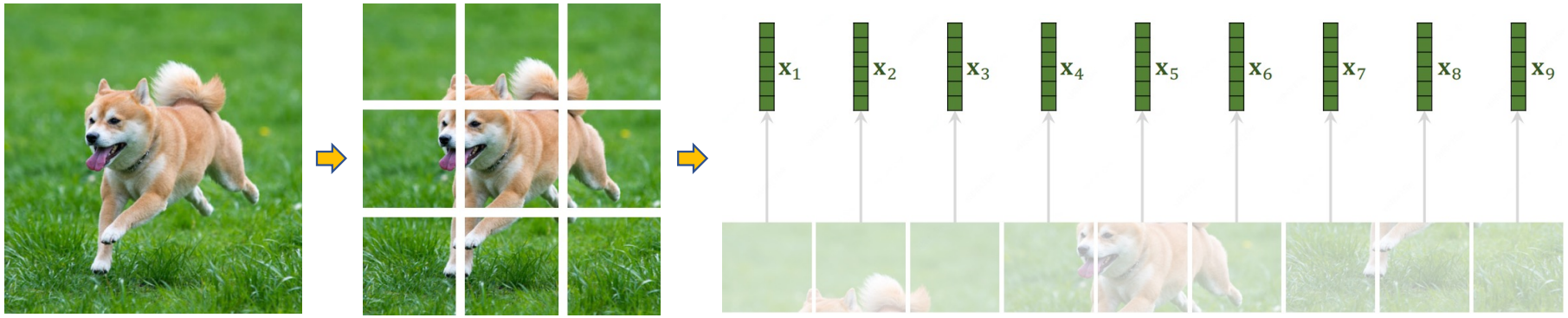


[An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021]

Development of Foundation Models

Large Vision Models (LVMs)

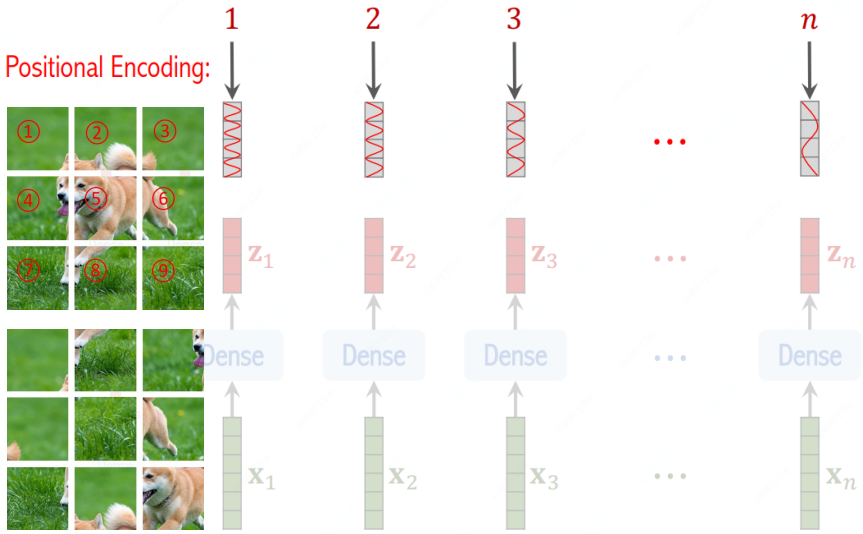
- ViT – Linear Projection of Flattened Patches
  - Split Image into Patches: patch 16 x16
  - Vectorization: If the patches are d1xd2xd3 tensors, then the vectors are d1d2d3×1

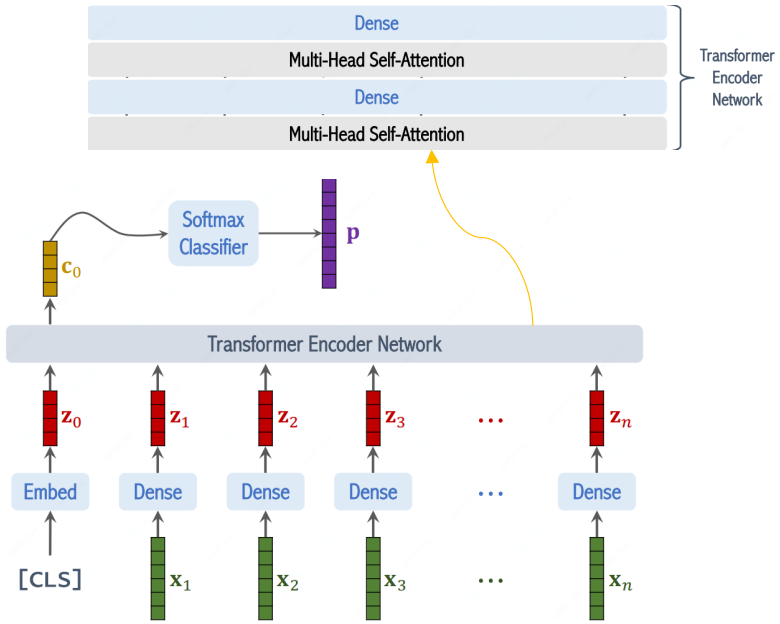


[An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021]

Development of Foundation Models

- **Large Vision Models (LVMs)**
  - ViT – Classification Task
    - Add Position Encoding
    - Connect to Transformer





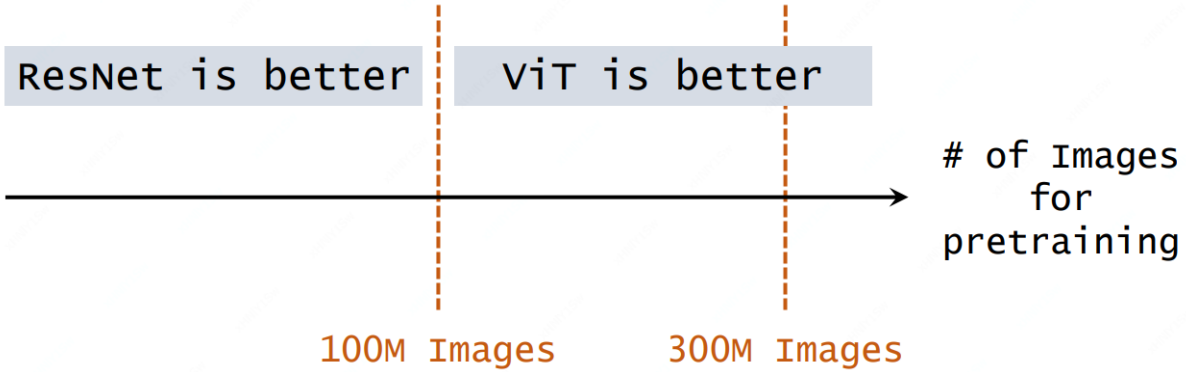
[An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021]

Development of Foundation Models

Large Vision Models (LVMs)

- ViT – Classification Task
  - Pretrain the model on Dataset A, fine-tune on Dataset B, then evaluate on Dataset B
  - Pretrained on ImageNet (small), ViT is slightly worse than ResNet.
  - Pretrained on ImageNet-21K (medium), ViT is comparable to ResNet.
  - Pretrained on JFT (large), ViT is slightly better than ResNet.

	# of Images	# of Classes
ImageNet (Small)	1.3 Million	1 Thousand
ImageNet-21K (Medium)	14 Million	21 Thousand
JFT (Big)	300 Million	18 Thousand

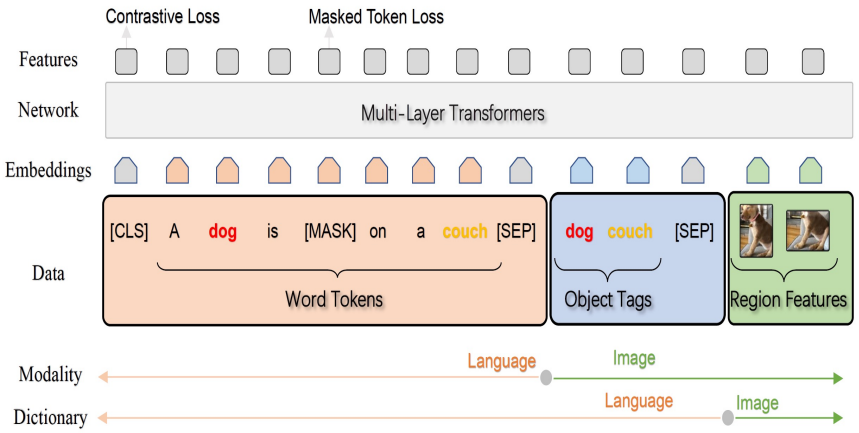


[An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021]

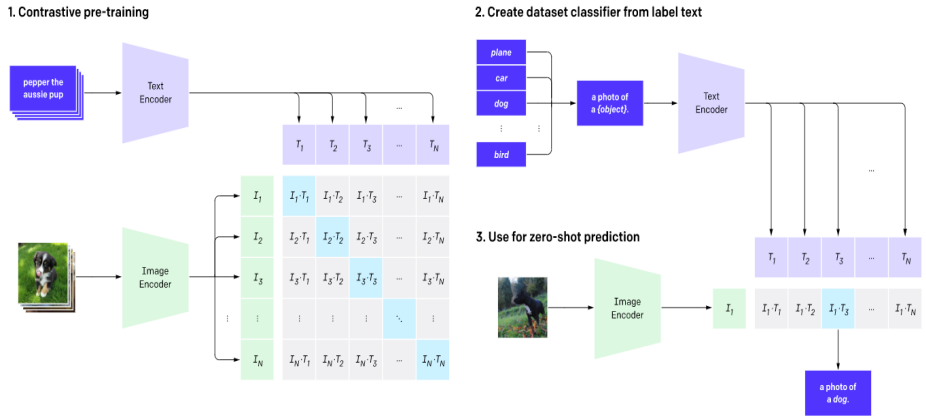
Development of Foundation Models

Vision Language Models (VLMs)

- Single stream（VisualBERT, VL-BERT, **Oscar**）- Early fusion
- Dual stream（ViLBERT, LXMERT, **CLIP**）- Late fusion



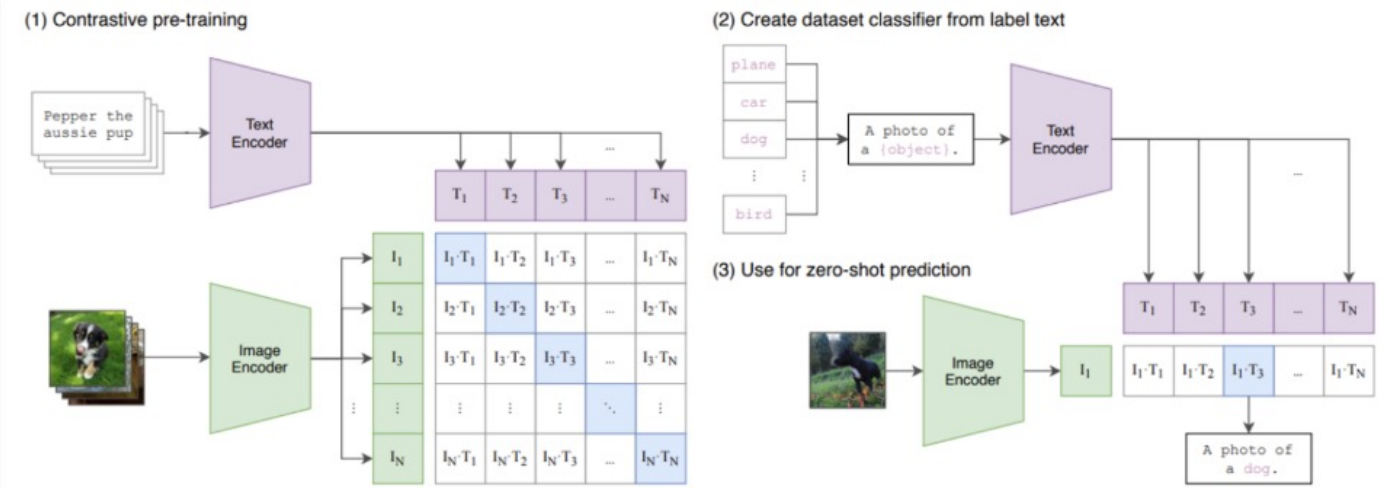
Oscar



CLIP

Development of Foundation Models

- **Vision Language Models (VLMs)**



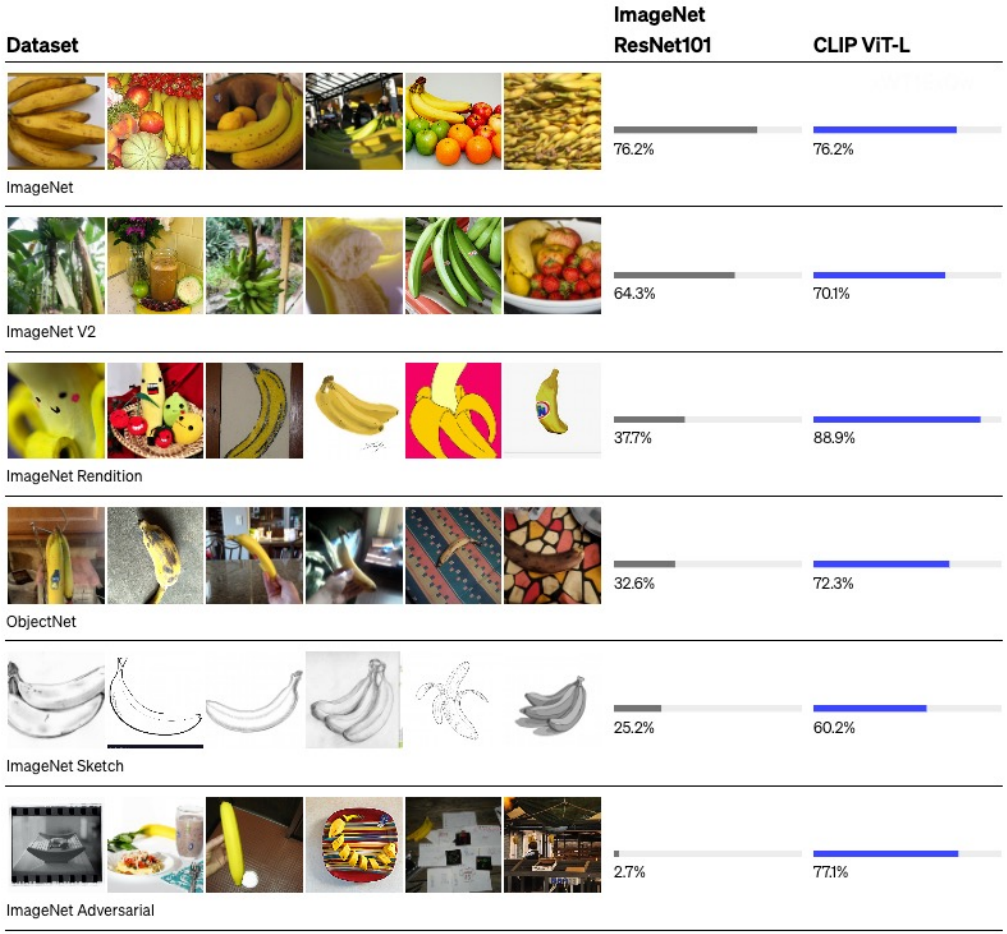
Contrastive Language-Image Pre-Training (CLIP)

- Rich pretraining data (400M image-text pairs)
- Mapping images and text to a shared embedding space
- Capability for zero-shot image classification and image-text retrieval

[Learning Transferable Visual Models From Natural Language Supervision, 2021]

Development of Foundation Models

- **Vision Language Models (VLMs)**

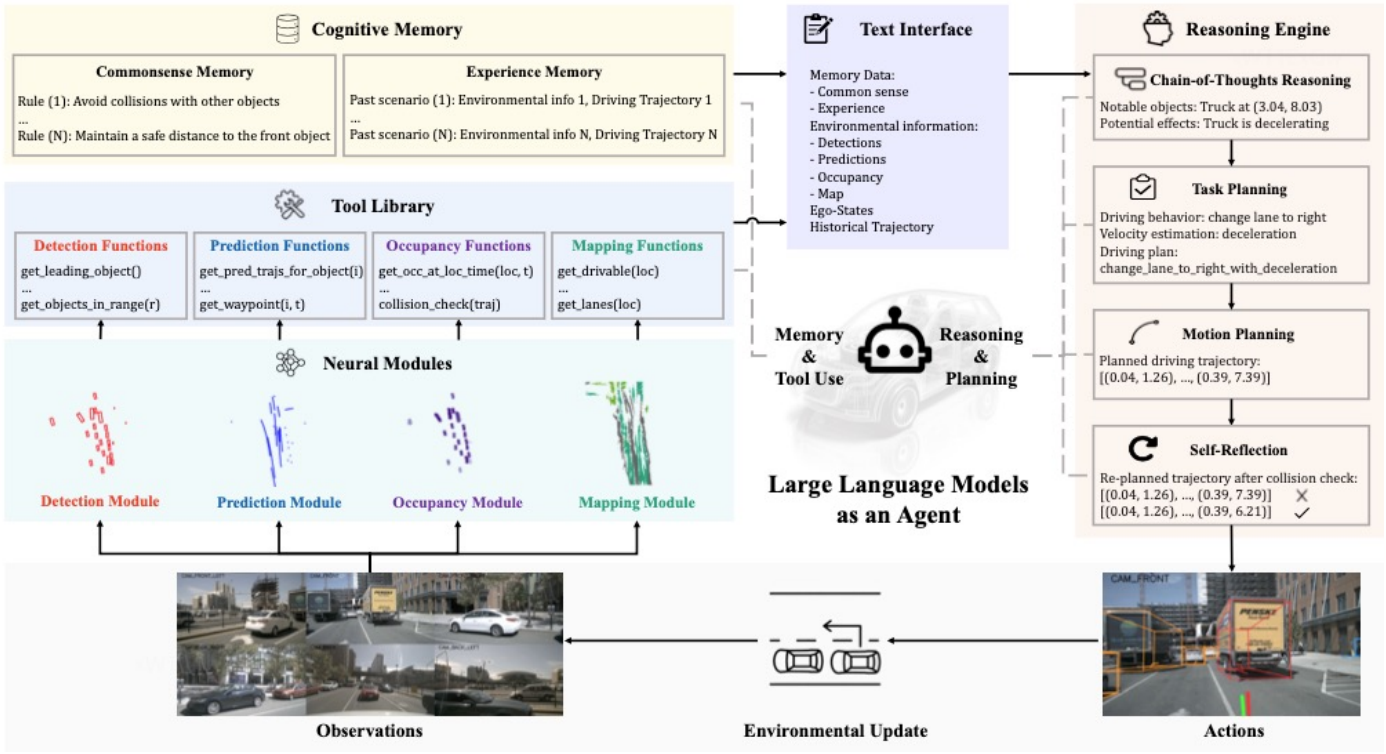


[Learning Transferable Visual Models From Natural Language Supervision, 2021]

How to Combine Spatio-Temporal Data and Foundation Models?

How to Combine?

- **Use Spatio-Temporal Data as Text or Image Formats**



Agent-Driver

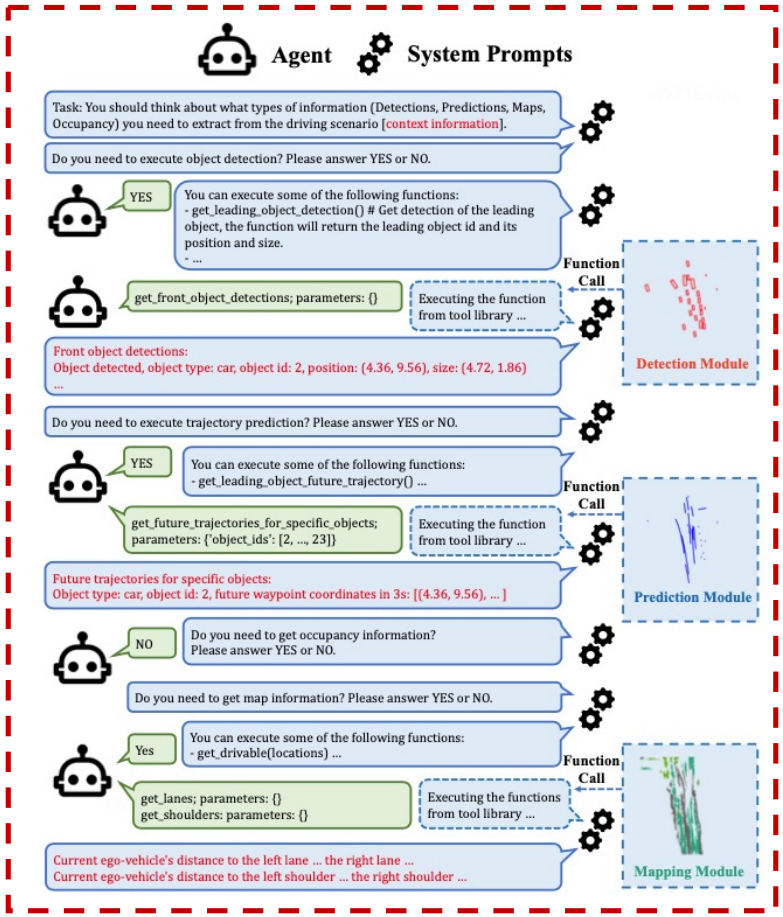
[A Language Agent for Autonomous Driving, 2023]

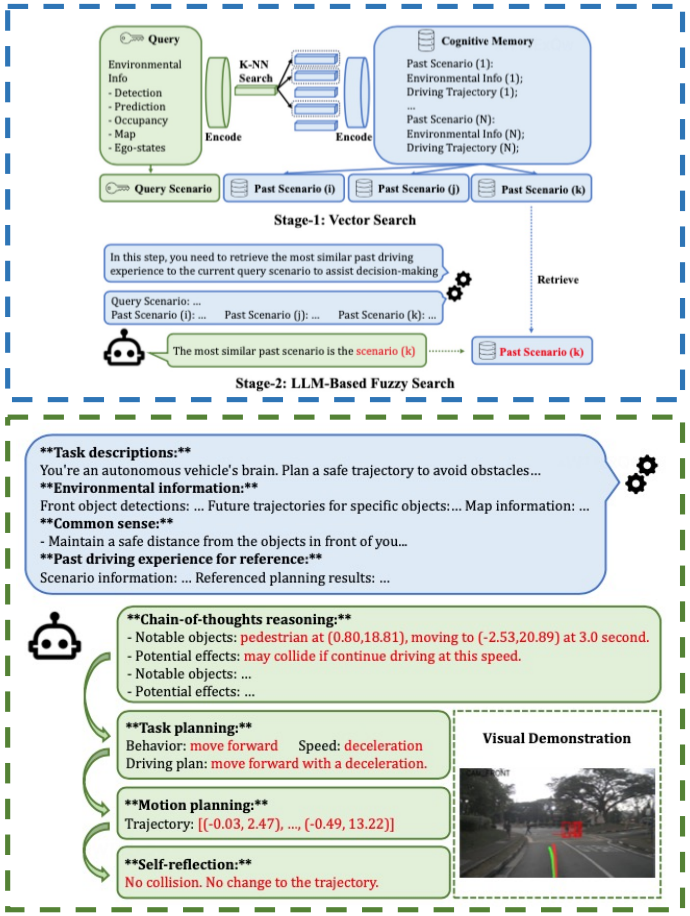
How to Combine?

- **Use Spatio-Temporal Data as Text or Image Formats**

Tool Library

Leverage text as unified interface





Cognitive Memory

A two-stage search algorithm

Reasoning Engine

Chain-of-Thought
Task Planning
Motion Planning
Self-Reflection

[A Language Agent for Autonomous Driving, 2023]

How to Combine?

- **Use Spatio-Temporal Data as Text or Image Formats**

*****Environmental information:*****
Front object detections:
Front object detected, object type: pedestrian, object id: 0, position: (-4.32, 13.85), size: (0.76, 0.91)
...
Front object detected, object type: pedestrian, object id: 9, position: (2.23, 19.39), size: (0.57, 0.69)
Future trajectories for specific objects:
Object type: pedestrian, object id: 0, future waypoint coordinates in 3s: [(-4.44, 14.50), ...]
...
Object type: pedestrian, object id: 9, future waypoint coordinates in 3s: [(1.74, 19.76), ...]
Map information (road shoulders):
Current ego-vehicle's distance to left shoulder is 0.5m and right shoulder is 4.5m

*****Common sense:*****
- Maintain a safe distance from the objects in front of you...
*****Past driving experience for reference:*****
Most similar driving experience from memory with similarity score: 0.77:
Scenario information: ...
The planned trajectory in this scenario for your reference:[(0.04,2.49), ... ,(0.30,12.26)]

*****Chain-of-thoughts reasoning:*****
- Notable objects: pedestrian at (0.80,18.81), moving to (-2.53,20.89) at 3.0 second
- Potential effects: may collide if continue driving at this speed.
...
*****Task planning:*****
Behavior: forward; Speed: deceleration; Driving plan: move forward with a deceleration
*****Motion planning:*****
Trajectory: [(-0.03, 2.47), (-0.10, 4.84), (-0.19, 7.10), (-0.29, 9.25), (-0.39, 11.29), (-0.49, 13.22)]
*****Self-reflection:*****
No collision. No change to the motion planning result.

*****Environmental information:*****
Front object detections:
Front object detected, object type: car, object id: 4, position: (-2.11, 14.95), size: (1.96, 4.76)
...
Front object detected, object type: car, object id: 6, position: (5.31, 32.79), size: (1.90, 4.48)
Future trajectories for specific objects:
Object type: pedestrian, object id: 4, future waypoint coordinates in 3s: [(-2.39, 14.80), ...]
...
Object type: pedestrian, object id: 6, future waypoint coordinates in 3s: [(5.32, 32.78), ...]
Map information (lanes):
Current ego-vehicle's distance to left lane is 1.5m and right lane is unknown

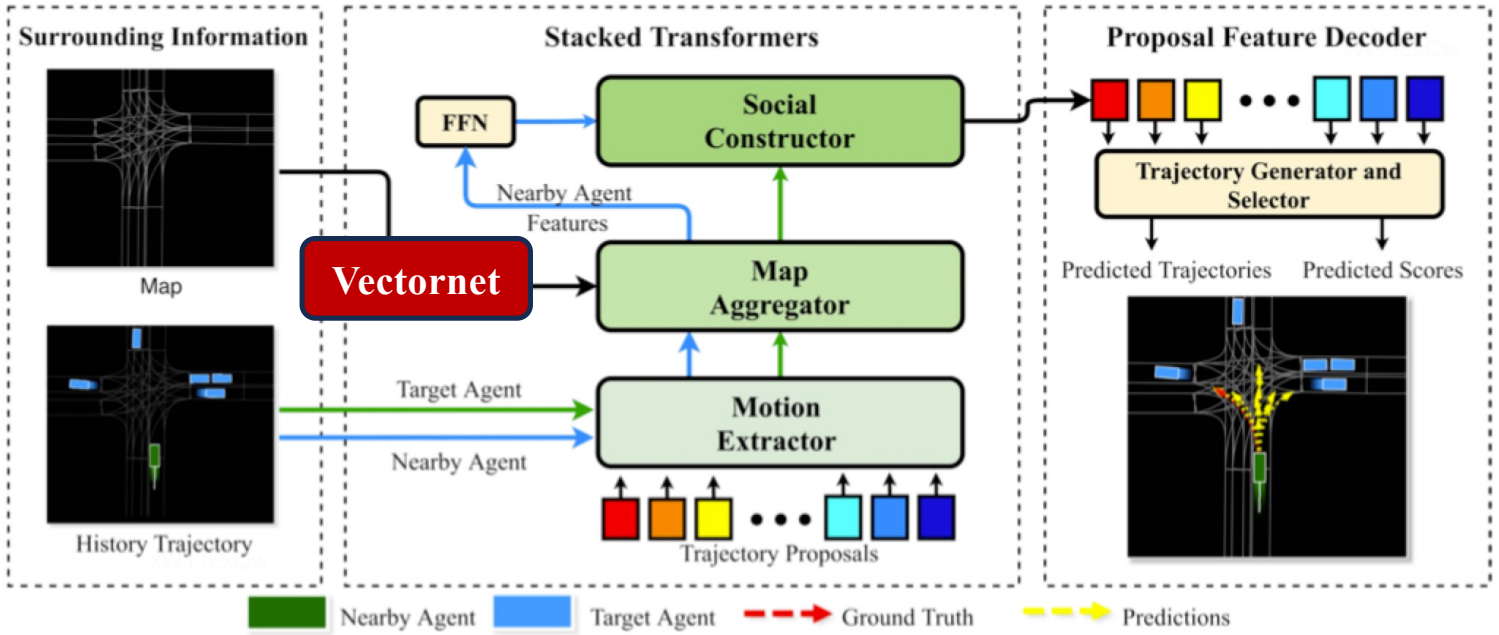
*****Common sense:*****
- Avoid collision with other objects...
*****Past driving experience for reference:*****
Most similar driving experience from memory with similarity score: 0.45:
Scenario information: ...
The planned trajectory in this scenario for your reference:[(-0.14,0.98), ... ,(-5.10,8.27)]

*****Chain-of-thoughts reasoning:*****
- Notable objects: car at (-2.11,14.95), moving to (-2.84,14.53) at 1.5 second
- Potential effects: inside the safety zone of the ego-vehicle at 1.5 second.
...
*****Task planning:*****
Behavior: turn left; Speed: deceleration; Driving plan: turn left with a deceleration
*****Motion planning:*****
Trajectory: [(-0.11,0.94), (-0.31,1.81), (-0.62,2.75), (-1.16,3.88), (-1.84,4.93), (-2.95,6.29)]
*****Self-reflection:*****
No collision. No change to the motion planning result.

[A Language Agent for Autonomous Driving, 2023]

How to Combine?

- **Use Spatio-Temporal Data as Text or Image Formats**

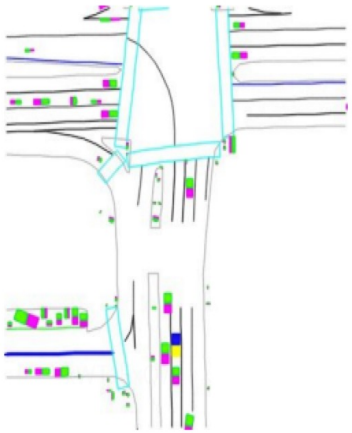


mmTransformer

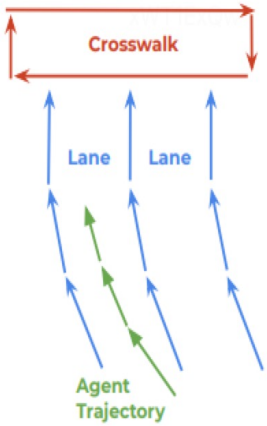
[Multimodal Motion Prediction with Stacked Transformers, 2021]

How to Combine?

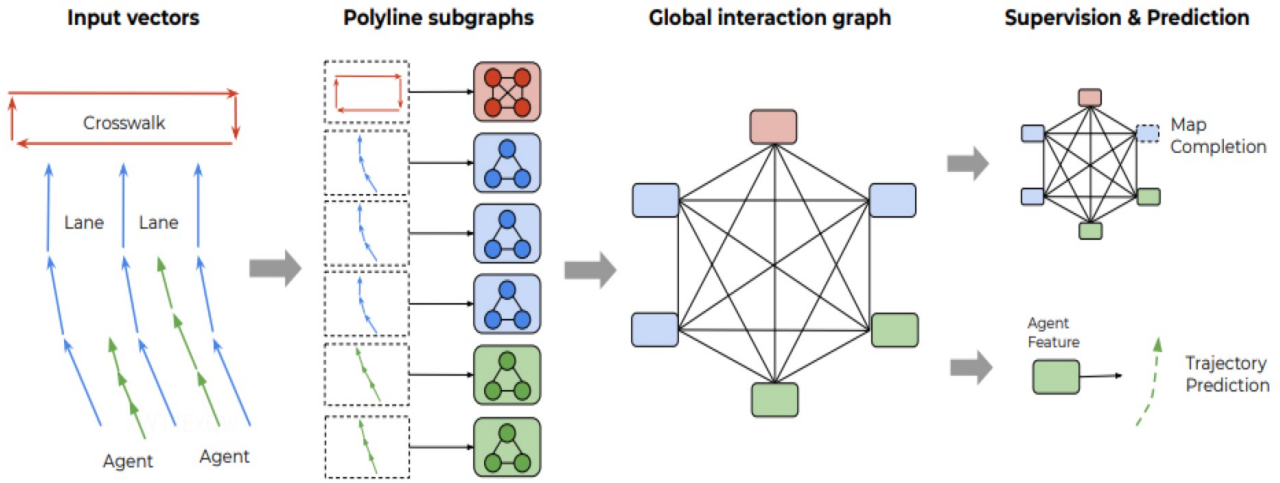
- **Use Spatio-Temporal Data as Text or Image Formats**



Rasterized Representation



Vectorized Representation



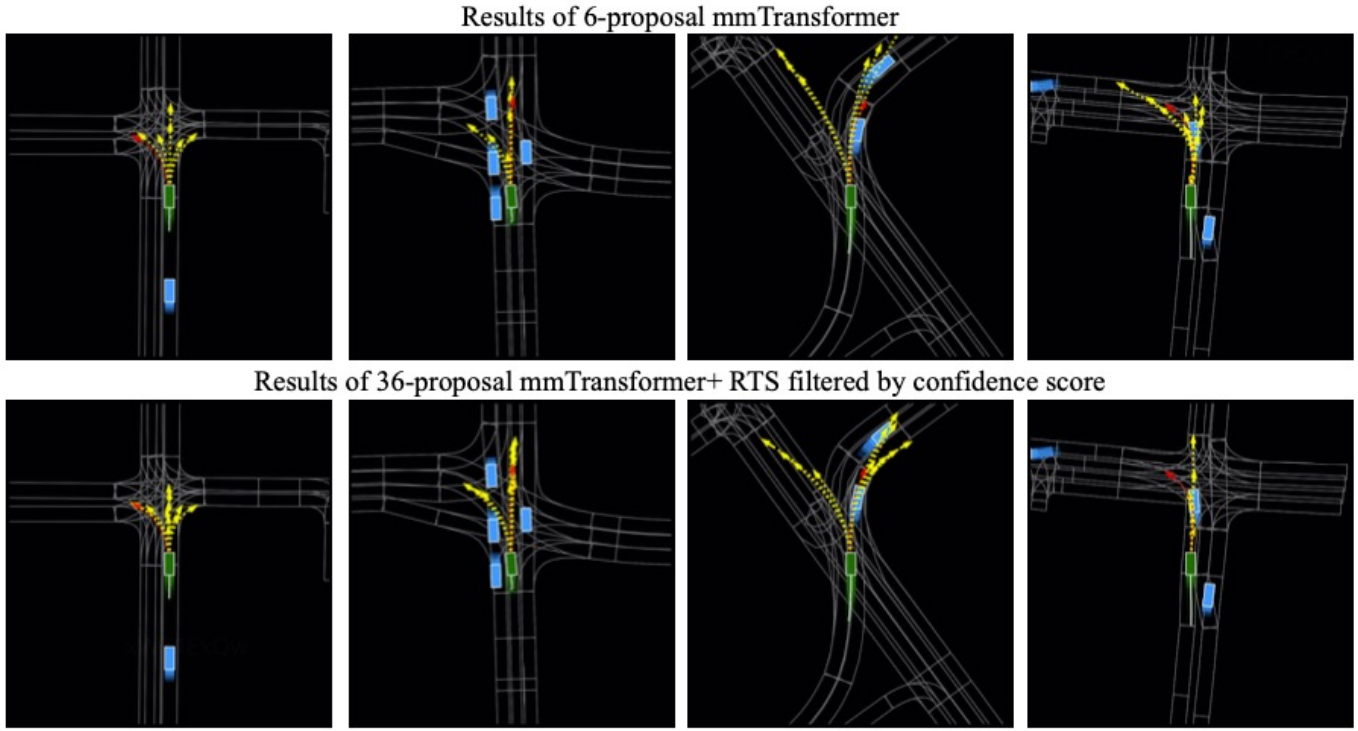
VectorNet

- ☐ Requires manual specifications
- ☐ Limited receptive fields

[VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation, 2020]

How to Combine?

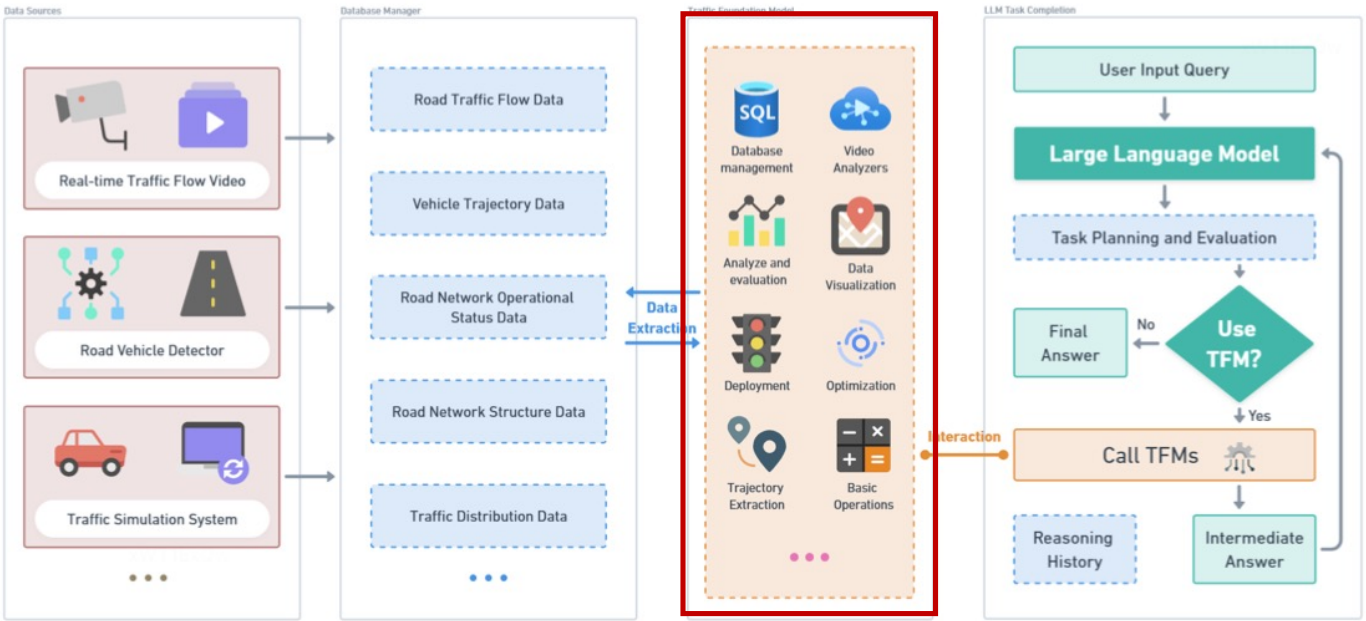
- **Use Spatio-Temporal Data as Text or Image Formats**



[Multimodal Motion Prediction with Stacked Transformers, 2021]

How to Combine?

- **Employ Specific Spatio-Temporal Models as Tools**

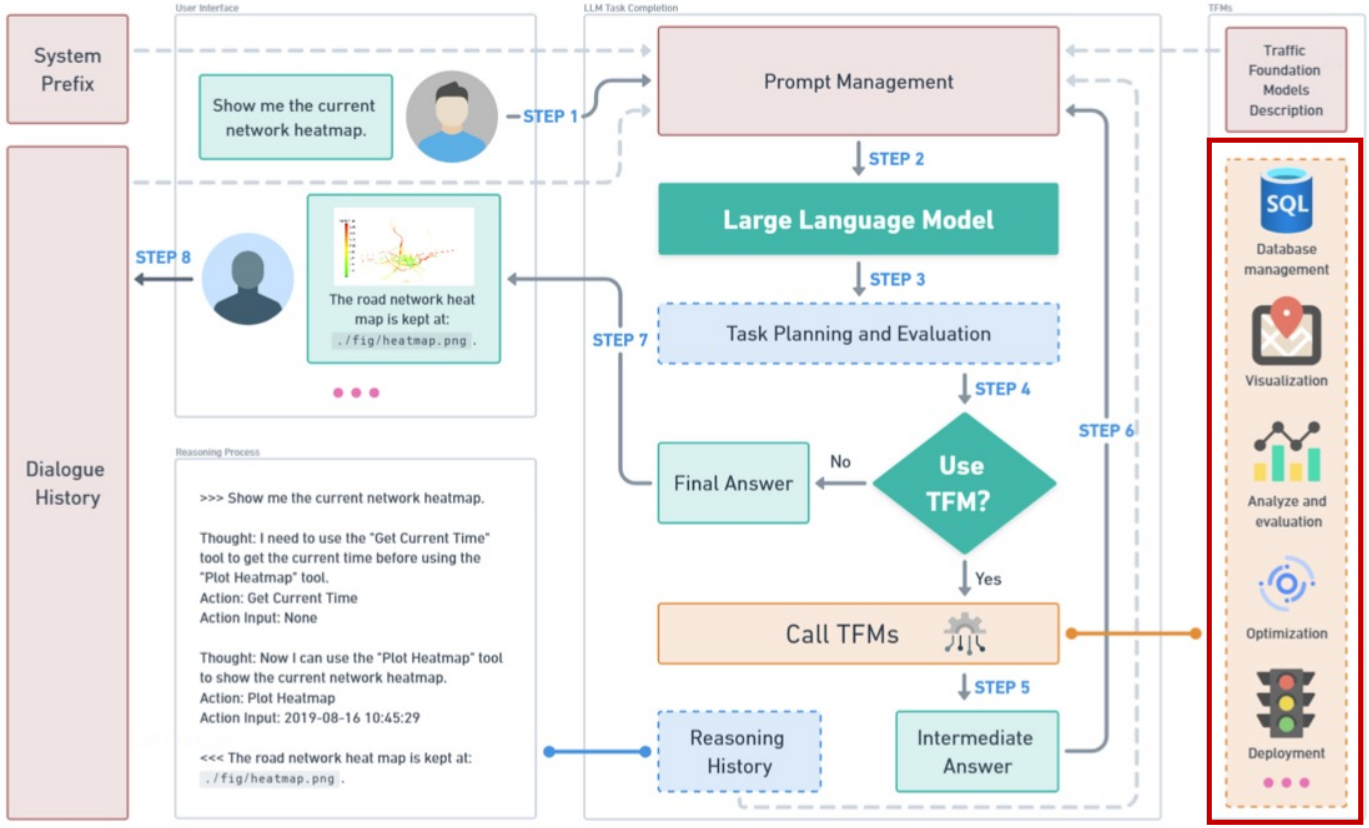


Tools

[TrafficGPT: Viewing, Processing and Interacting with Traffic Foundation Models, 2023]

How to Combine?

- **Employ Specific Spatio-Temporal Models as Tools**

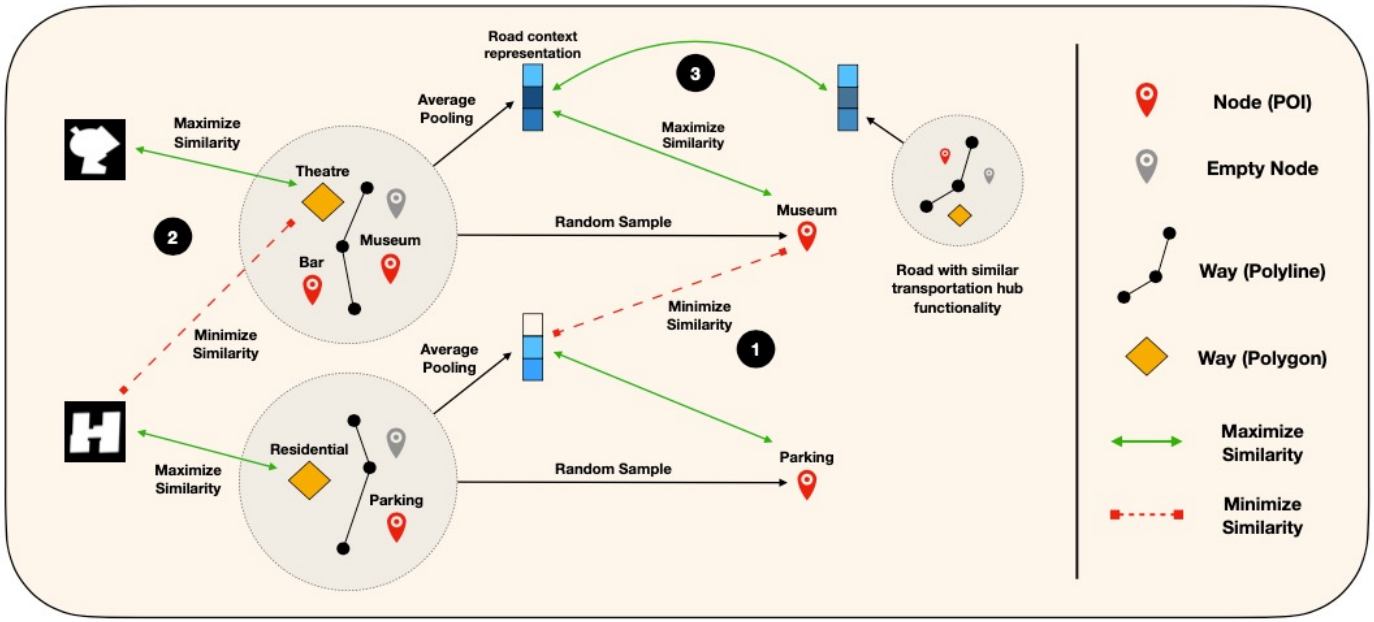


Tools

[TrafficGPT: Viewing, Processing and Interacting with Traffic Foundation Models, 2023]

How to Combine?

- **Pre-train a Foundation Model with Spatio-Temporal Data**



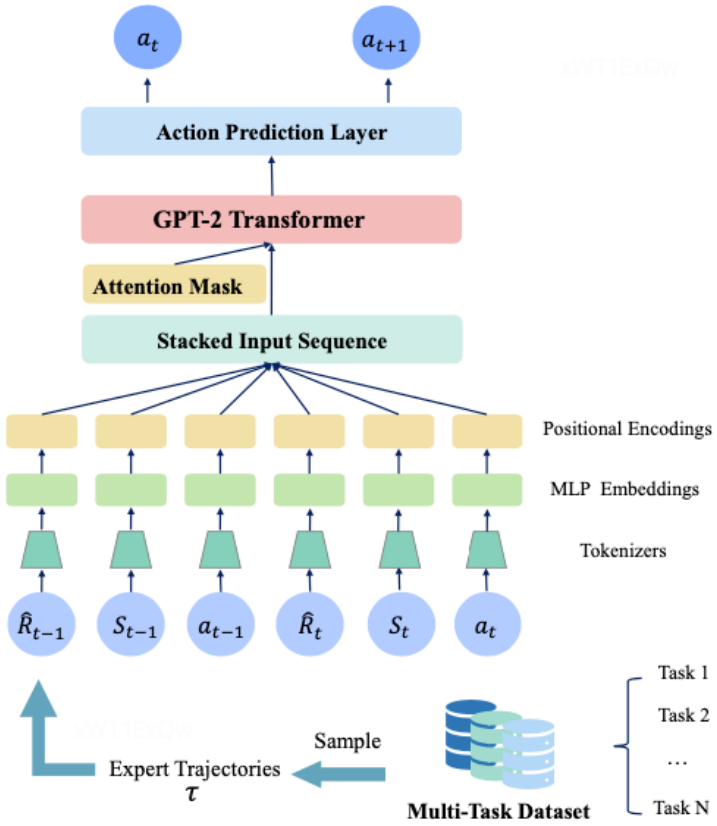
Input: OSM data

Objectives:
① Text-based objective
② Vision-Language multimodal objective
③ Road-based objective.

[CityFM: City Foundation Models to Solve Urban Challenges, 2023]

How to Combine?

- **Pre-train a Foundation Model with Spatio-Temporal Data**



- Abstract the problem of multi-task decision-making for autonomous driving as a sequence modeling and prediction task.

- Transform the ”state-action-reward” tuples from expert data into a token format similar to NLP task to match the input format of the GPT model

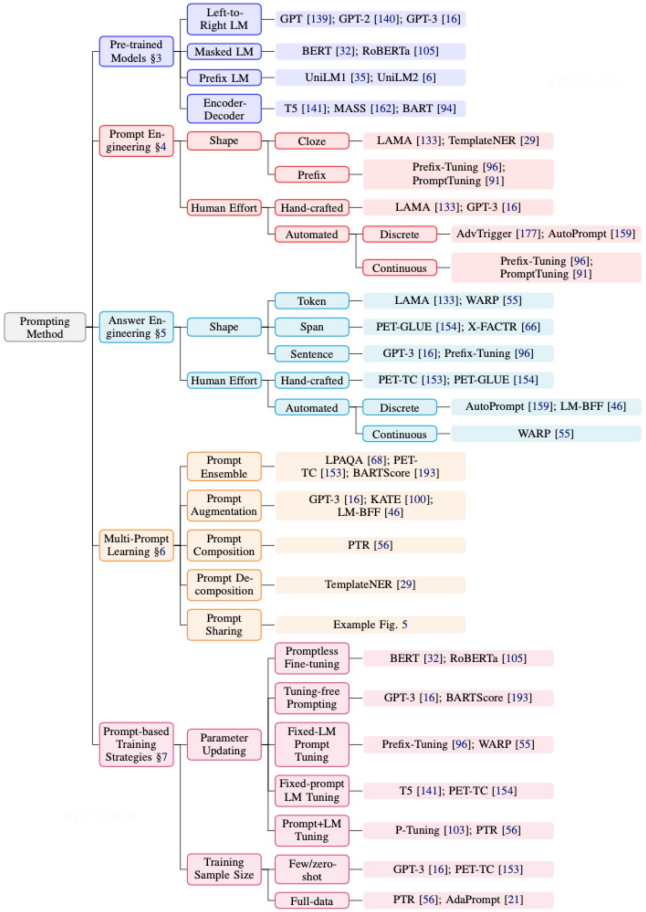
Input: $\tau' = \left(s_1, a_1, g_1, s_2, a_2, g_2, \ldots, s_T, a_T, g_T\right)$

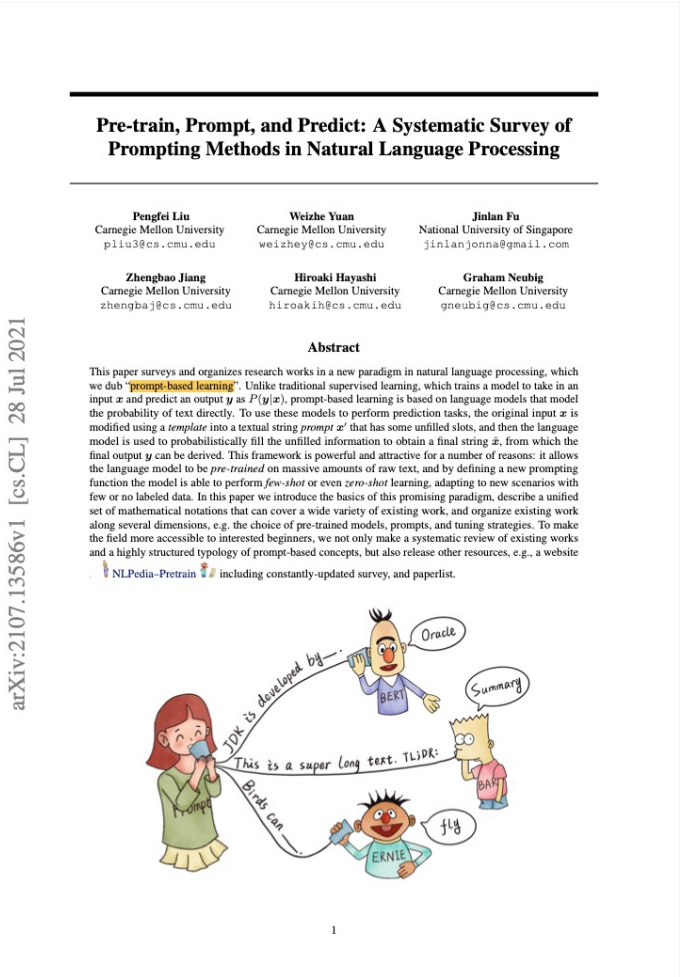
s: Sequence of state
a: Sequence of action
g: Sequence of function of trajectory

[MTD-GPT: A Multi-Task Decision-Making GPT Model for Autonomous Driving at Unsignalized Intersections, 2023]

Others

- **Prompt Learning**





arXiv:2107.13586v1 [cs.CL] 28 Jul 2021

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

Pengfei Liu
Carnegie Mellon University
pliu3@cs.cmu.edu

Weizhe Yuan
Carnegie Mellon University
weizhey@cs.cmu.edu

Jinlan Fu
National University of Singapore
jinlanjonna@gmail.com

Zhengbao Jiang
Carnegie Mellon University
zhengbaj@cs.cmu.edu

Hiroaki Hayashi
Carnegie Mellon University
hiroakih@cs.cmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

Abstract

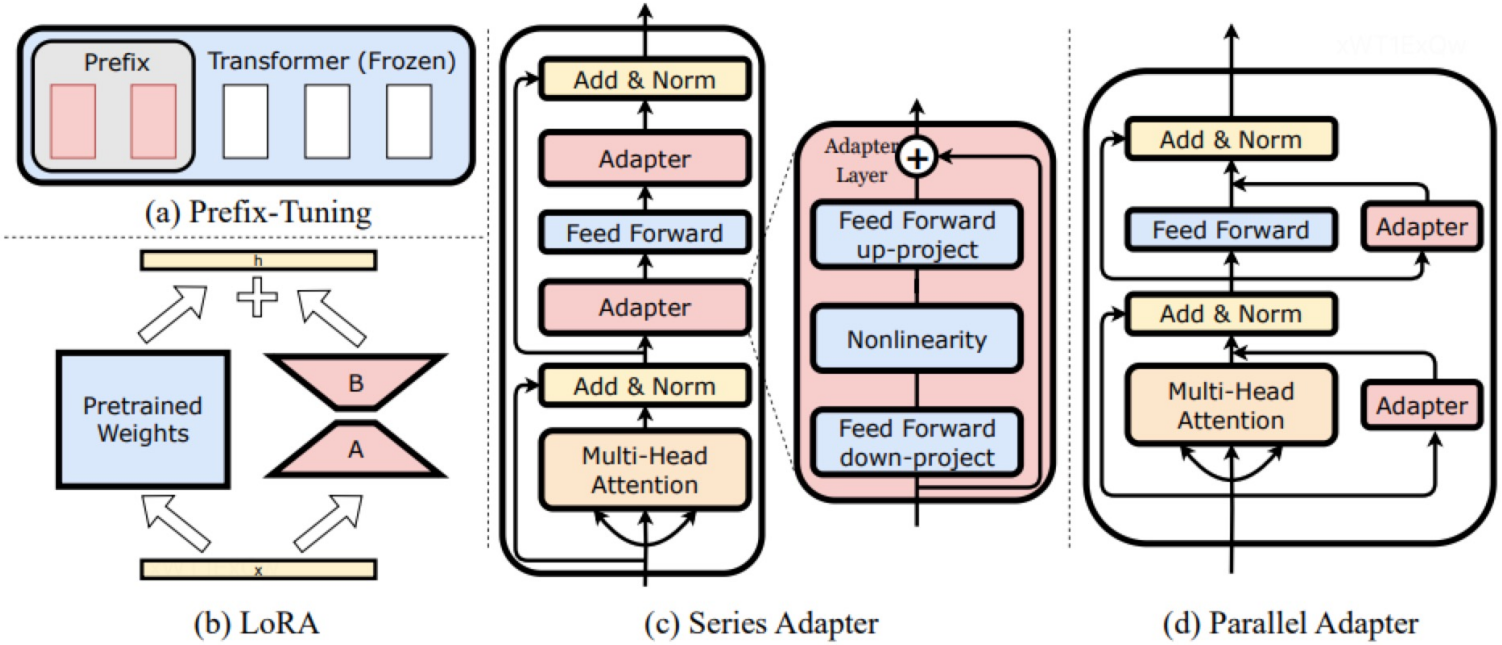
This paper surveys and organizes research works in a new paradigm in natural language processing, which we dub "prompt-based learning". Unlike traditional supervised learning, which trains a model to take in an input x and predict an output y as $P(y|x)$, prompt-based learning is based on language models that model the probability of text directly. To use these models to perform prediction tasks, the original input x is modified using a *template* into a textual string *prompt* x' that has some unfilled slots, and then the language model is used to probabilistically fill the unfilled information to obtain a final string $\hat{x}$, from which the final output y can be derived. This framework is powerful and attractive for a number of reasons: it allows the language model to be *pre-trained* on massive amounts of raw text, and by defining a new prompting function the model is able to perform *few-shot* or even *zero-shot* learning, adapting to new scenarios with few or no labeled data. In this paper we introduce the basics of this promising paradigm, describe a unified set of mathematical notations that can cover a wide variety of existing work, and organize existing work along several dimensions, e.g. the choice of pre-trained models, prompts, and tuning strategies. To make the field more accessible to interested beginners, we not only make a systematic review of existing works and a highly structured typology of prompt-based concepts, but also release other resources, e.g., a website NLPedia–Pretrain including constantly-updated survey, and paperlist.

1

[Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, 2021]

Others

- **Parameter-Efficient Fine-Tuning (PEFT)**

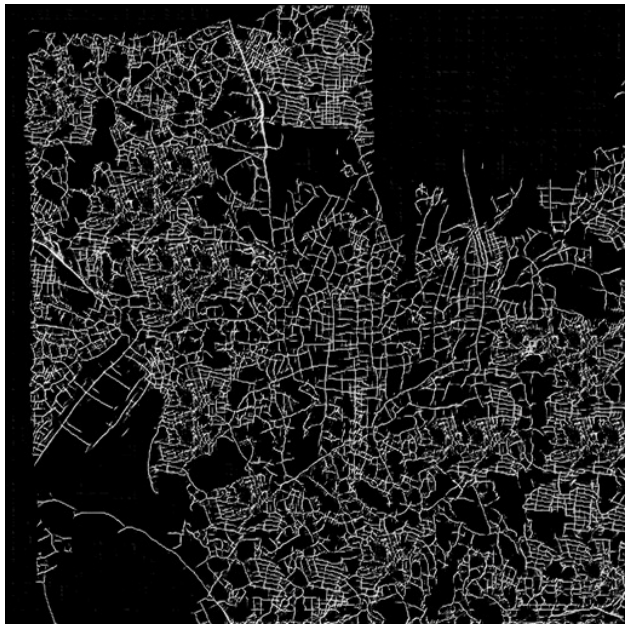


[LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models, 2023]

Future Directions

- **Traffic Scene Understanding**
- **Traffic Description Generation**
- **Traffic Simulation**
- **Trajectory/Road Network Generation**

…



Thanks for Listening

Jingtian Ma