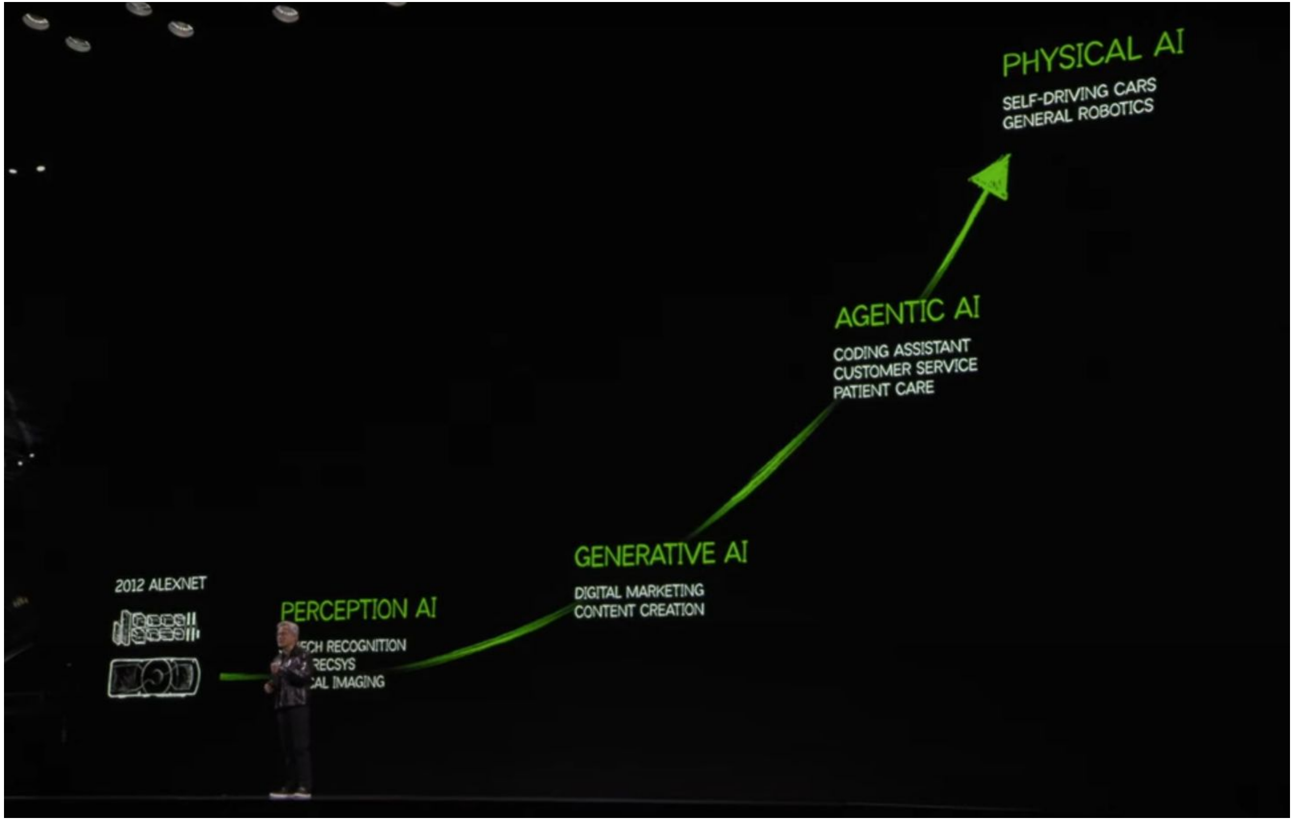
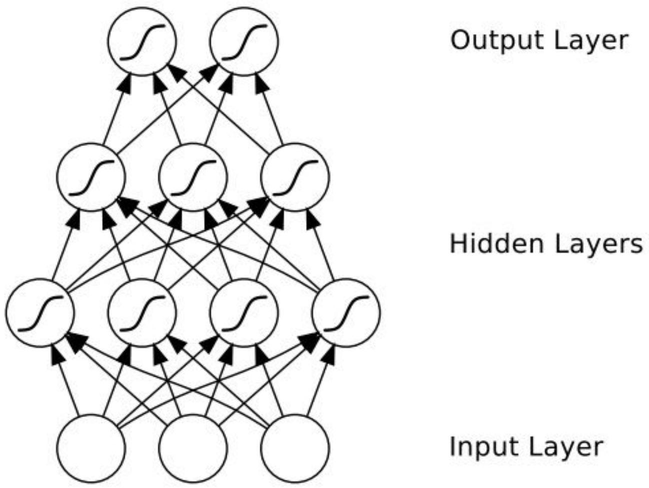

PHYSICAL AI
SELF-DRIVING CARS
GENERAL ROBOTICS
AGENTIC AI
CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE
GENERATIVE AI
DIGITAL MARKETING
CONTENT CREATION
2012 ALEXNET
PERCEPTION AI

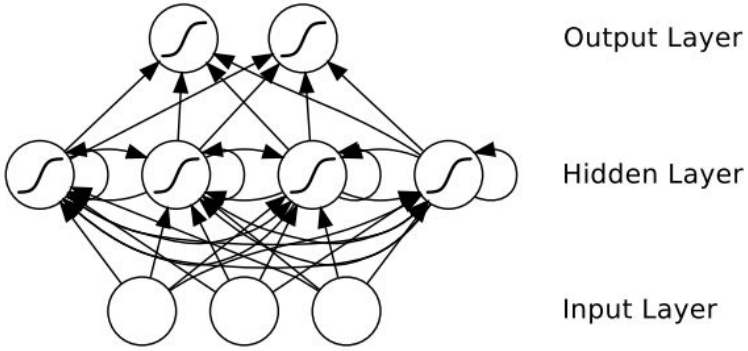
[Link](#)

RNN & LSTM (Memory Networks)

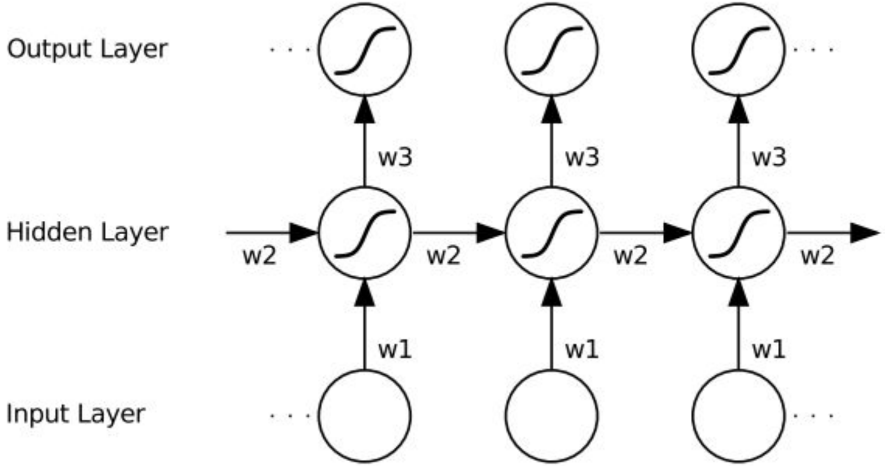
Robotics Summer School
Sanket Kalwar
(31/05/2025)
Day 11

Neural Networks & RNN’s





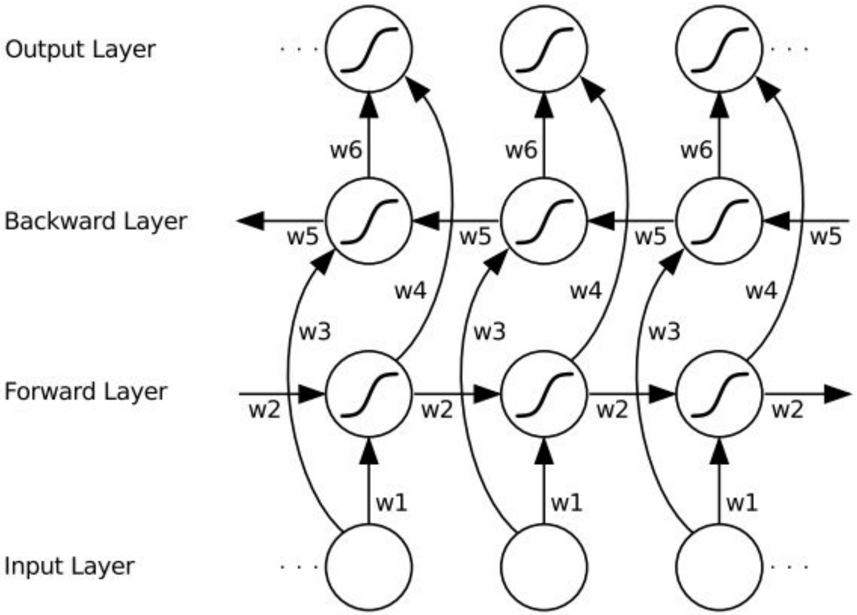
RNN:



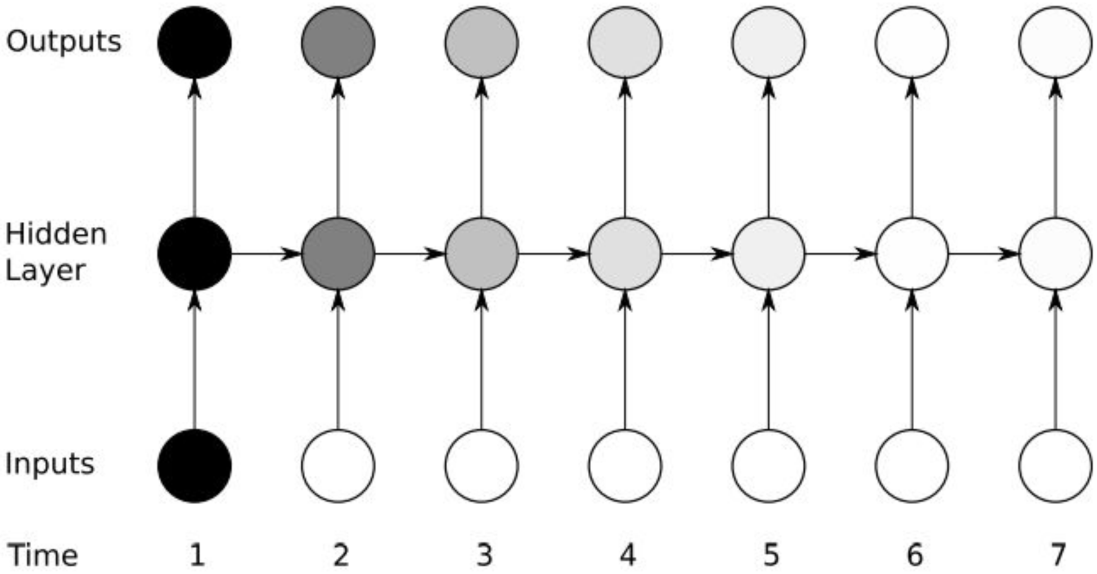
$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

[Code]

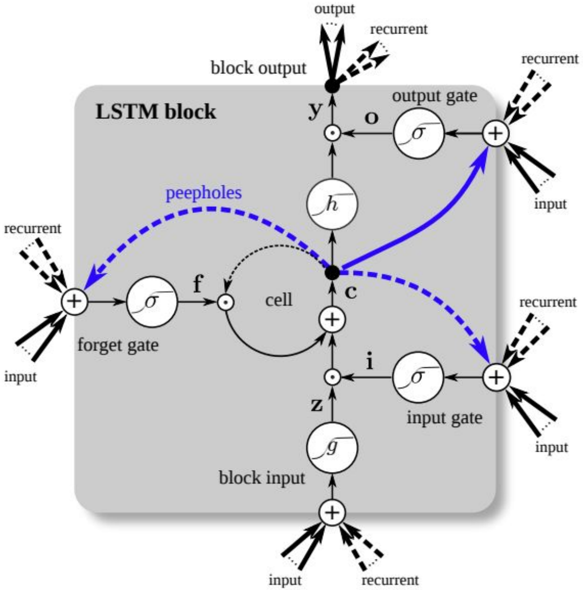
Bidirectional RNN:

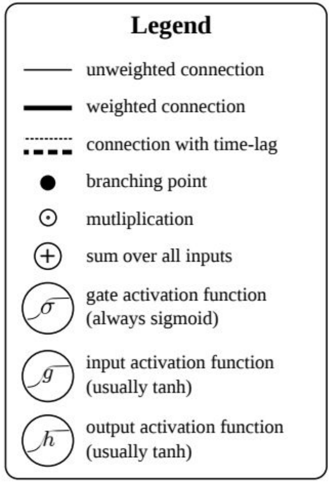


Vanishing gradient Issue with RNN:



LSTM Architecture:

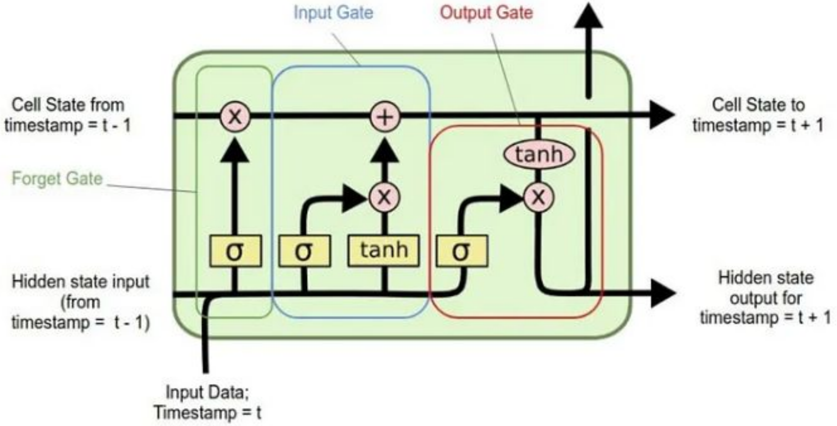


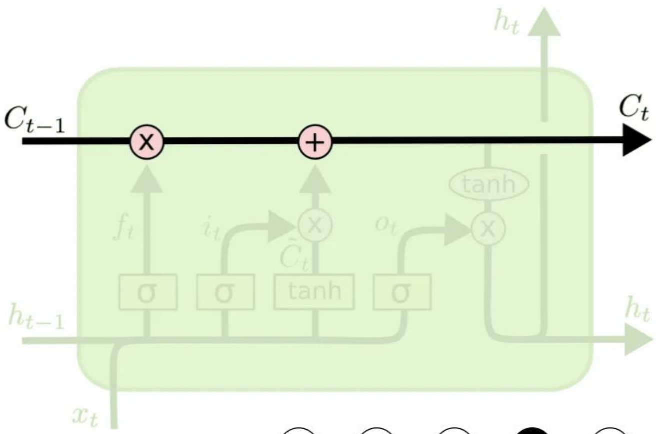


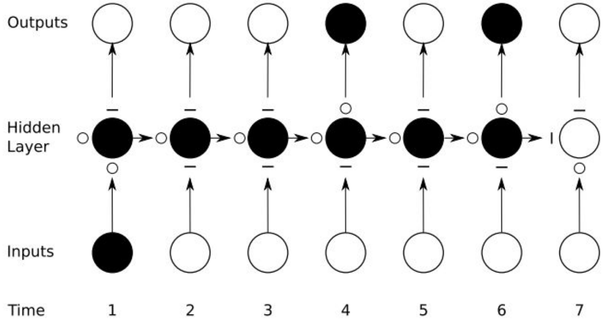
$$
\begin{aligned}
\bar{\mathbf{z}}^t &= \mathbf{W}_z\mathbf{x}^t + \mathbf{R}_z\mathbf{y}^{t-1} + \mathbf{b}_z && \\
\mathbf{z}^t &= g(\bar{\mathbf{z}}^t) && \textit{block input} \\
\bar{\mathbf{i}}^t &= \mathbf{W}_i\mathbf{x}^t + \mathbf{R}_i\mathbf{y}^{t-1} + \mathbf{p}_i \odot \mathbf{c}^{t-1} + \mathbf{b}_i && \\
\mathbf{i}^t &= \sigma(\bar{\mathbf{i}}^t) && \textit{input gate} \\
\bar{\mathbf{f}}^t &= \mathbf{W}_f\mathbf{x}^t + \mathbf{R}_f\mathbf{y}^{t-1} + \mathbf{p}_f \odot \mathbf{c}^{t-1} + \mathbf{b}_f && \\
\mathbf{f}^t &= \sigma(\bar{\mathbf{f}}^t) && \textit{forget gate} \\
\mathbf{c}^t &= \mathbf{z}^t \odot \mathbf{i}^t + \mathbf{c}^{t-1} \odot \mathbf{f}^t && \textit{cell} \\
\bar{\mathbf{o}}^t &= \mathbf{W}_o\mathbf{x}^t + \mathbf{R}_o\mathbf{y}^{t-1} + \mathbf{p}_o \odot \mathbf{c}^t + \mathbf{b}_o && \\
\mathbf{o}^t &= \sigma(\bar{\mathbf{o}}^t) && \textit{output gate} \\
\mathbf{y}^t &= h(\mathbf{c}^t) \odot \mathbf{o}^t && \textit{block output}
\end{aligned}
$$

[Paper] [Code]

Skip Connections (Addressing Vanishing Gradient Issue):







Blog

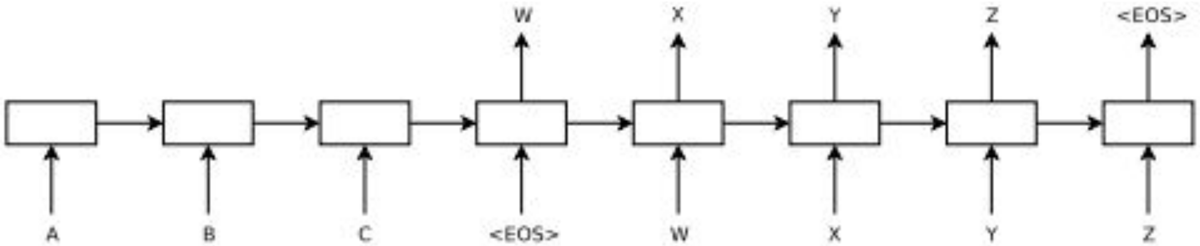
Seq2Seq Model:

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

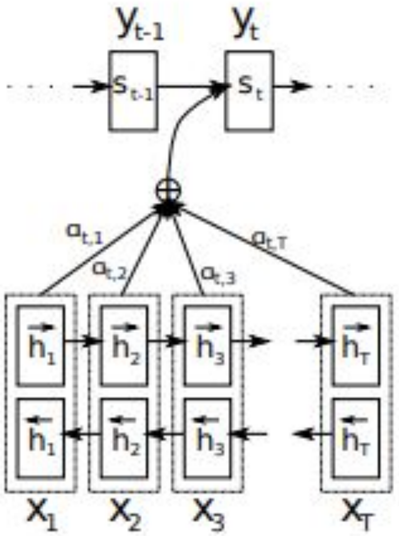
Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com



[Paper](#) [Code](#)

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE



NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

ABSTRACT

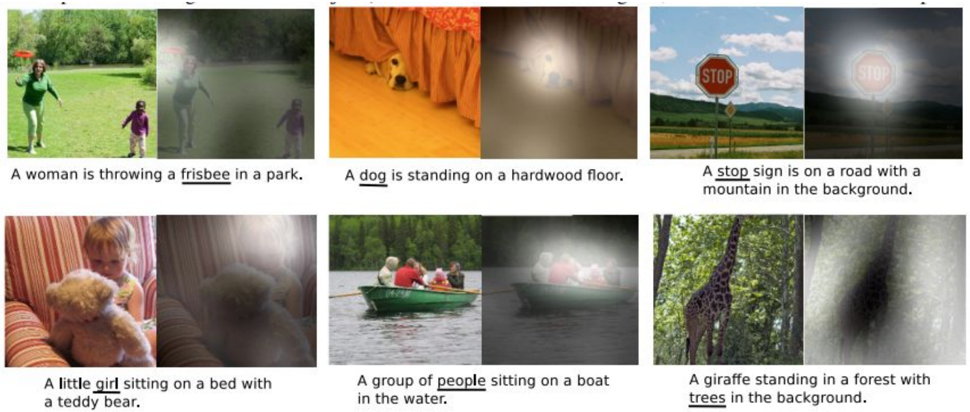
Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

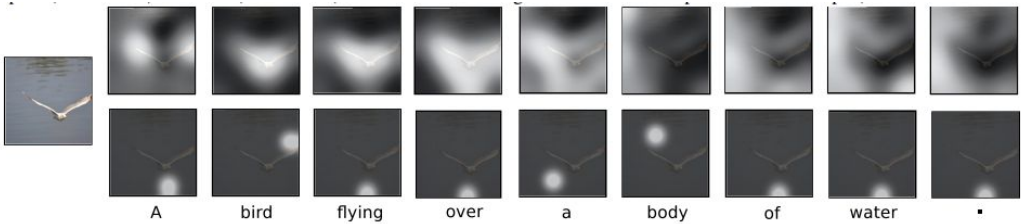
[Paper](#) [Code](#)

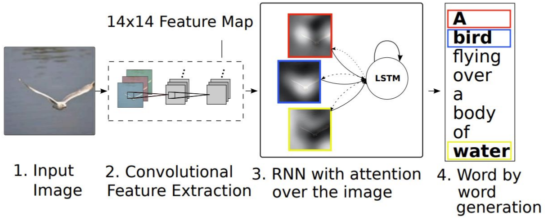
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu
Jimmy Lei Ba
Ryan Kiros
Kyunghyun Cho
Aaron Courville
Ruslan Salakhutdinov
Richard S. Zemel
Yoshua Bengio







Paper

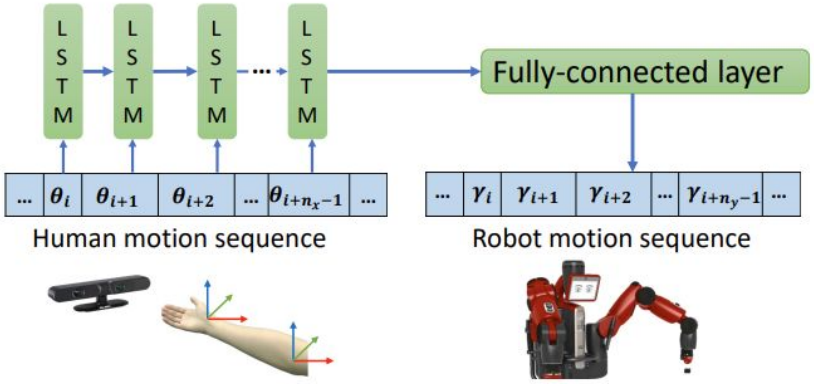
Code

Application To Robotics:

Collaborative Human-Robot Motion Generation using LSTM-RNN

Xuan Zhao, Sakmongkon Chumkamon, Shuanda Duan, Juan Rojas, and Jia Pan[†]

Abstract—We propose a deep learning based method for fast and responsive human-robot handovers that generate robot motion according to human motion observations. Our method learns an offline human-robot interaction model through a Recurrent Neural Network with Long Short-Term Memory units (LSTM-RNN). The robot uses the learned network to respond appropriately to novel online human motions. Our method is tested both on pre-recorded data and real-world human-robot handover experiments. Our method achieves robot motion accuracies that outperform the baseline. In addition, our method demonstrates a strong ability to adapt to changes in velocity of human motions.



Paper

Prior Work From RRC Using LSTM:

ATPPNet: Attention based Temporal Point cloud Prediction Network

Kaustab Pal[*1], Aditya Sharma[*1], Avinash Sharma[2], K. Madhava Krishna[1]

Abstract—Point cloud prediction is an important yet challenging task in the field of autonomous driving. The goal is to predict future point cloud sequences that maintain object structures while accurately representing their temporal motion. These predicted point clouds help in other subsequent tasks like object trajectory estimation for collision avoidance or estimating locations with the least odometry drift. In this work, we present ATPPNet, a novel architecture that predicts future point cloud sequences given a sequence of previous time step point clouds obtained with LiDAR sensor. ATPPNet leverages Conv-LSTM along with channel-wise and spatial attention dually complemented by a 3*D*-CNN branch for extracting an enhanced spatio-temporal context to recover high quality fidel predictions of future point clouds. We conduct extensive experiments on publicly available datasets and report impressive performance outperforming the existing methods. We also conduct a thorough ablative study of the proposed architecture and provide an application study that highlights the potential of our model for tasks like odometry estimation.

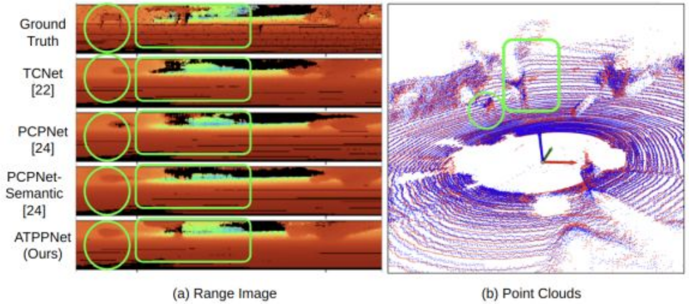
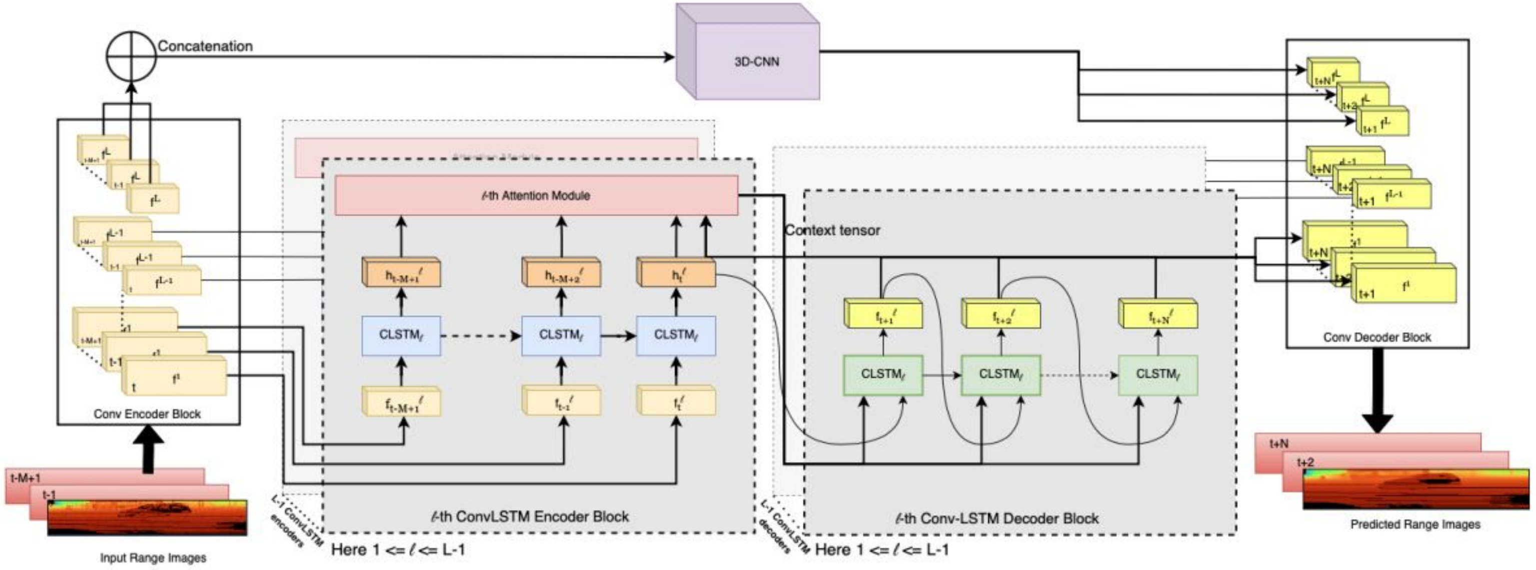


Fig. 1: (a) Predicted range images by our ATPPNet and existing methods in comparison to ground truth and, (b) the 3*D* rendering of the predicted point cloud by ATPPNet (blue) and ground-truth (red). Green circle/rectangle highlights regions where ATPPNet's predictions are superior.

(e.g., CNNs) and sequence prediction (e.g., LSTMs) cannot be directly employed as they cannot process spatially unordered data. Another key challenge is that the LiDAR point clouds are extremely sparse making it difficult to capture

[Paper](#)

ATPPNet: Attention based Temporal Point cloud Prediction Network



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

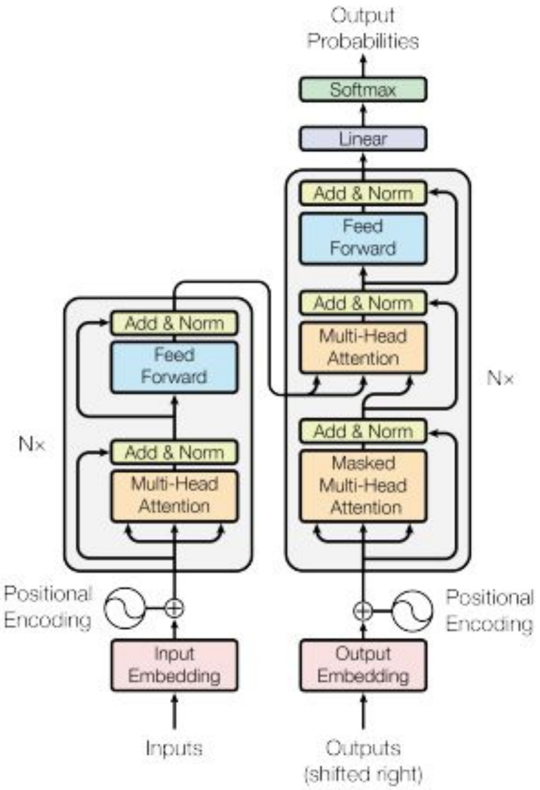
Aidan N. Gomez[*, †]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*, ‡]
illia.polosukhin@gmail.com

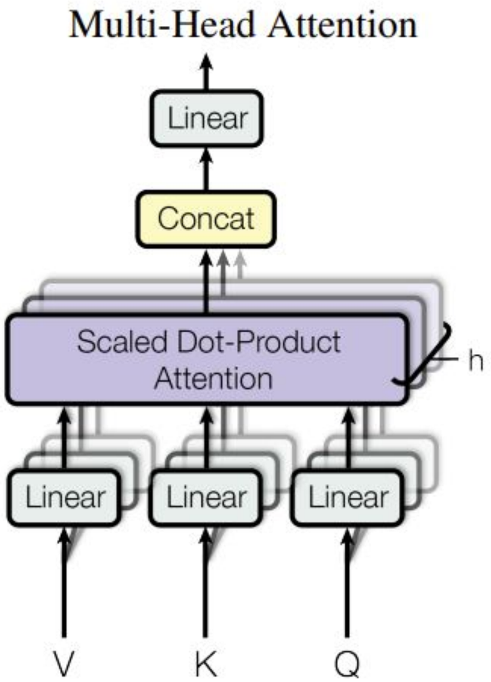
Transformer (Memory Networks)

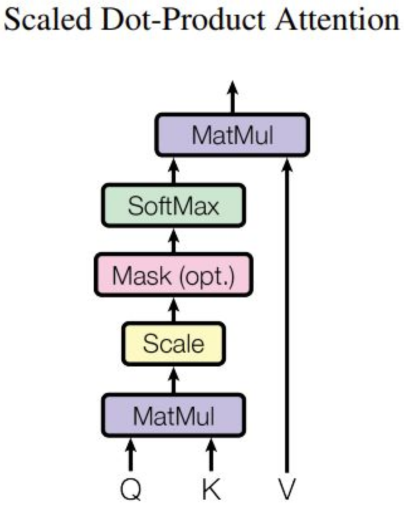
Attention is ALL You Need



[Paper](#) [Code](#)

MHA Layer:


Multi-Head Attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q


Scaled Dot-Product Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V

Need Some Attention!

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where}\ \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Let’s Talk about Compute:

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

1st Attempt To Make transformer work on Images:

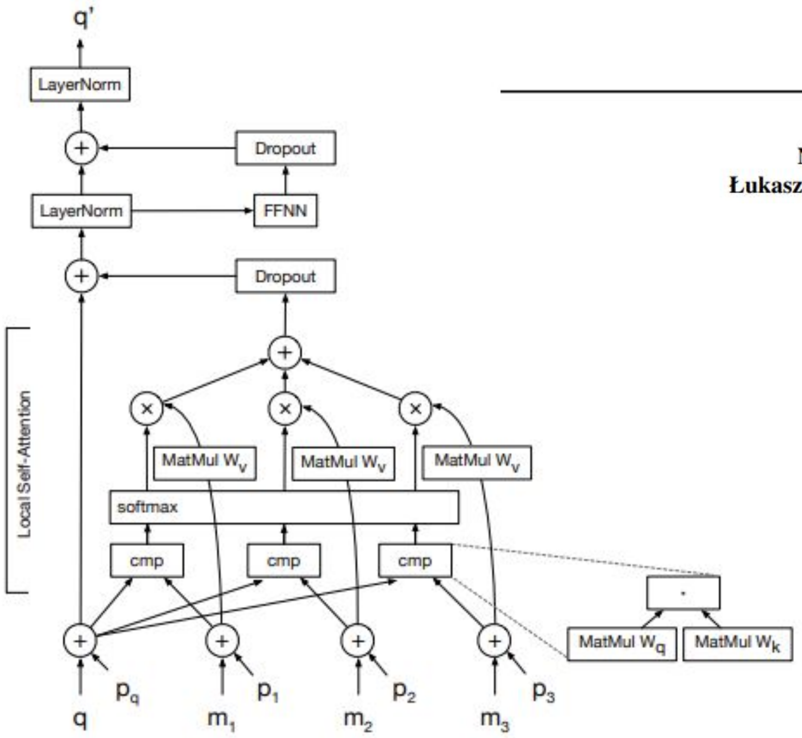



Image Transformer

Niki Parmar * [1] Ashish Vaswani * [1] Jakob Uszkoreit [1]
Łukasz Kaiser [1] Noam Shazeer [1] Alexander Ku [2 3] Dustin Tran [4]

[Paper]

2nd Attempt To Make transformer work on Images:

Non-local Neural Networks

Xiaolong Wang[1,2*] Ross Girshick[2] Abhinav Gupta[1] Kaiming He[2]
[1]Carnegie Mellon University [2]Facebook AI Research

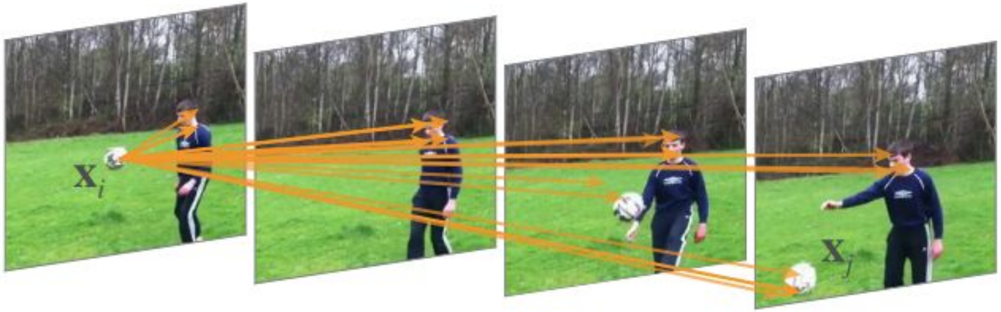
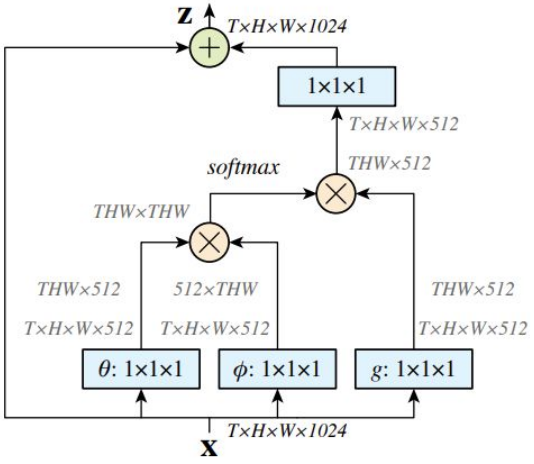
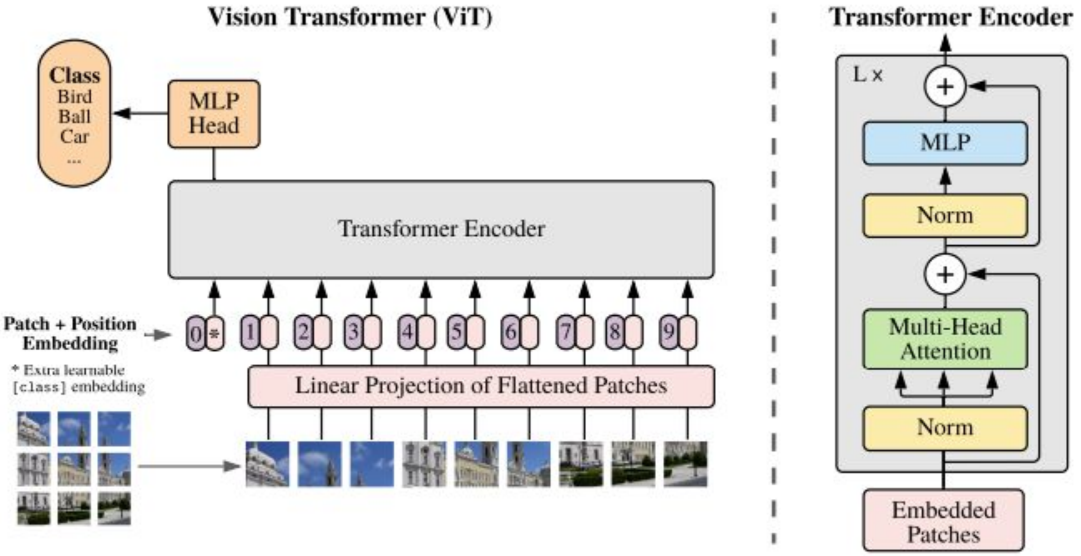


Figure 1. A spacetime ***non-local*** operation in our network trained for video classification in Kinetics. A position $\mathbf{x}_i$'s response is computed by the weighted average of the features of *all* positions $\mathbf{x}_j$ (only the highest weighted ones are shown here). In this example computed by our model, note how it relates the ball in the first frame to the ball in the last two frames. More examples are in Figure 3.



Paper Code

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE



Paper Code

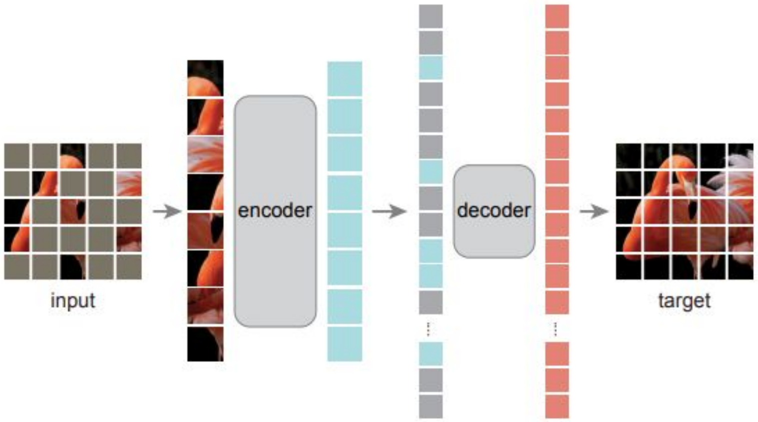
Masked Autoencoder (MAE):

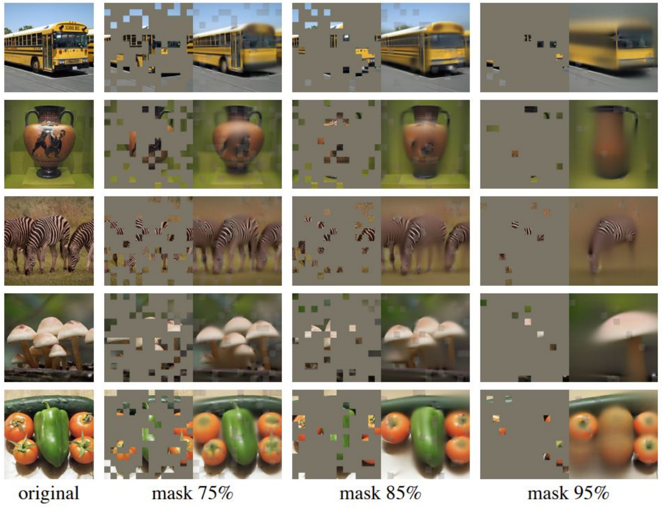
Masked Autoencoders Are Scalable Vision Learners

Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

*equal technical contribution †project lead

Facebook AI Research (FAIR)



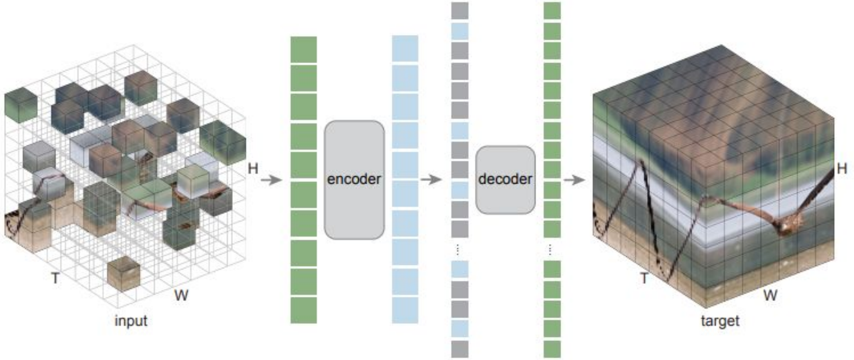


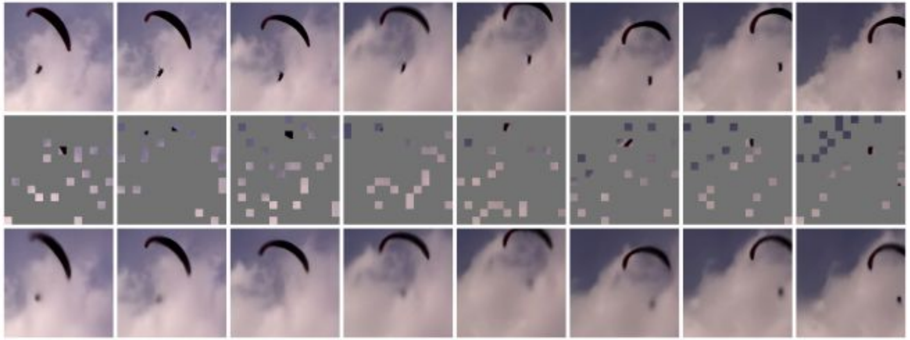
[Paper] [Code]

Masked Autoencoders As Spatiotemporal Learners:

Masked Autoencoders As Spatiotemporal Learners

Christoph Feichtenhofer*   Haoqi Fan*   Yanghao Li   Kaiming He
Meta AI, FAIR

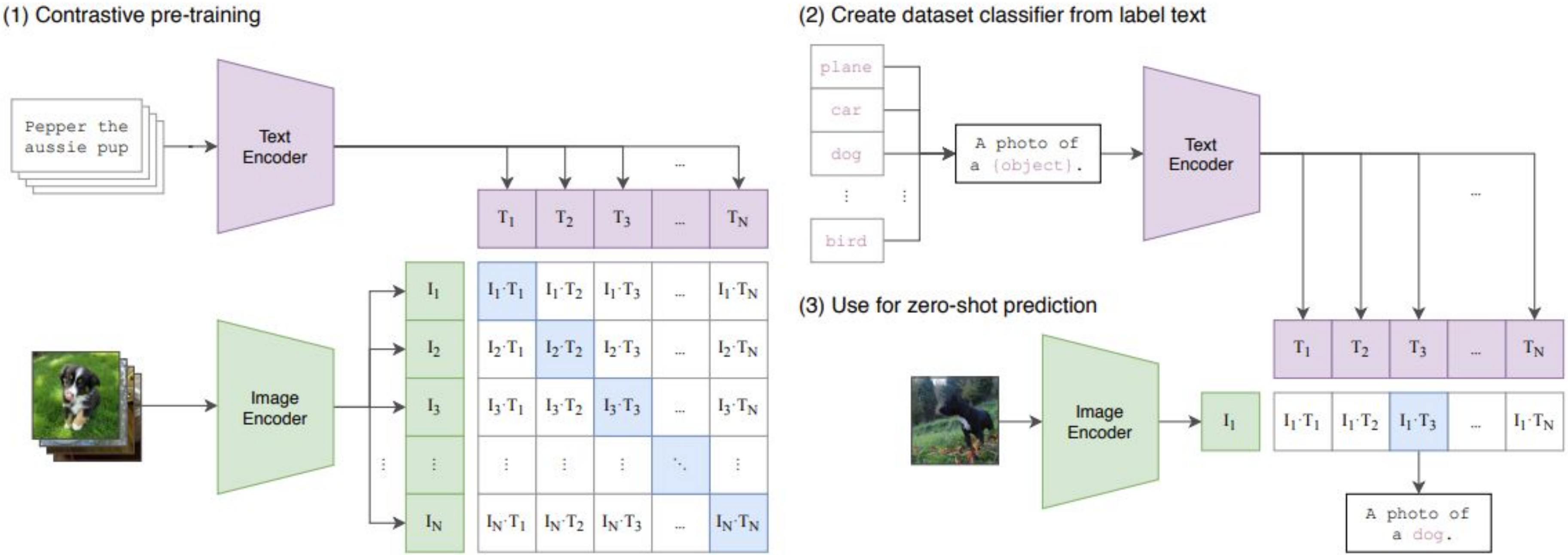




[Paper](#)   [Code](#)

Visual-Language Model’s

CLIP:

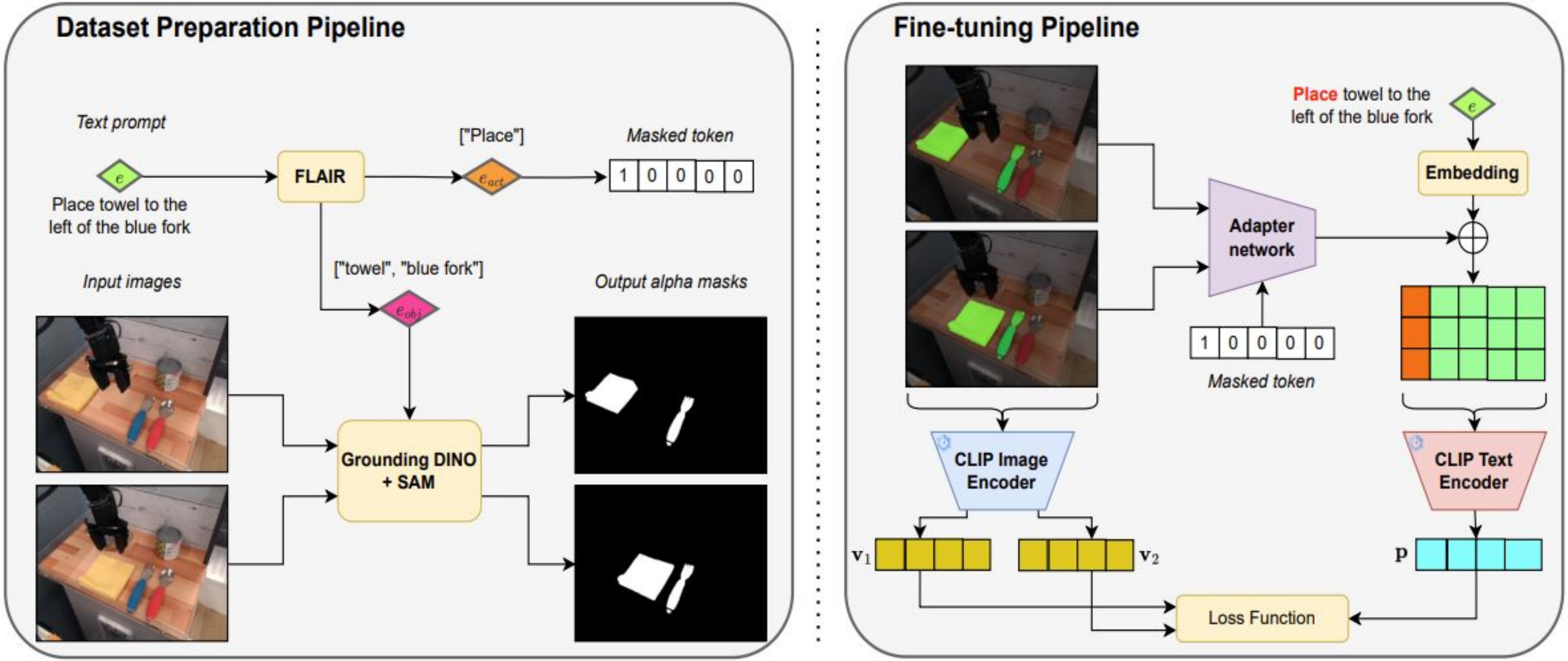


[Paper](#) [Code](#)

Training Pseudo code:

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Robot-CLIP:

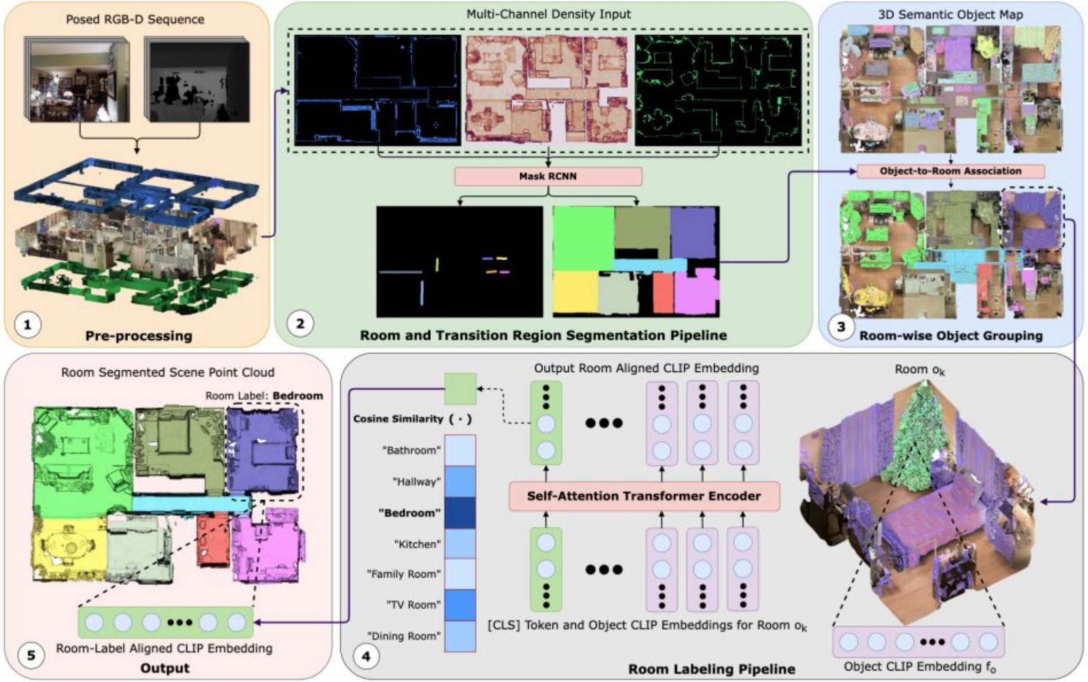


[Paper](#)

Prior Work From RRC:

QueSTMaps: Queryable Semantic Topological Maps for 3D Scene Understanding

Yash Mehan[1*], Kumaraditya Gupta[1*], Rohit Jayanti[1*], Anirudh Govil[1], Sourav Garg[2], and Madhava Krishna[1]

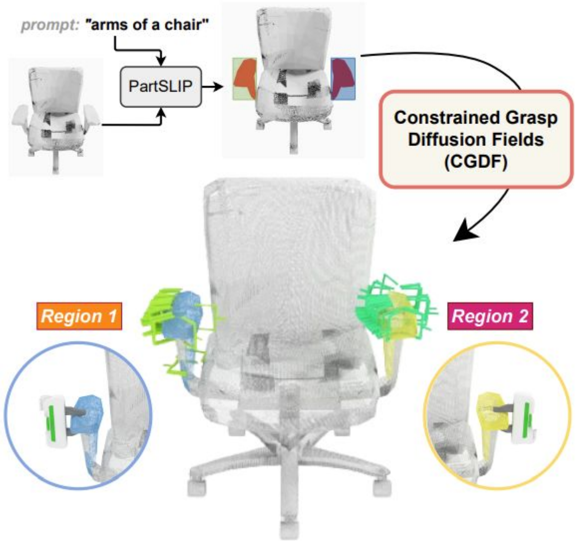


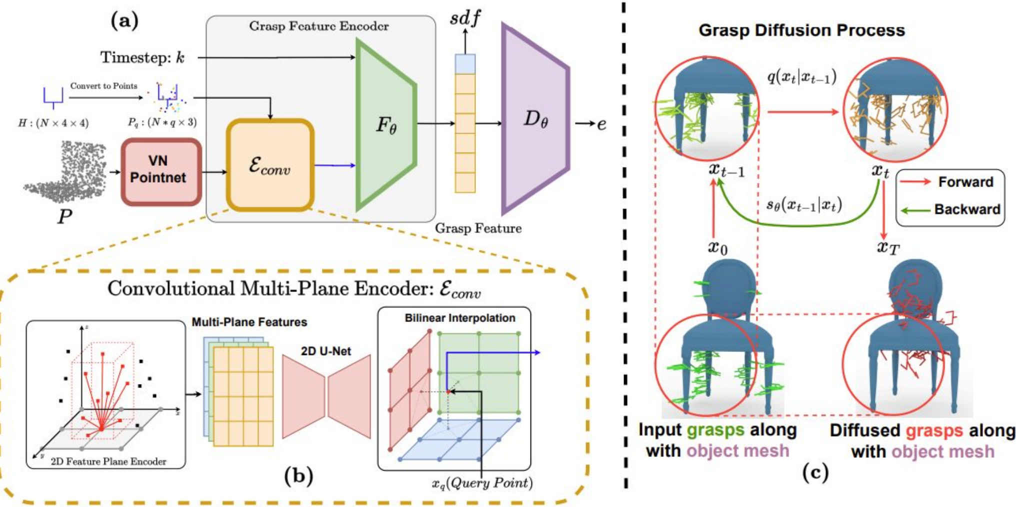
[Paper](#)

Prior Work From RRC:

Constrained 6-DoF Grasp Generation on Complex Shapes for Improved Dual-Arm Manipulation

Gaurav Singh[1*], Sanket Kalwar[1*], Md Faizal Karim[1], Bipasha Sen[2], Nagamanikandan Govindan[1], Srinath Sridhar[3] and K Madhava Krishna[1]





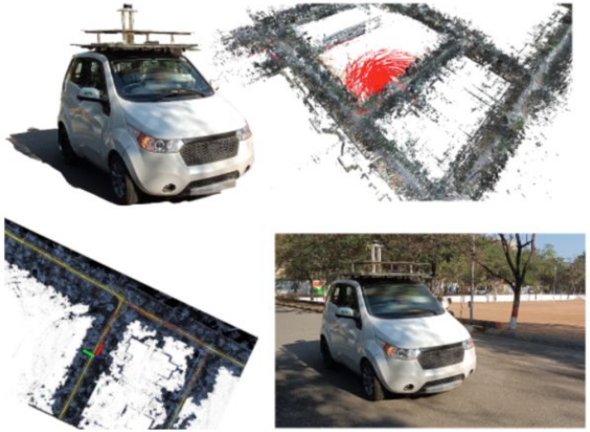
Code Paper

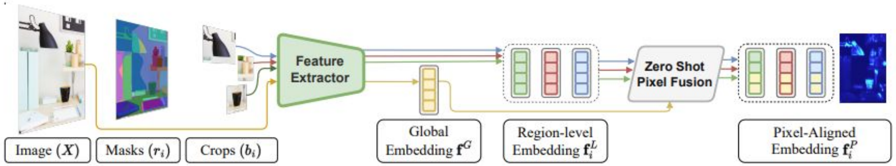
Prior Work From RRC:

ConceptFusion: Open-set Multimodal 3D Mapping

Krishna Murthy Jatavallabhula[1], Alihusein Kuwajerwala[2,†], Qiao Gu[3,†], Mohd Omama[4,†], Tao Chen[1], Alaa Maalouf[1], Shuang Li[1], Ganesh Iyer[7,‡], Soroush Saryazdi[8], Nikhil Keetha[5], Ayush Tewari[1], Joshua B. Tenenbaum[1], Celso Miguel de Melo[6], K. Madhava Krishna[4], Liam Paull[2], Florian Shkurti[3], and Antonio Torralba[1]

[1]MIT, [2]Université de Montréal, [3]University of Toronto, [4]IIIT Hyderabad, [5]CMU, [6]DEVCOM Army Research Lab, [7]Amazon, [8]Concordia University, †Co-second authors, ‡Work done prior to Amazon





[Paper]

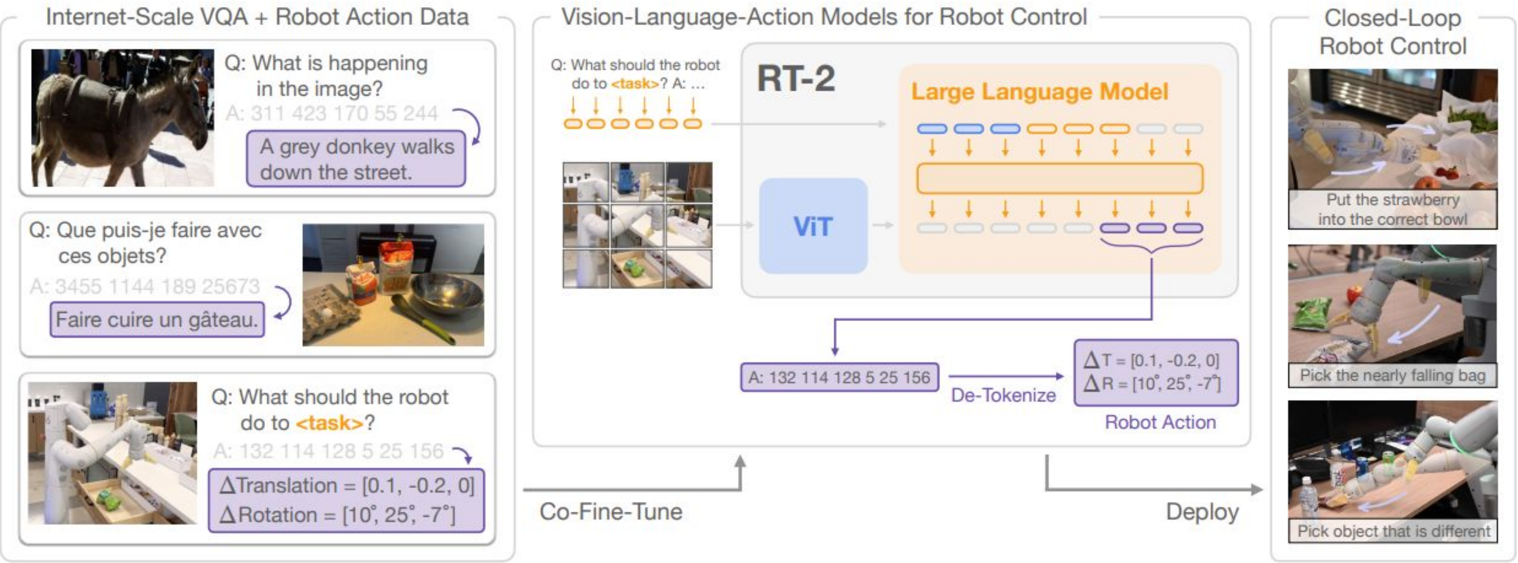
Application To Robotics:

Google DeepMind

https://robotics-transformer2.github.io

28 July 2023

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control



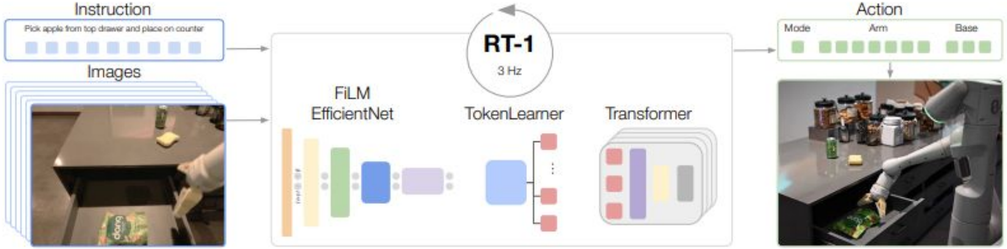
[Paper]

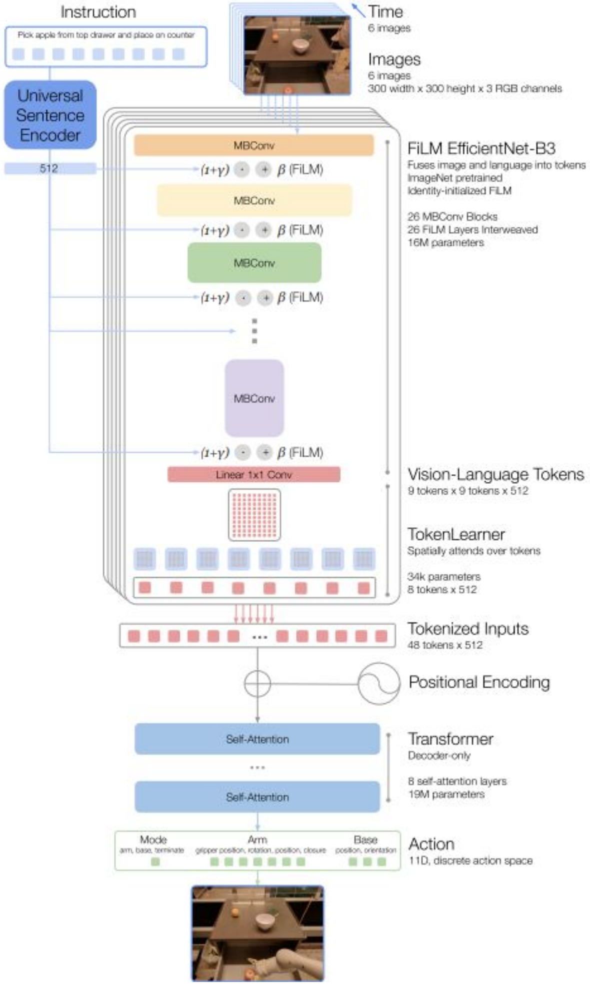
Application To Robotics:

RT-1: Robotics Transformer for Real-World Control at Scale

"pick water bottle from the bottom drawer and put it on the counter"

"move sponge to green jalapeno chips"





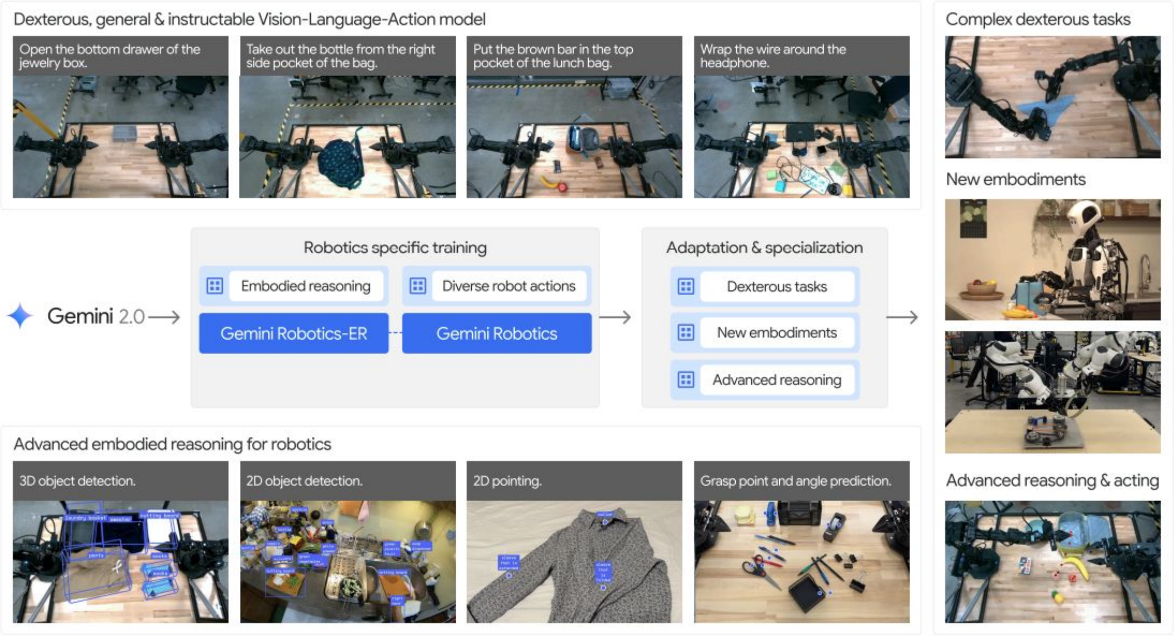
Paper

Application To Robotics:

Google DeepMind

Gemini Robotics: Bringing AI into the Physical World

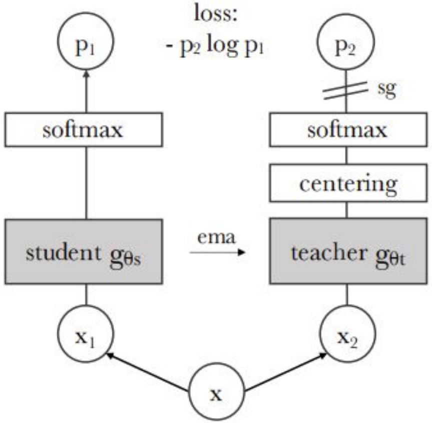
Gemini Robotics Team, Google DeepMind[1]



DINO & DINOV2

DINO:

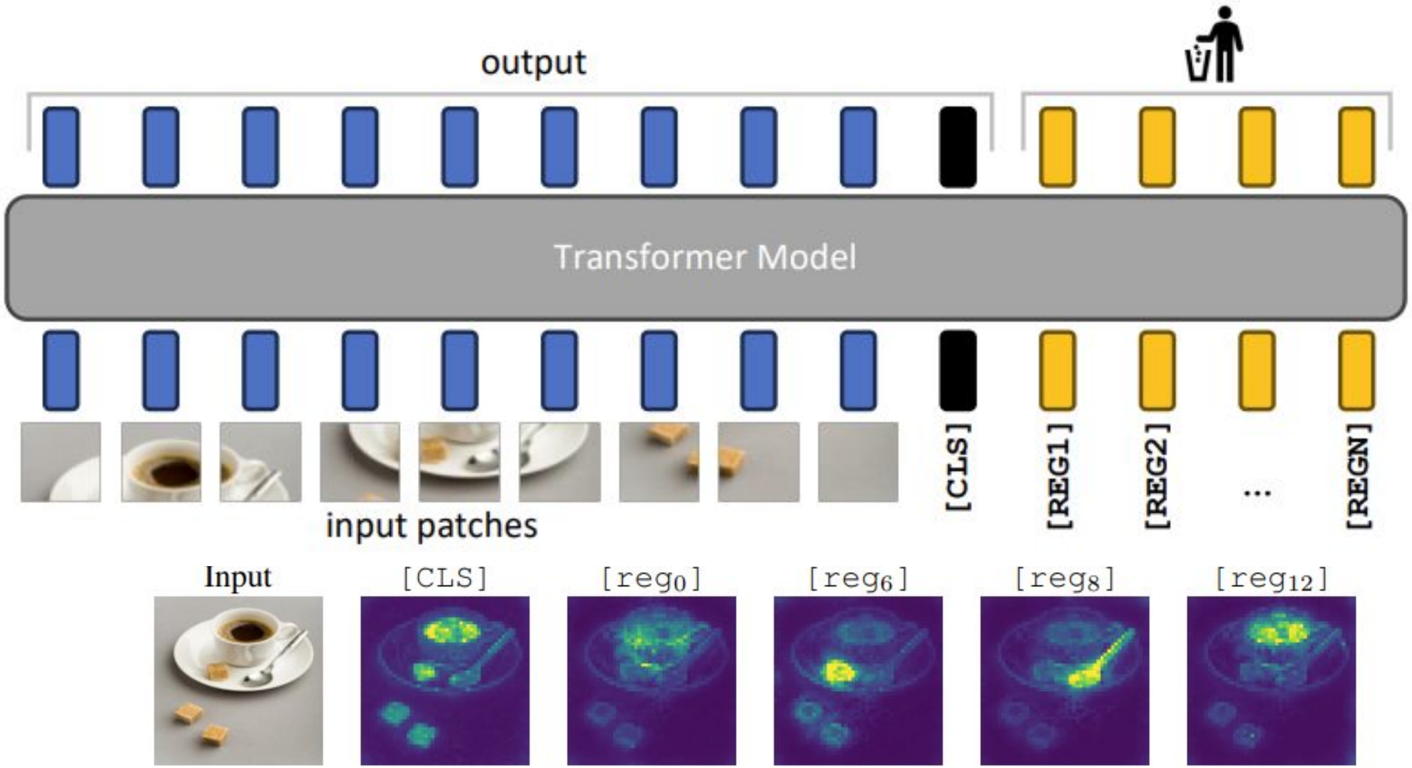
	Random	Supervised	DINO
ViT-S/16	22.0	27.3	45.9
ViT-S/8	21.8	23.7	44.7



$$\min_{\theta_s} H(P_t(x), P_s(x))$$

[Code](#) [Paper](#)

DINO2:



[Paper](#) [Code](#)

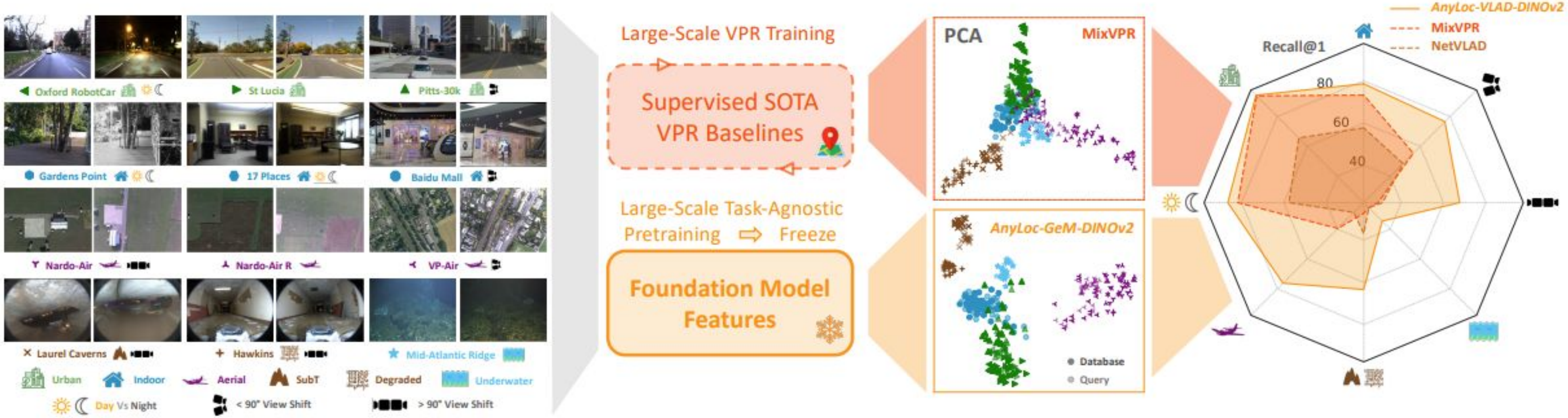
Prior Work From RRC:

AnyLoc: Towards Universal Visual Place Recognition

https://anyloc.github.io/

Nikhil Keetha*[1], Avneesh Mishra*[2], Jay Karhade*[1], Krishna Murthy Jatavallabhula[3], Sebastian Scherer[1], Madhava Krishna[2], and Sourav Garg[4]

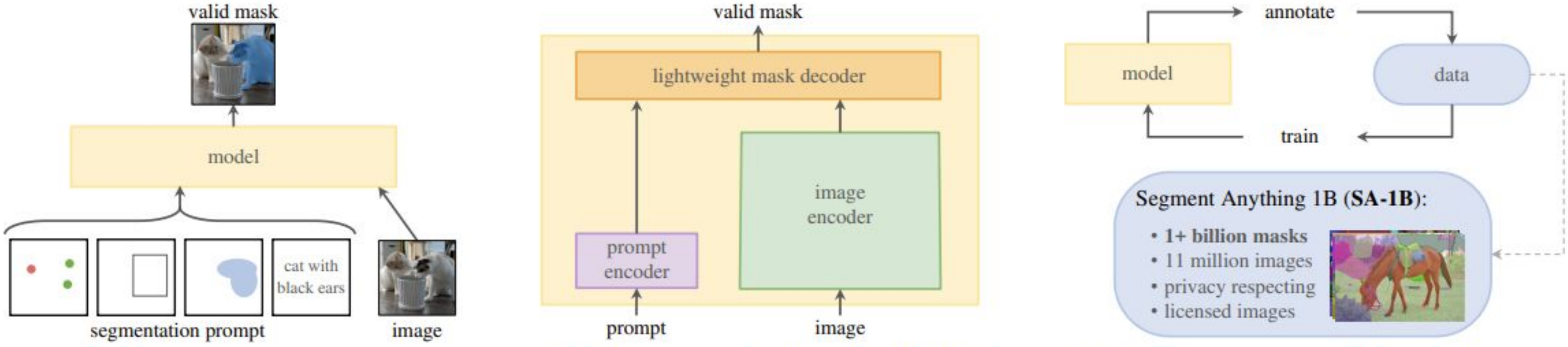
[1]CMU, [2]IIIT Hyderabad, [3]MIT, [4]University of Adelaide



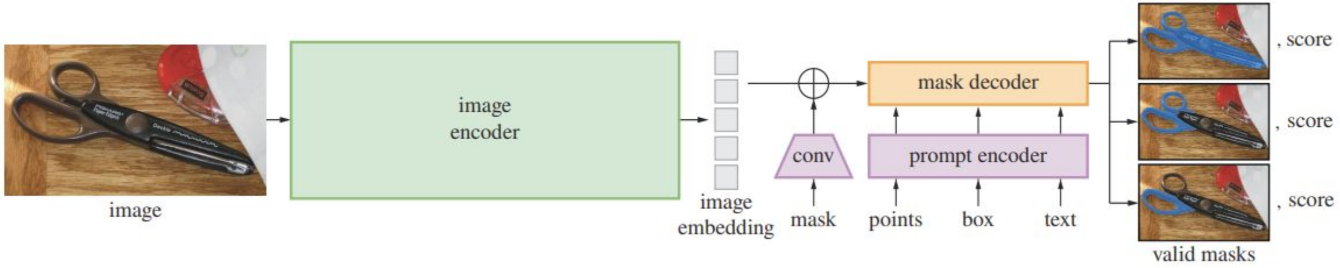
[Paper](#)

SAM & SAM2

Segment Anything:

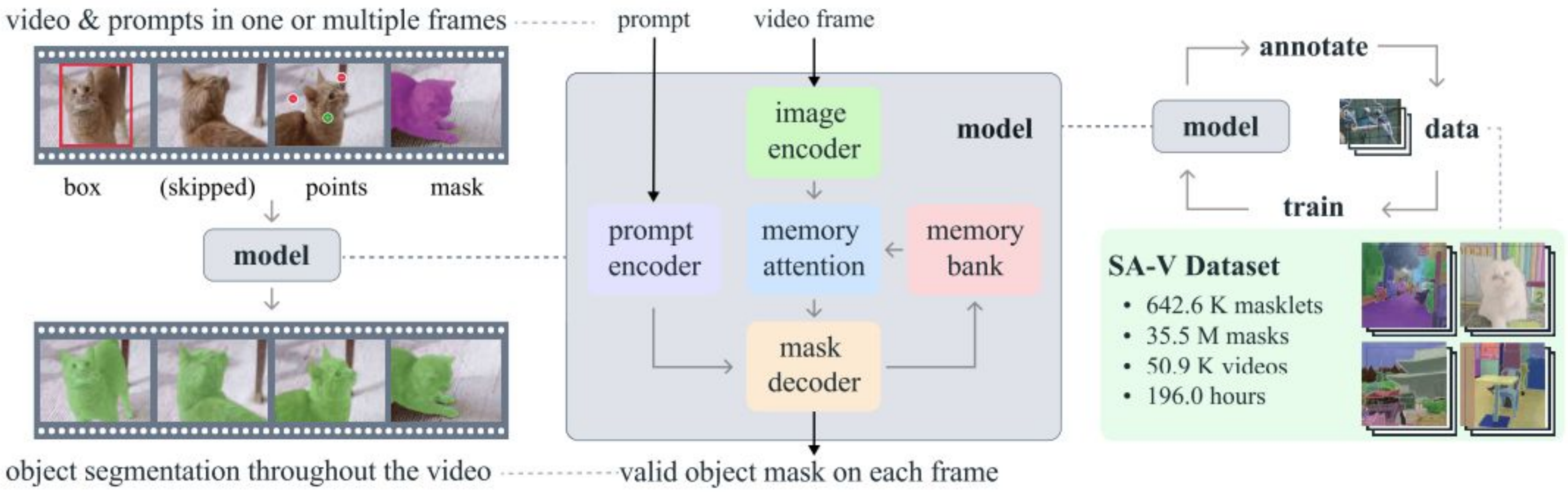


(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)



[Paper] [Code]

SAM2:



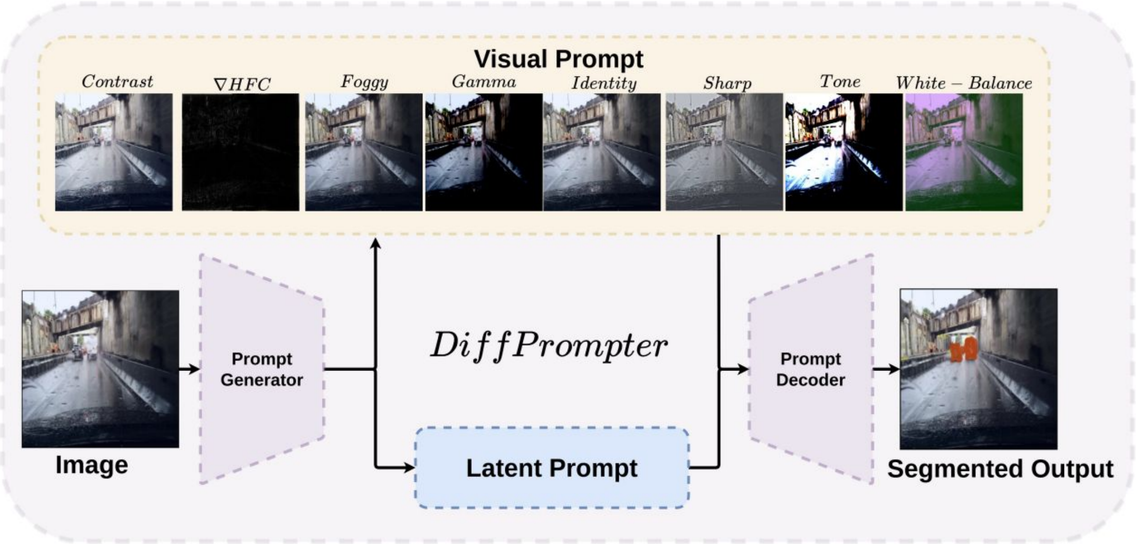
(a) Task: promptable visual segmentation **(b) Model:** Segment Anything Model 2 **(c) Data:** data engine and dataset

[Paper](#) [Code](#)

Prior Work From RRC:

DiffPrompter: Differentiable Implicit Visual Prompts for Semantic-Segmentation in Adverse Conditions

Sanket Kalwar*[1], Mihir Ungarala*[1], Shruti Jain*[1], Aaron Monis[1], Krishna Reddy Konda[3]
Sourav Garg[2], K Madhava Krishna[1]



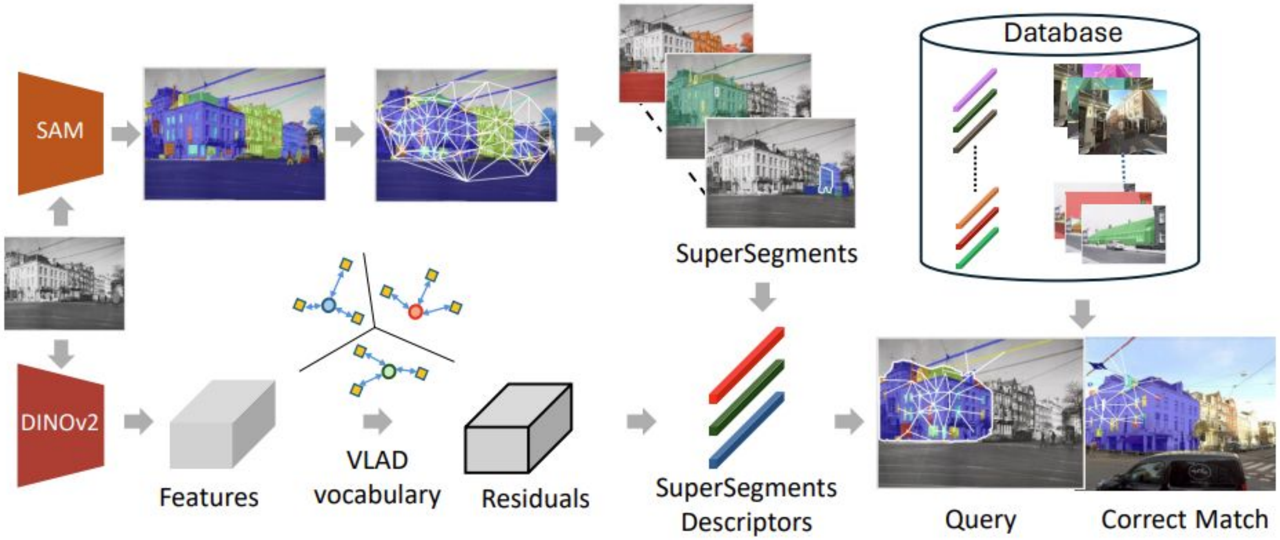
[Paper](#)

[Code](#)

Prior Work From RRC:

Revisit Anything: Visual Place Recognition via Image Segment Retrieval

Kartik Garg*[1], Sai Shubodh Puligilla*[2], Shishir Kolathaya[1], Madhava Krishna[2], and Sourav Garg[3]



[Paper](#)

[Code](#)

Thank You!