Physics → ML @ Google

Kanishka Rao,
Google Deepmind

Aug 2024

Machine Learning for Fundamental Physics School 2024

- My experience going from physics to machine learning in industry

- Machine learning concepts and takeaways

Disclaimer: YMMV

Google has ~175,000 employees worldwide

Machine learning is happening at many places: large corporations, mid-size corporations, startups

Hiring and work experiences change frequently

About Me

- BSc at UCLA, Astrophysics. 2007

- PhD at UC Irvine, experimental particle physics. 2013

- Google, 2013-present
  Speech Recognition, 2013-2018
  Robotics, 2018-present

PhD, UC Irvine

Advised by Daniel Whiteson

Searches for new physics in the top, higgs and dark matter sectors at high energy particle experiments

Selected Work

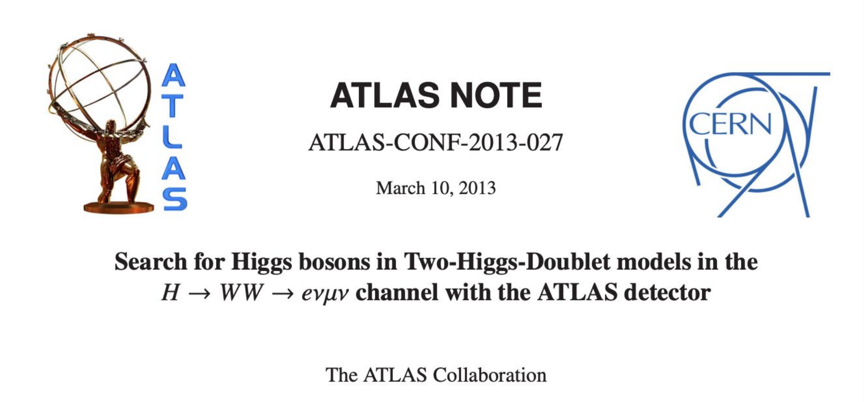
Search for a heavy particle decaying to a top quark and a light quark in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV

CDF Collaboration

We present a search for a new heavy particle M produced in association with a top quark, $p\bar{p} \rightarrow t(M \rightarrow \bar{t}q)$ or $p\bar{p} \rightarrow \bar{t}(\bar{M} \rightarrow t\bar{q})$, where q stands for up quarks and down quarks. Such a particle may explain the recent anomalous measurements of top-quark forward-backward asymmetry. If the light-flavor quark (q) is reconstructed as a jet (j), this gives a $\bar{t} + j$ or $t + j$ resonance in $t\bar{t}$+jet events, a previously unexplored experimental signature. In a sample of events with exactly one lepton, missing transverse momentum and at least five jets, corresponding to an integrated luminosity of 8.7 fb^{-1} collected by the CDF II detector, we find the data to be consistent with the standard model. We set cross-section upper limits on the production ($p\bar{p} \rightarrow Mt$ or $\bar{M}\bar{t}$) at 95% confidence level from 0.61 pb to 0.02 pb for M masses ranging from 200 GeV/c^2 to 800 GeV/c^2, respectively.

- Learned basic C++
- Data analysis concepts like filtering, plotting, fitting hypothesis
- Working with a large collaboration

Selected Work



ATLAS NOTE

ATLAS-CONF-2013-027

March 10, 2013

Search for Higgs bosons in Two-Higgs-Doublet models in the $H \to WW \to e\nu\mu\nu$ channel with the ATLAS detector

The ATLAS Collaboration

- First introduction to neural networks
- Complicated software infrastructure used by a lot of people
- Learned how to effectively communicate findings

Jobs: Google?

- Physics classmate had joined Google
- Why would Google hire physics PhDs?
- What would it like working at a software company?

Google Roles?

Roles

Software Engineer

Your work is at the core of everything we build. Develop massive, complex software systems that scale globally.

Network Engineer

Design and implement enterprise and carrier network systems and architecture vital to Google's operations.

Research Scientist

You move beyond the lab, working closely with Software Engineers and others to discover, invent, and build at the largest scale.

Product Manager

Architect the future of our products by bridging engineering and business as you manage a product's full lifecycle, from strategic planning to development and launch.

Technical Program Manager

Rely on your strong technical experience to oversee all the essential activities of a particular program, including planning, communications, and execution.

Security/Privacy Engineer

Hack Google... if you can. Work on finding security flaws, building secure infrastructure, or ensuring data privacy as part of a diverse engineering team.

Sourcing/Supply Chain

Own relationships with Google's strategic internal and external partners in order to manage Google's inventory and procurement needs.

UX Specialist

Our passionate, interdisciplinary UX specialists and designers work across platforms while exemplifying one of Google's core principles: "Focus on the user and all else will follow."

Systems Integrator

Integrate systems :)

Data Center Technician

Install, test, and maintain hardware and systems software for Google's data centers.

Systems Engineer

Drive systems and software reliability by engineering tools to manage the efficiency of Google services across the globe.

Solutions Consultant

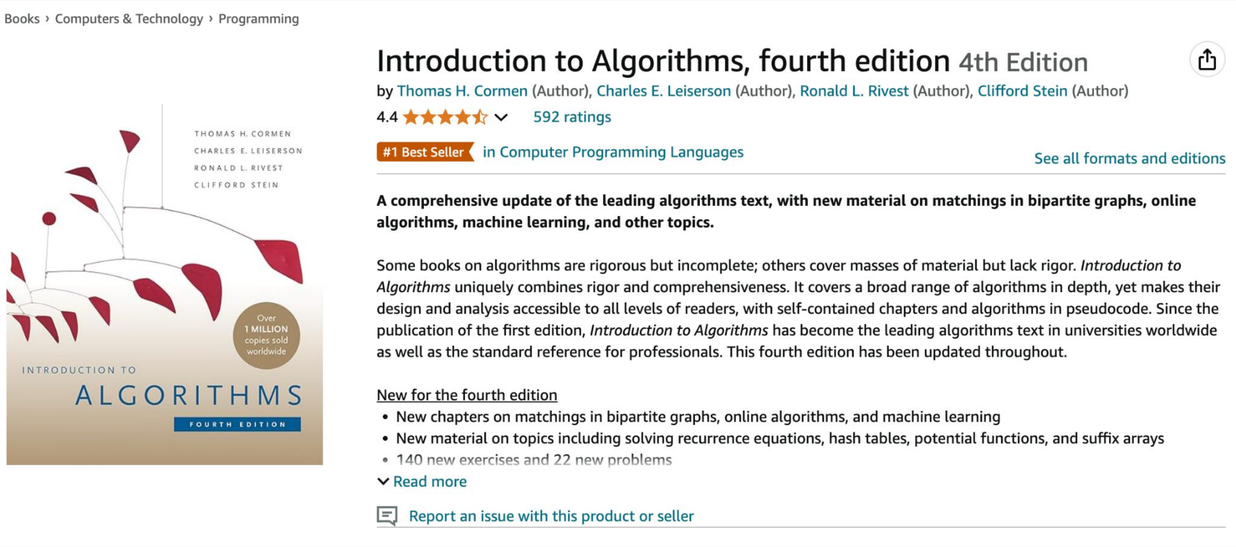
Manage vital business relationships and troubleshoot complex engineering problems in this hybrid Tech/Business role.

Google: Software Engineer

- Research Scientist (RS): requires research presentations and interviews
- New Grad software engineer: generic software engineering interviews
  - Requires CS or similar technical PhD
- Don't apply for a specific team, match with teams after clearing the interview bar
- Apply for roles at large campuses like Mountain View or New York for best team matches
- Rewrite your resume for Industry
- Prepare for the coding interviews!

The interview

- Most physicists fail this step
- You will need to study to pass this interview
  - I spent 3 months prepping



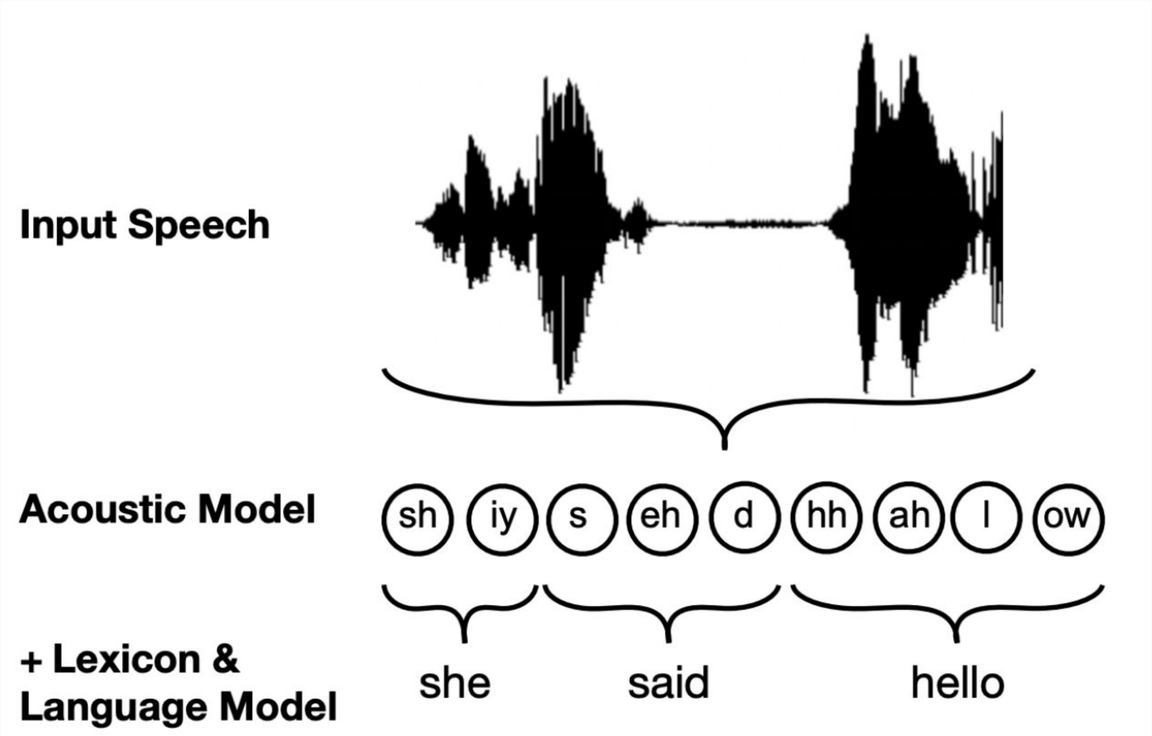
- Once you are prepared, go wide with your interviews

The Offer

- Once you get a Google offer things are actually a lot flexible
  - Recruiter will help find you a team
    - Team that is doing machine learning and publishes papers
  - People change teams often
  - You can also change your role → RS later
  - Easier to move to other industry positions

Speech Recognition

Automatic Speech Recognition



Acoustic Models

- Larger machines available for DNN training
- More speech data available

Expert handcrafted functions → Deep Learning

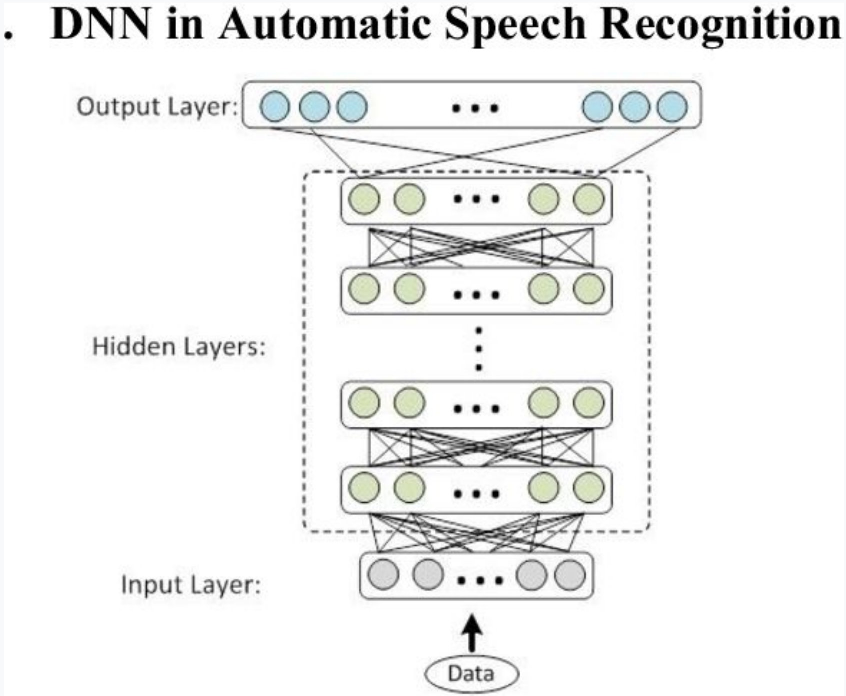


Fig. 1 DNN used in ASR systems

Expert handcrafted functions → Deep Learning

- Speech is a 50-year old academic field
- Machine learning breakthroughs became Speech Recognition breakthroughs
- Speech expertise was not as critical
  - Worked with people with PhDs in Speech research, Signal Processing, Linguistics, Language
- Generalists thrived over specialists

Acoustic Models

DNN are not enough

. **DNN in Automatic Speech Recognition**

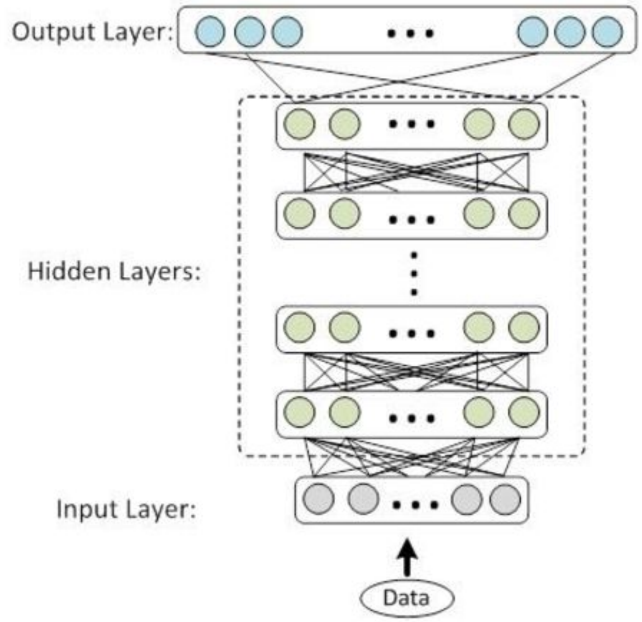


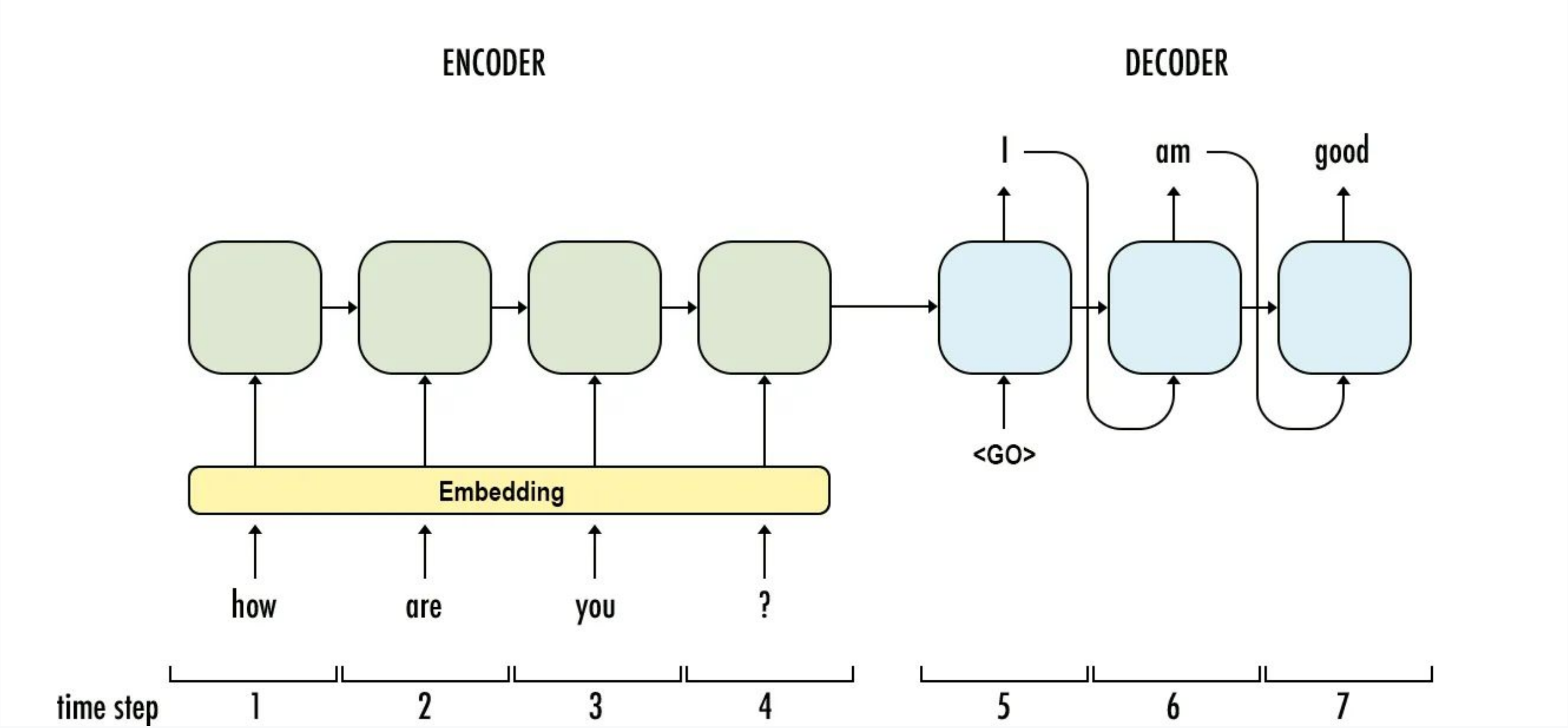
Fig. 1 DNN used in ASR systems

Sequence Problems

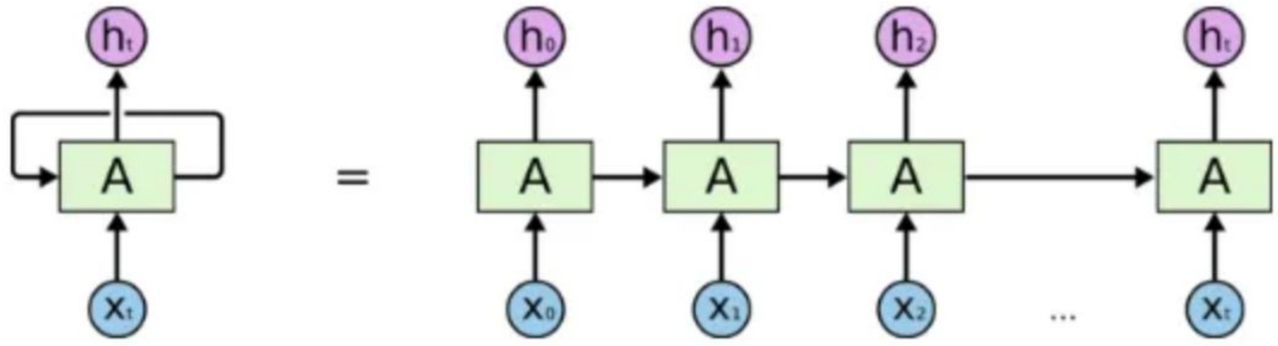
Temporal context matters: current outputs depend on the history

- Speech recognition
- Translate
- Language modeling
- Playing starcraft
- Playing Go
- Robotics

Sequence Models

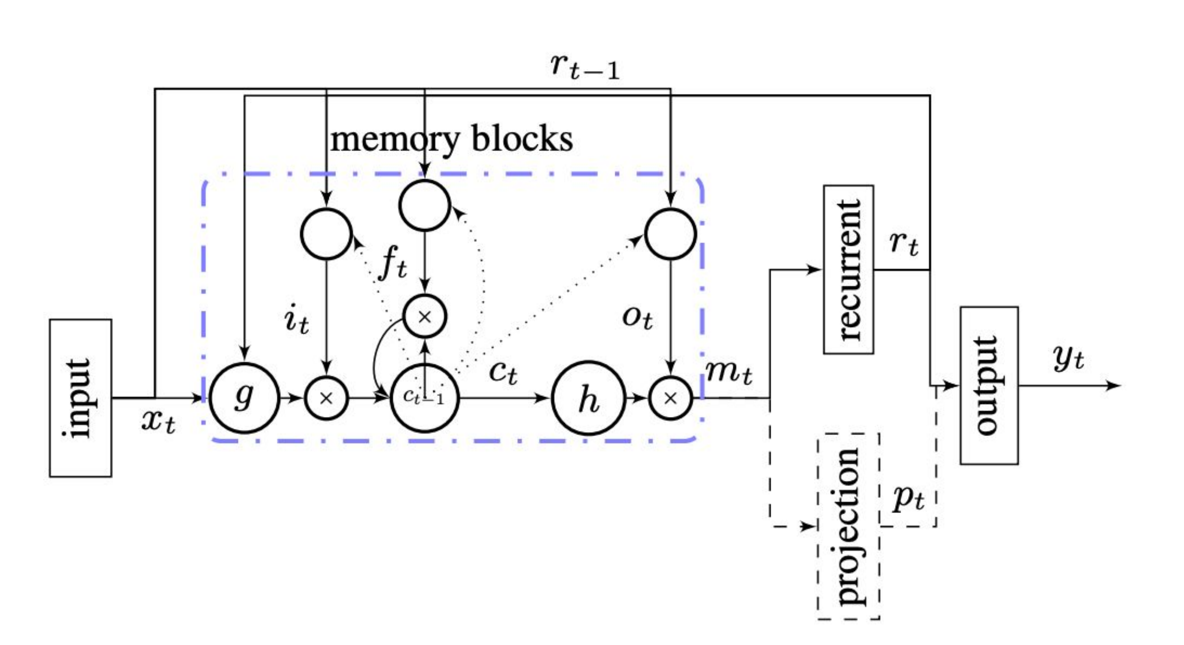


Recurrent Neural Networks



An unrolled recurrent neural network.

LSTM

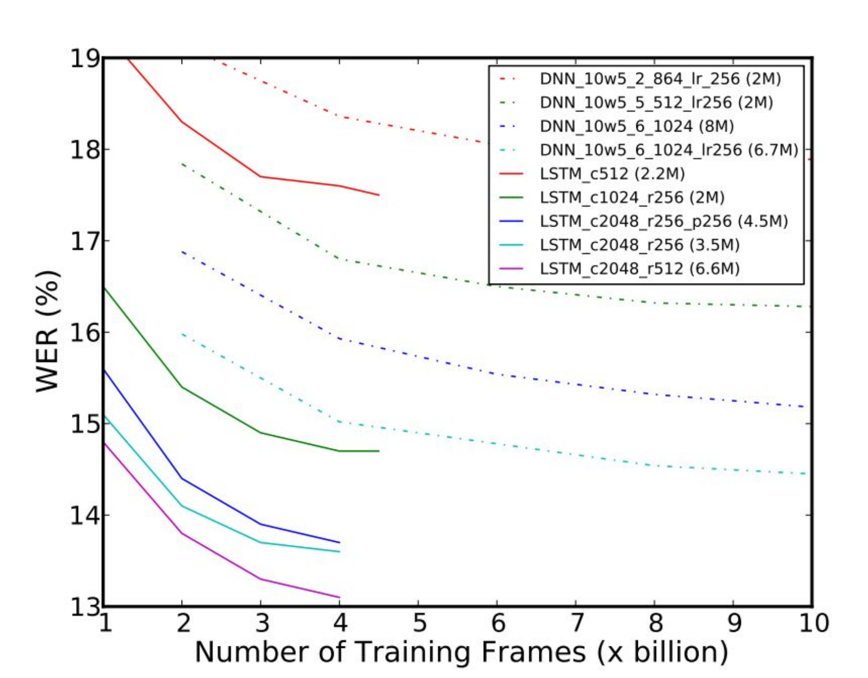

r_{t-1}
memory blocks
f_t
i_t
o_t
input
x_t
g
c_{t-1}
c_t
h
m_t
recurrent
r_t
output
y_t
projection
p_t

LONG SHORT-TERM MEMORY BASED RECURRENT NEURAL NETWORK ARCHITECTURES FOR LARGE VOCABULARY SPEECH RECOGNITION

Haşim Sak, Andrew Senior, Françoise Beaufays

Word Error Rates dropping to near 10-15%

Most pre-ML error rates were >25%



Starter Project

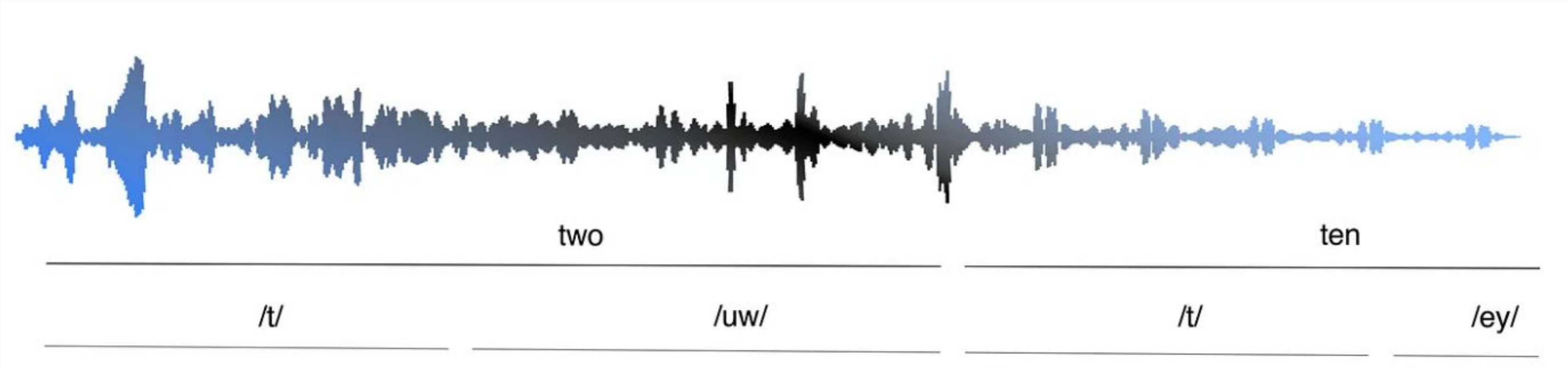
Implement software for LSTM training

- Code reviews
- Collaborating with many other code contributors
- Testing
- Software for others
- Software design
- Quality and documentation

The Alignment Problem

Not every input needs an output.

Sample audio at 10 samples per second → a few phonemes a second

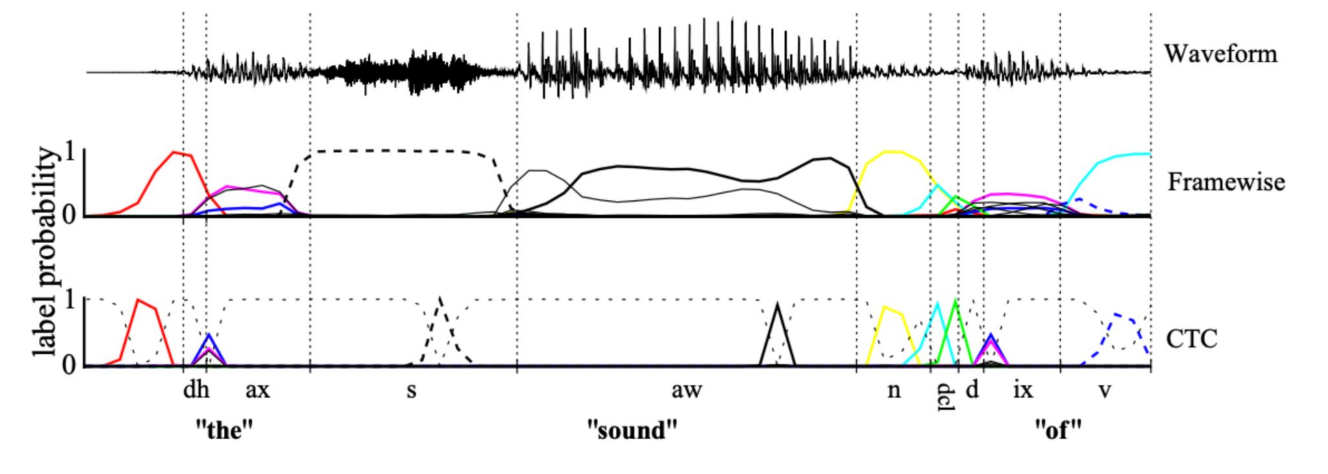


Training requires data alignment between input speech signals and output phonemes

CTC Loss

A special loss function that introduces a *blank* label

Loss sums up likelihoods of the correct output sequence with all possible alignments



Model now learns phoneme recognition and alignment

⟶ Make deep learning models do more of the problem.

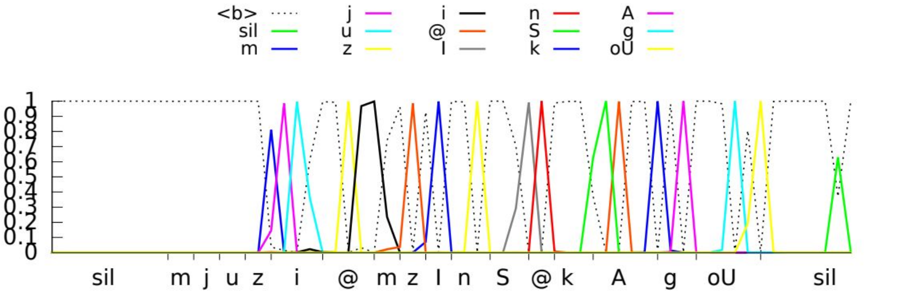
Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition

Haşim Sak, Andrew Senior, Kanishka Rao, Françoise Beaufays

Google
{hasim,andrewsenior,kanishkarao,fsb}@google.com

- 12% WER
- Paper writing, making plots
- Peer review
- Presented at conference: InterSpeech

Vocabulary	OOV	WER (%)	In vocab. WER (%)
25k Word	4.8	19.5	14.5
7k Word	13	26.8	11.8



(a) unidirectional phone CTC + sMBR

Presented at major speech conferences

INTERSPEECH 2024

Scientific Areas and Topics

Interspeech 2024 embraces a broad range of science and technology in speech, language and communication, including – but not limited to – the following topics:

• Speech Perception, Production and Acquisition	• Speech Synthesis
• Phonetics, Phonology, and Prosody	• Spoken Language Generation
• Paralinguistics in Speech and Language	• Automatic Speech Recognition
• Analysis of Conversation	• Spoken Dialogue and Conversational AI Systems
• Speech, Voice, and Hearing Disorders	• Spoken Language Translation, Information Retrieval, Summarization
• Speaker and Language Identification	• Technologies and Systems for New Applications
• Speech and Audio Signal Analysis	• Resources and Evaluation
• Speech Coding and Enhancement	• Beyond traditional speech topics (not limited to the provided list)

https://interspeech2024.org

Translate

- Google translate was similar improvements from sequence modeling improvements
- Introduction of the *attention mechanism*

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

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

Attention Mechanism

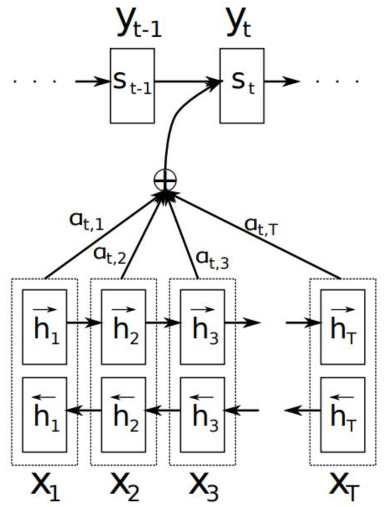


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

- The network can *decide* which inputs to *pay attention* to for the current output

The context vector c_i is, then, computed as a weighted sum of t annotations h_i:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

The weight α_{ij} of each annotation h_j is computed by

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

where

$$e_{ij} = a(s_{i-1}, h_j)$$

Attention Mechanism

- Did not need to pass information via time (LSTM)
- Very long range context
- Alignment problem naturally solved
  - You can see heat map of activations to see alignment
- Easily stacked
- Fewer complicated loss functions

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

We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [16]	23.75			
Deep-Att + PosUnk [35]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [34]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [29]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [35]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [34]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

STATE-OF-THE-ART SPEECH RECOGNITION WITH SEQUENCE-TO-SEQUENCE MODELS

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, Michiel Bacchiani

Google, USA

{chungchengc,tsainath,yonghui,prabhavalkar,drpng,zhifengc,anjuli ronw,kanishkarao,kgonina,ndjaitly,boboli,chorowski,michiel}@google.com

Exp-ID	Model	WER
E2	WPM	9.0
E3	+ MHA	8.0
E4	+ Sync	7.7
E5	+ SS	7.1
E6	+ LS	6.7
E7	+ MWER	5.8

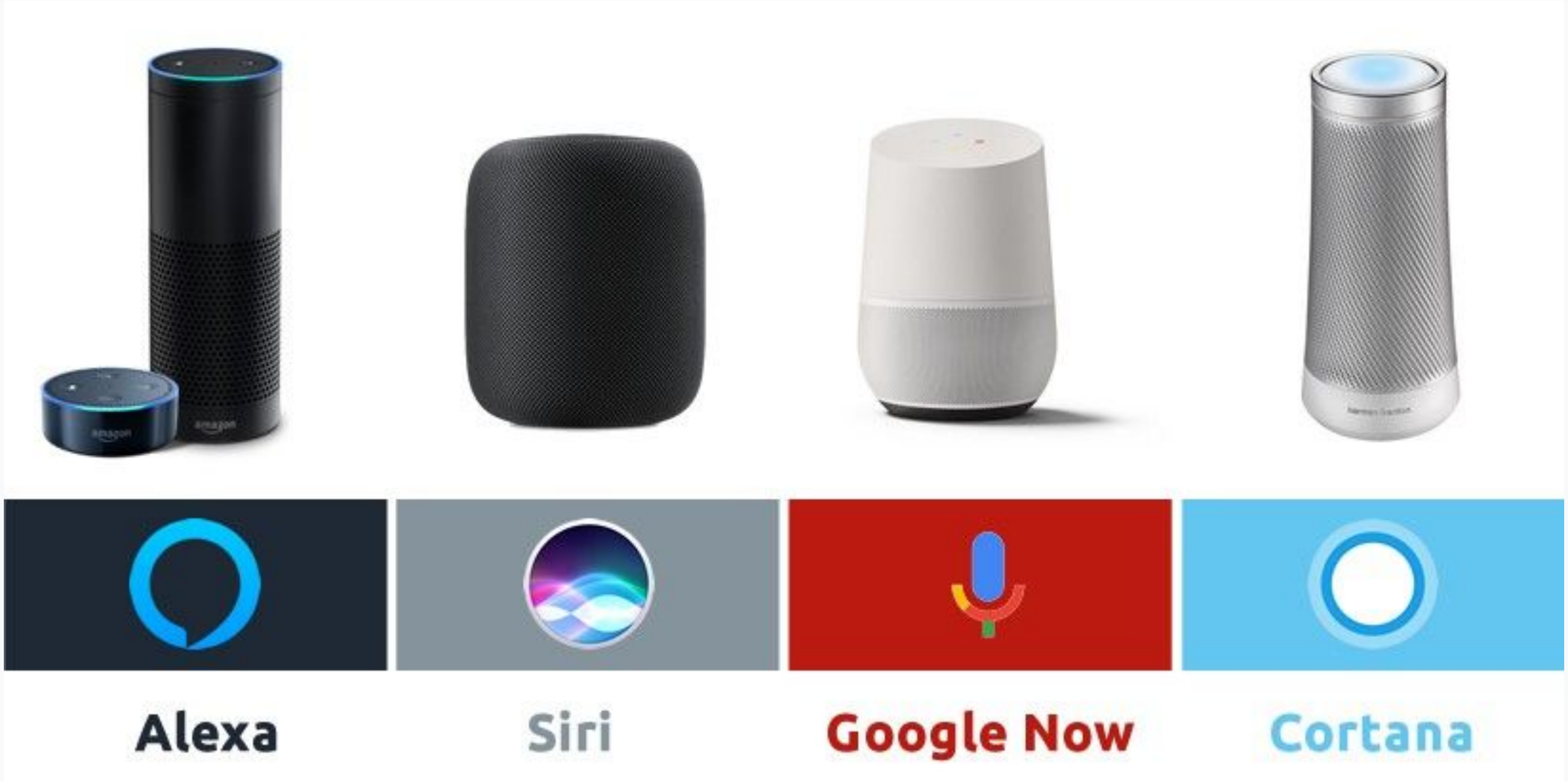
- After 50 years of speech research finally WER were <10%
- Research problem ⟶ useful product

Google I/O: Launch of Google Home

Proprietary + Confidential

Google

Speech Recognition was in the world

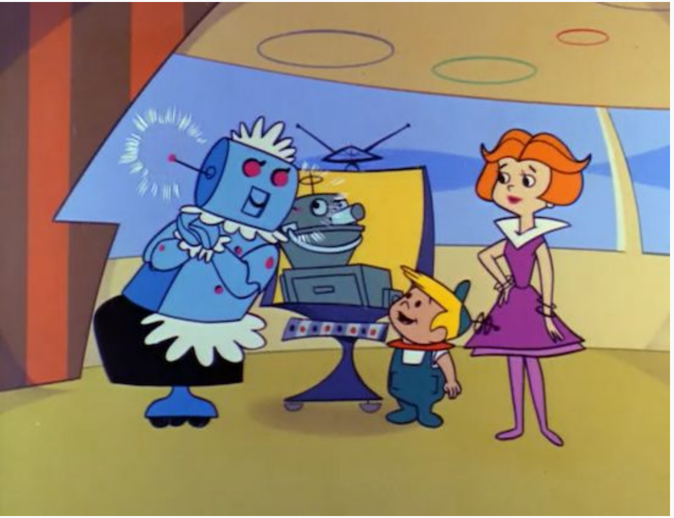


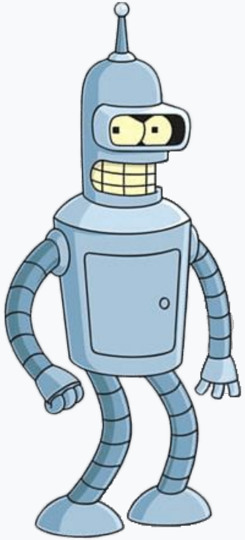
Research → Useful thing

- The paper is only 50% of the breakthroughs
    - Works in noisy environments
    - Needs to be fast
    - Needs to deal with multiple accents
    - Needs to be computationally cheap
- The ultimate test is usefulness
    - 100M queries a week

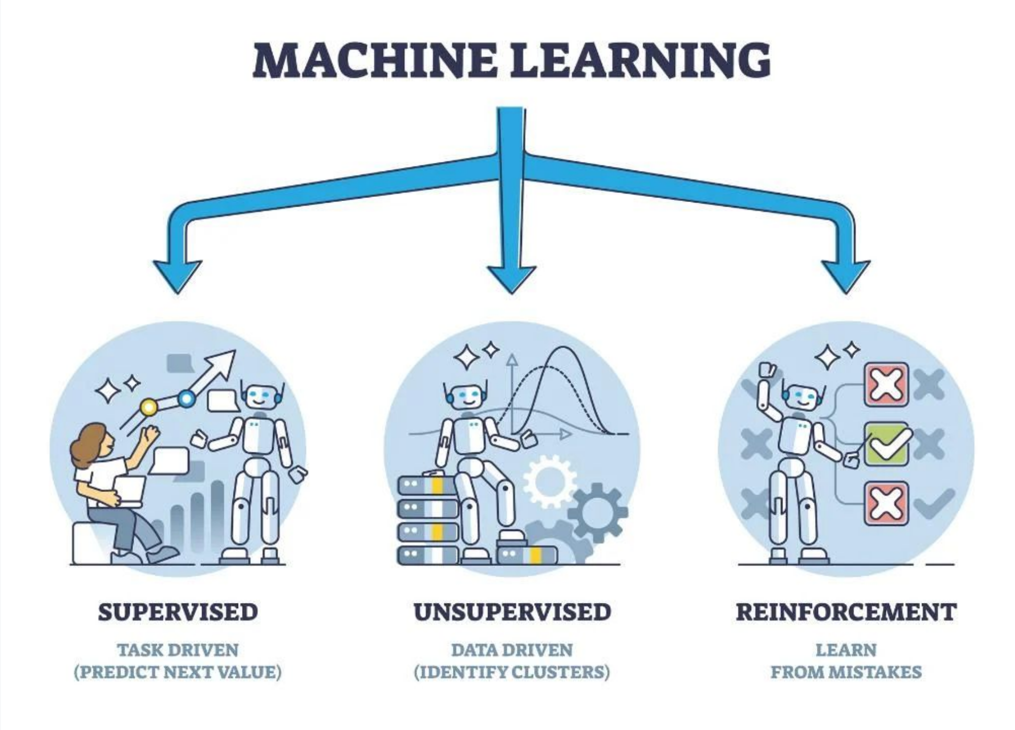
Robotics

General Purpose Robotics

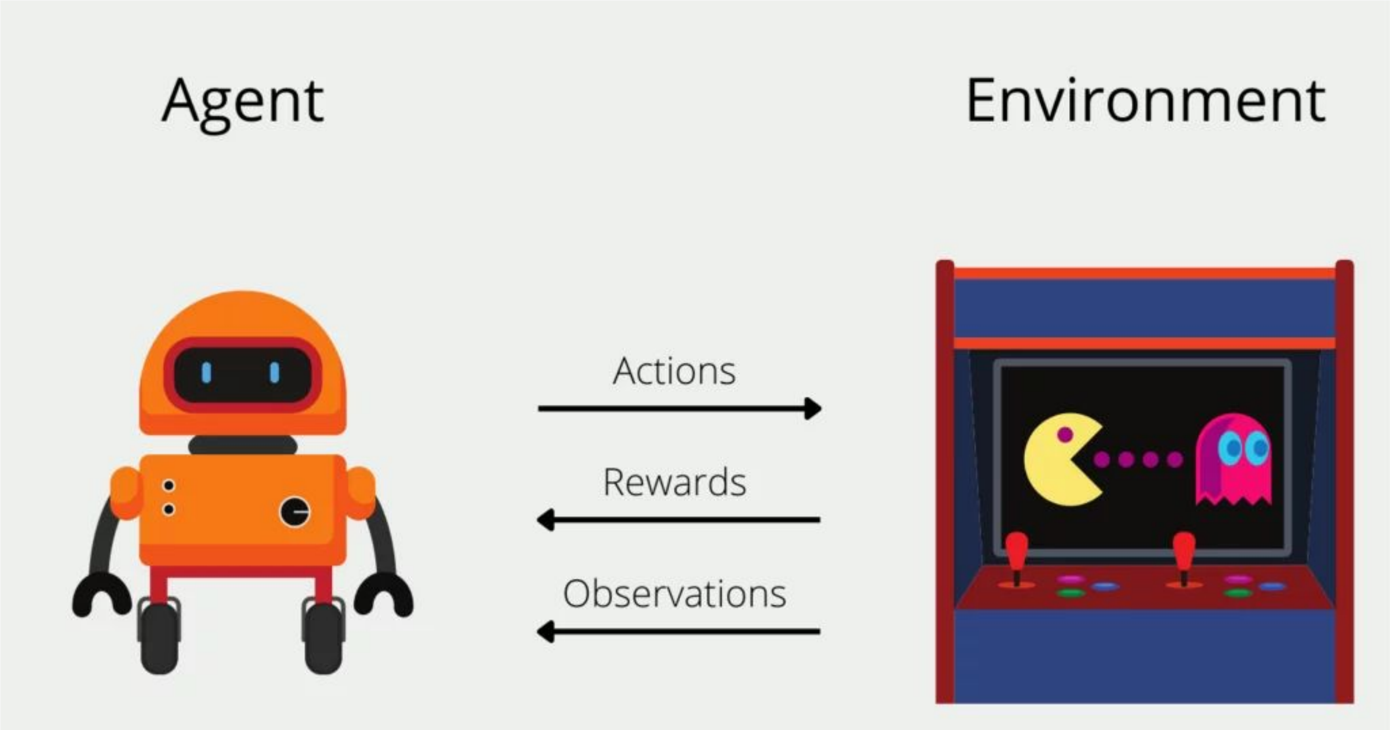




Reinforcement Learning



Reinforcement Learning



Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih **Koray Kavukcuoglu** **David Silver** **Alex Graves** **Ioannis Antonoglou**

Daan Wierstra **Martin Riedmiller**

DeepMind Technologies

{vlad,koray,david,alex.graves,ioannis,daan,martin.riedmiller} @ deepmind.com

Article

Grandmaster level in StarCraft II using multi-agent reinforcement learning

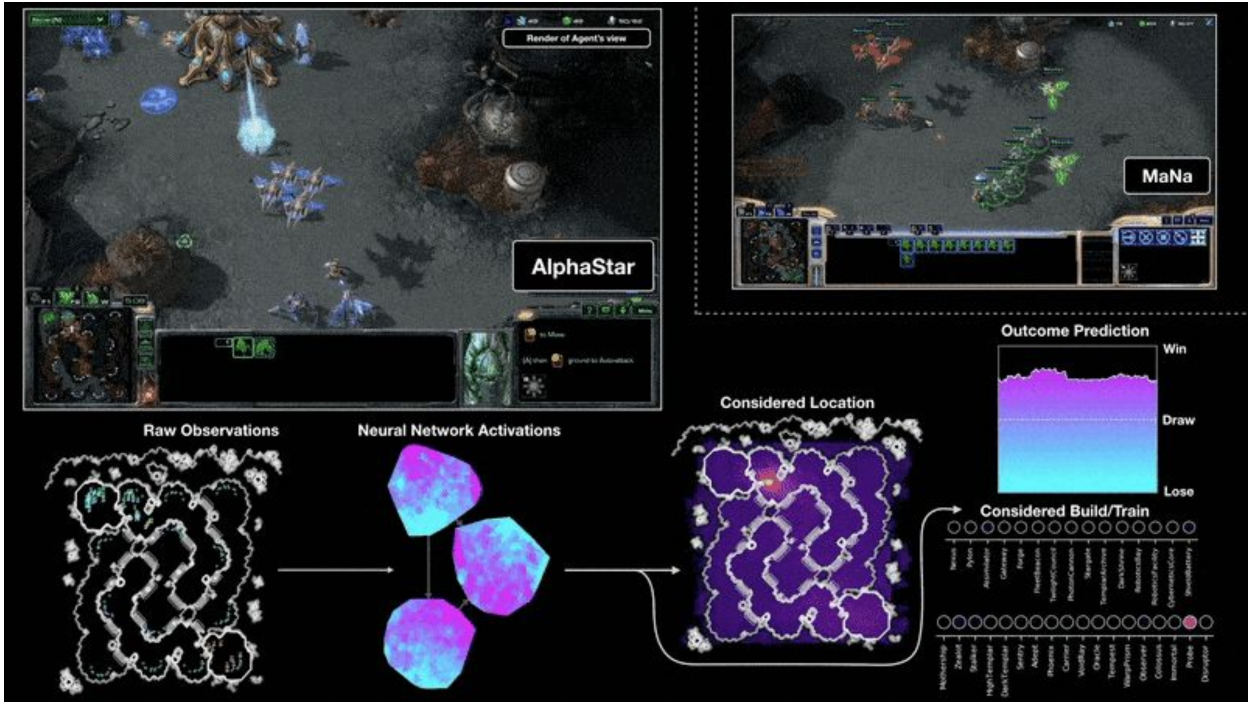
https://doi.org/10.1038/s41586-019-1724-z

Received: 30 August 2019

Accepted: 10 October 2019

Published online: 30 October 2019

Oriol Vinyals[1,3]*, Igor Babuschkin[1,3], Wojciech M. Czarnecki[1,3], Michaël Mathieu[1,3], Andrew Dudzik[1,3], Junyoung Chung[1,3], David H. Choi[1,3], Richard Powell[1,3], Timo Ewalds[1,3], Petko Georgiev[1,3], Junhyuk Oh[1,3], Dan Horgan[1,3], Manuel Kroiss[1,3], Ivo Danihelka[1,3], Aja Huang[1,3], Laurent Sifre[1,3], Trevor Cai[1,3], John P. Agapiou[1,3], Max Jaderberg[1], Alexander S. Vezhnevets[1], Rémi Leblond[1], Tobias Pohlen[1], Valentin Dalibard[1], David Budden[1], Yury Sulsky[1], James Molloy[1], Tom L. Paine[1], Caglar Gulcehre[1], Ziyu Wang[1], Tobias Pfaff[1], Yuhuai Wu[1], Roman Ring[1], Dani Yogatama[1], Dario Wünsch[2], Katrina McKinney[1], Oliver Smith[1], Tom Schaul[1], Timothy Lillicrap[1], Koray Kavukcuoglu[1], Demis Hassabis[1], Chris Apps[1,3] & David Silver[1,3]*

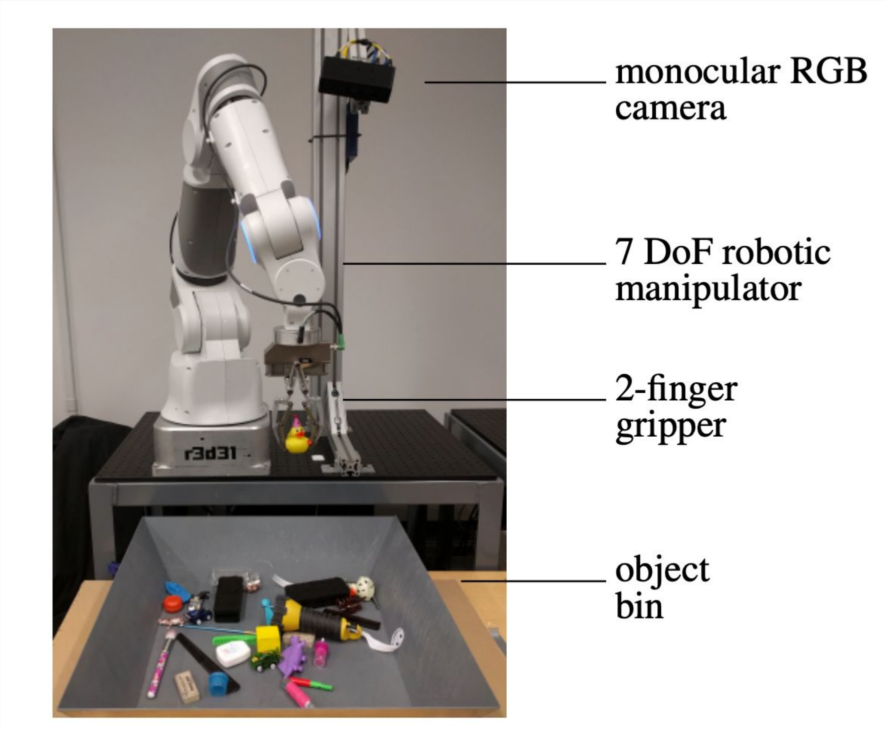


QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation

Dmitry Kalashnikov[1], Alex Irpan[1], Peter Pastor[2], Julian Ibarz[1], Alexander Herzog[2], Eric Jang[1], Deirdre Quillen[3], Ethan Holly[1], Mrinal Kalakrishnan[2], Vincent Vanhoucke[1], Sergey Levine[1,3]
{dkalashnikov, alexirpan, julianibarz, ejang, eholly, vanhoucke, slevine}@google.com,
{peterpastor, alexherzog, kalakris}@x.team, {deirdrequillen}@berkeley.edu

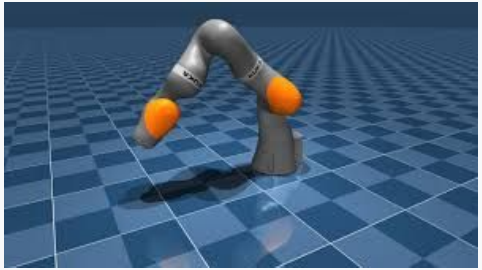
Robot Manipulation

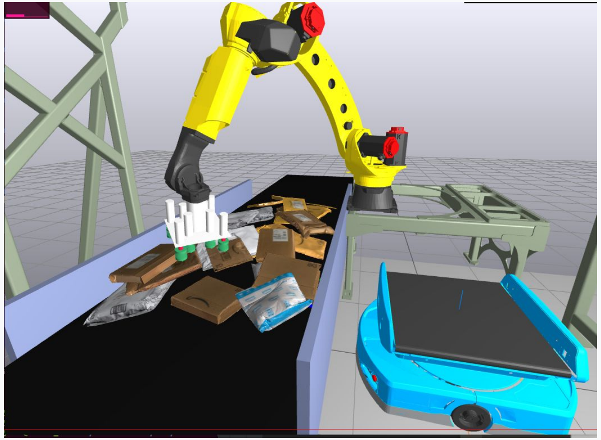
Hand-eye coordination problem → continuous image classification problem

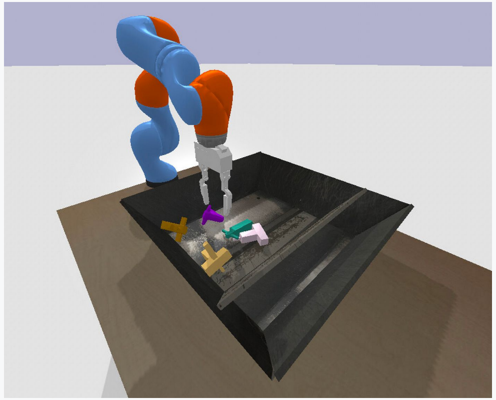


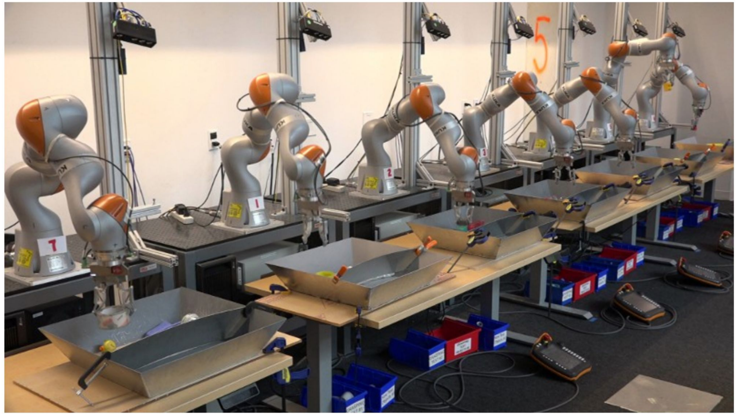
Starter Project

Use simulation data for training real robots: *sim2real gap*



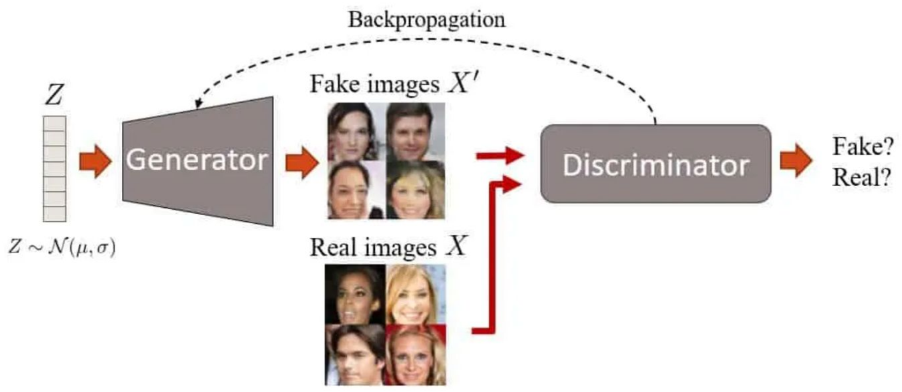






Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7



source

LARGE SCALE GAN TRAINING FOR HIGH FIDELITY NATURAL IMAGE SYNTHESIS

Andrew Brock[*, †]
Heriot-Watt University
ajb5@hw.ac.uk

Jeff Donahue[†]
DeepMind
jeffdonahue@google.com

Karen Simonyan[†]
DeepMind
simonyan@google.com

Figure 1: Class-conditional samples generated by our model.

RL-CycleGAN: Reinforcement Learning Aware Simulation-To-Real

Kanishka Rao[1], Chris Harris[1], Alex Irpan[1], Sergey Levine[1, 2], Julian Ibarz[1], and Mohi Khansari[3]

[1]Google Brain, Mountain View, USA
[2]University of California Berkeley, Berkeley, USA
[3]X, The Moonshot Factory, Mountain View, USA
{kanishkarao, ckharris, alexirpan, slevine, julianibarz}@google.com, khansari@x.team

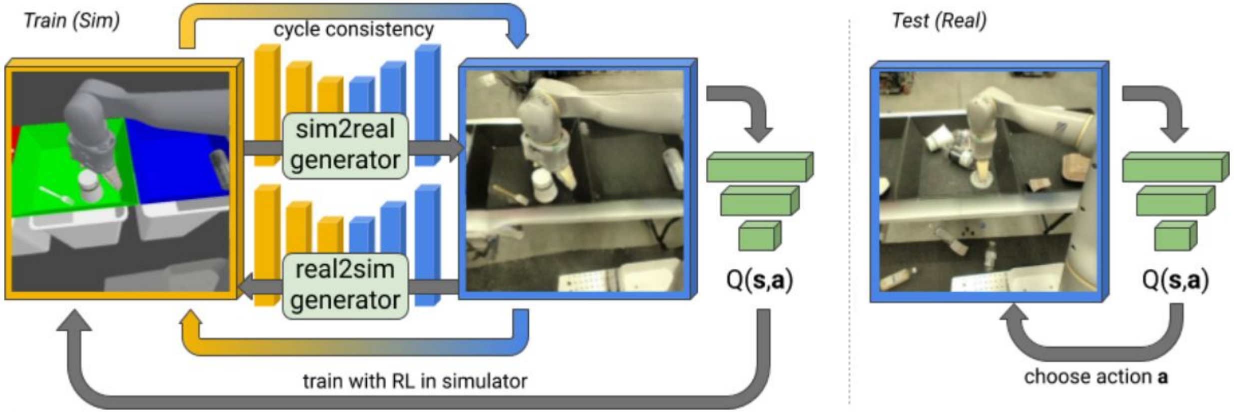


Figure 1. RL-CycleGAN trains a CycleGAN which maps an image from the simulator (left) to a realistic image (middle), a jointly trained RL task ensures that these images are useful for that specific task. At test time, the RL model may be transferred to real robot (right).

[redacted]

Robotics Starter Project

- Crash course on
  - Reinforcement learning
  - Simulation
  - Robot manipulation
  - Image generation
- Worked with leading experts in the field
  - Robotics PhDs
  - Visiting researchers from Universities
  - Peers at Google and Deepmind
- Presented at Robotics conferences

Machine learning breakthroughs → Robotics breakthroughs

Deep RL at Scale: Sorting Waste in Office Buildings with a Fleet of Mobile Manipulators

Alexander Herzog[*†], Kanishka Rao[*‡], Karol Hausman[*‡], Yao Lu[*‡], Paul Wohlhart[*†],
Mengyuan Yan[†], Jessica Lin[†], Montserrat Gonzalez Arenas[‡], Ted Xiao[‡], Daniel Kappler[†], Daniel Ho[†],
Jarek Rettinghouse[†] Yevgen Chebotar[‡], Kuang-Huei Lee[‡], Keerthana Gopalakrishnan[‡], Ryan Julian[‡], Adrian Li[†],
Chuyuan Kelly Fu[†], Bob Wei[†], Sangeetha Ramesh[†], Khem Holden[‡], Kim Kleiven[†], David Rendleman[‡],
Sean Kirmani[†], Jeff Bingham[†], Jon Weisz[†], Ying Xu[†], Wenlong Lu[†], Matthew Bennice[†], Cody Fong[†],
David Do[†], Jessica Lam[†], Yunfei Bai[†], Benjie Holson[†], Michael Quinlan[†], Noah Brown[‡],
Mrinal Kalakrishnan[†], Julian Ibarz[‡], Peter Pastor[†], Sergey Levine[‡]
[*]Authors with equal contribution [†]Everyday Robots [‡]Robotics at Google

Google

Deep RL at Scale: Sorting Waste in Office Buildings with a Fleet of Mobile Manipulators

Alexander Herzog[*†], Kanishka Rao[*‡], Karol Hausman[*‡], Yao Lu[*‡], Paul Wohlhart[*†],
Mengyuan Yan[†], Jessica Lin[†], Montserrat Gonzalez Arenas[‡], Ted Xiao[‡], Daniel Kappler[†], Daniel Ho[†],
Jarek Rettinghouse[†] Yevgen Chebotar[‡], Kuang-Huei Lee[‡], Keerthana Gopalakrishnan[‡], Ryan Julian[‡], Adrian Li[†],
Chuyuan Kelly Fu[†], Bob Wei[†], Sangeetha Ramesh[†], Khem Holden[‡], Kim Kleiven[†], David Rendleman[‡],
Sean Kirmani[†], Jeff Bingham[†], Jon Weisz[†], Ying Xu[†], Wenlong Lu[†], Matthew Bennice[†], Cody Fong[†],
David Do[†], Jessica Lam[†], Yunfei Bai[†], Benjie Holson[†], Michael Quinlan[†], Noah Brown[‡],
Mrinal Kalakrishnan[†], Julian Ibarz[‡], Peter Pastor[†], Sergey Levine[‡]
[*]Authors with equal contribution [†]Everyday Robots [‡]Robotics at Google

Robotics Data

Deep learning requires large amounts of data → Robot data is very expensive to collect

Trash sorting requires learning many concepts:

- What objects look like trash?
- Are they in the wrong bin?
- How would you pick them up?
- Try picking them up, what happens if you fail?
- Place them in the correct bin.
- Did you succeed?

Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

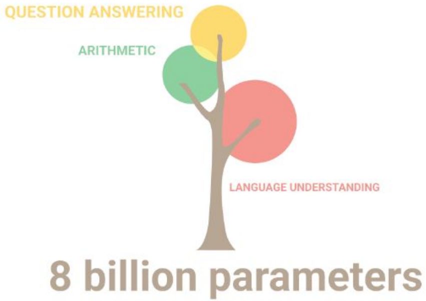
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Large language models have general world understanding

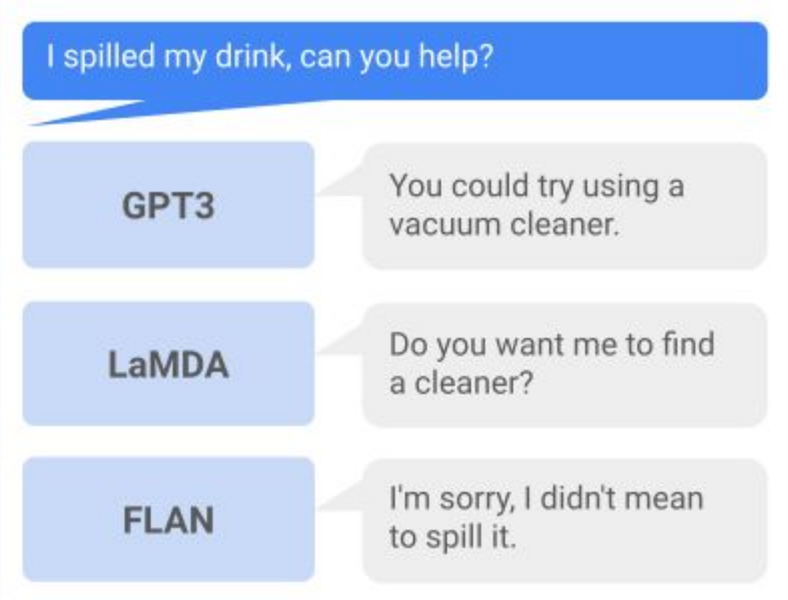
Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

April 4, 2022 · Posted by Sharan Narang and Aakanksha Chowdhery, Software Engineers, Google Research



Transformer based large language models were showing emergent properties

Robotics and Language Models



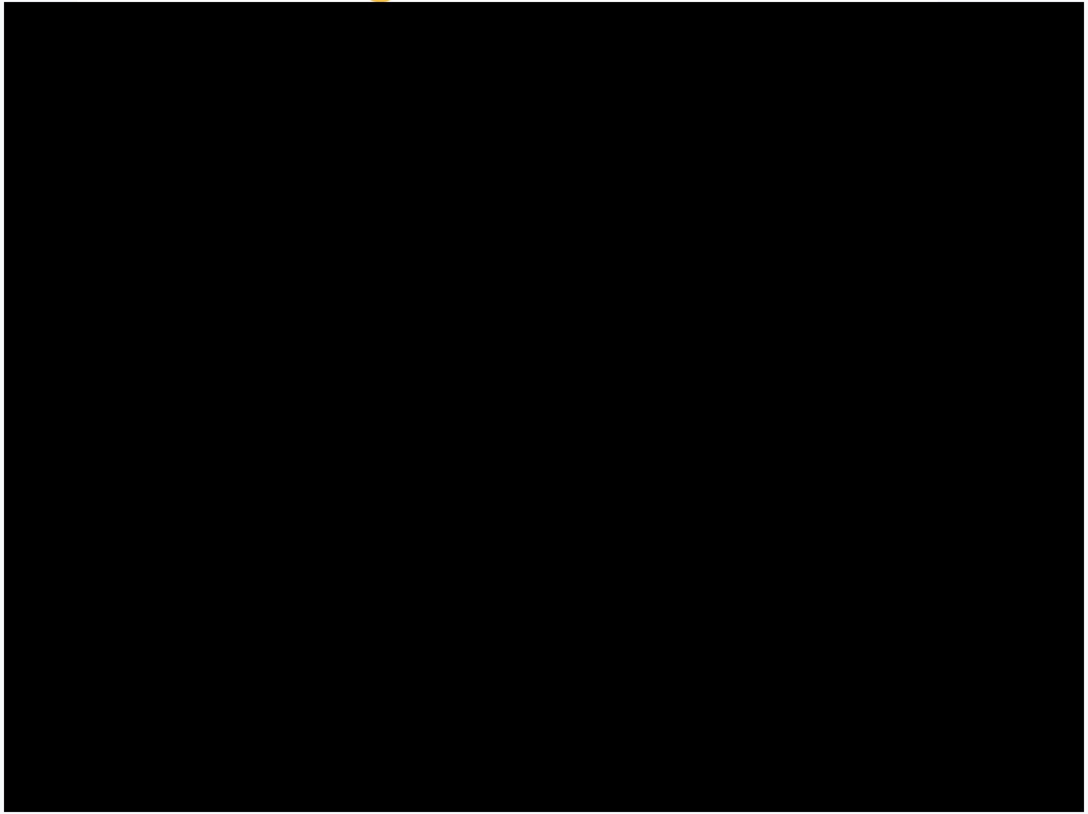
Robotics and Language Models

Do As I Can, Not As I Say:

Grounding Language in Robotic Affordances

Michael Ahn* Anthony Brohan* Noah Brown* Yevgen Chebotar* Omar Cortes* Byron David* Chelsea Finn*
Chuyuan Fu* Keerthana Gopalakrishnan* Karol Hausman* Alex Herzog* Daniel Ho* Jasmine Hsu* Julian Ibarz*
Brian Ichter* Alex Irpan* Eric Jang* Rosario Jauregui Ruano* Kyle Jeffrey* Sally Jesmonth* Nikhil Joshi*
Ryan Julian* Dmitry Kalashnikov* Yuheng Kuang* Kuang-Huei Lee* Sergey Levine* Yao Lu* Linda Luu* Carolina Parada*
Peter Pastor* Jornell Quiambao* Kanishka Rao* Jarek Rettinghouse* Diego Reyes* Pierre Sermanet* Nicolas Sievers*
Clayton Tan* Alexander Toshev* Vincent Vanhoucke* Fei Xia* Ted Xiao* Peng Xu* Sichun Xu* Mengyuan Yan* Andy Zeng*

Robotics at Google Everyday Robots



Robotics and Large Language Models

- Made the robot more useful
  - Can do more tasks
  - You can talk to it
- Grounded the language models in the real world
  - Embodied LLM
- LLMs solved the task planning problem for robotics

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

Anthony Brohan Noah Brown Justice Carbajal Yevgen Chebotar Joseph Dabis Chelsea Finn Keerthana Gopalakrishnan
Karol Hausman Alex Herzog Jasmine Hsu Julian Ibarz Brian Ichter Alex Irpan Tomas Jackson
Sally Jesmonth Nikhil Joshi Ryan Julian Dmitry Kalashnikov Yuheng Kuang Isabel Leal Kuang-Huei Lee
Sergey Levine Yao Lu Utsav Malla Deeksha Manjunath Igor Mordatch Ofir Nachum Carolina Parada
Jodilyn Peralta Emily Perez Karl Pertsch Jornell Quiambao Kanishka Rao Michael Ryoo Grecia Salazar
Pannag Sanketi Kevin Sayed Jaspiar Singh Sumedh Sontakke Austin Stone Clayton Tan Huong Tran
Vincent Vanhoucke Steve Vega Quan Vuong Fei Xia Ted Xiao Peng Xu Sichun Xu Tianhe Yu Brianna Zitkovich

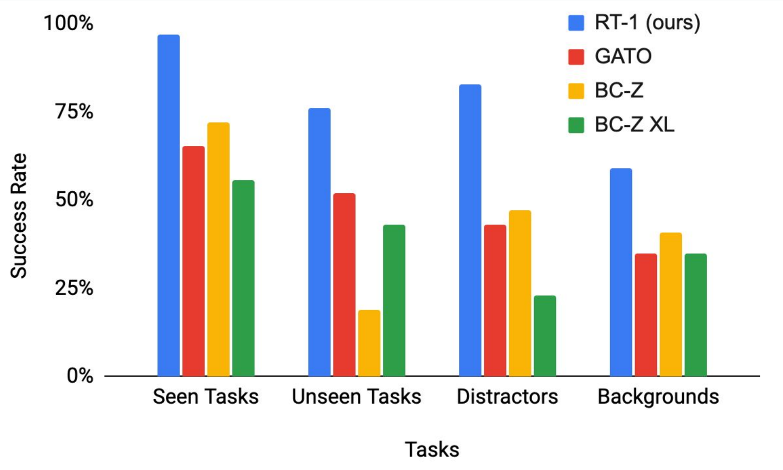
Authors listed in alphabetical order (see paper appendix for contribution statement).

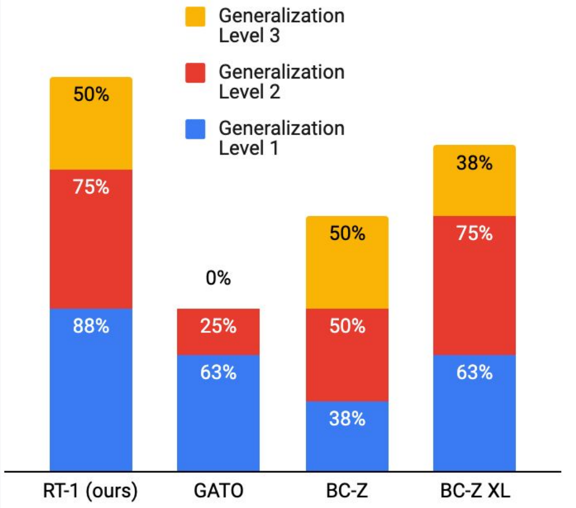
Google

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

Anthony Brohan Noah Brown Justice Carbajal Yevgen Chebotar Joseph Dabis Chelsea Finn Keerthana Gopalakrishnan
Karol Hausman Alex Herzog Jasmine Hsu Julian Ibarz Brian Ichter Alex Irpan Tomas Jackson
Sally Jesmonth Nikhil Joshi Ryan Julian Dmitry Kalashnikov Yuheng Kuang Isabel Leal Kuang-Huei Lee
Sergey Levine Yao Lu Utsav Malla Deeksha Manjunath Igor Mordatch Ofir Nachum Carolina Parada
Jodilyn Peralta Emily Perez Karl Pertsch Jornell Quiambao Kanishka Rao Michael Ryoo Grecia Salazar
Pannag Sanketi Kevin Sayed Jaspiar Singh Sumedh Sontakke Austin Stone Clayton Tan Huong Tran
Vincent Vanhoucke Steve Vega Quan Vuong Fei Xia Ted Xiao Peng Xu Sichun Xu Tianhe Yu Brianna Zitkovich

Authors listed in alphabetical order (see paper appendix for contribution statement).





Vision Transformers

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

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

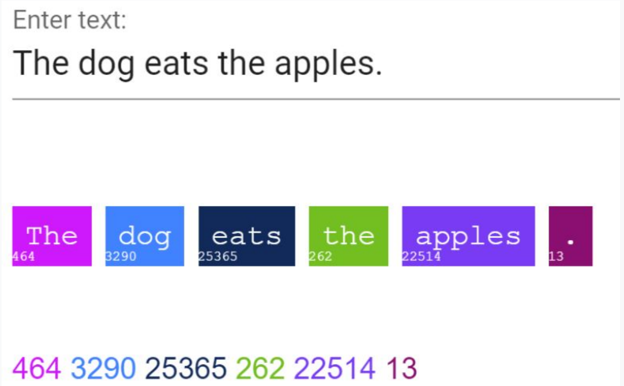
Vision + Language Transformers

CoCa: Contrastive Captioners are Image-Text Foundation Models

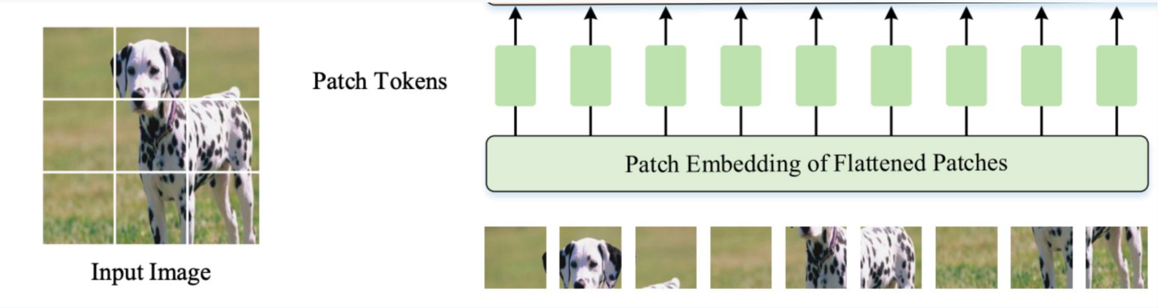
Jiahui Yu[†] **Zirui Wang[†]**
{jiahuiyu, ziruiw}@google.com
Vijay Vasudevan Legg Yeung Mojtaba Seyedhosseini Yonghui Wu
Google Research

Tokens are all you need

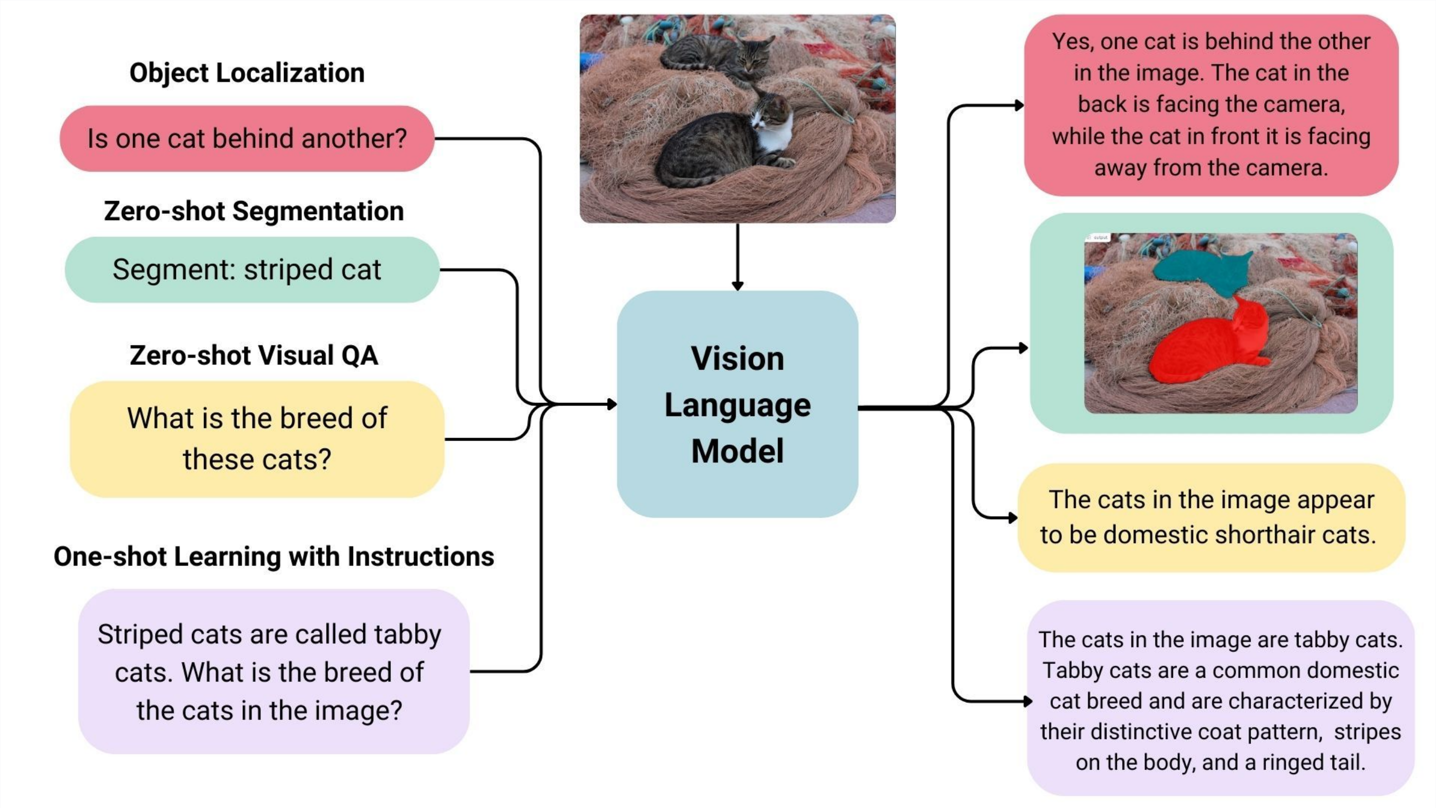
Language



Vision



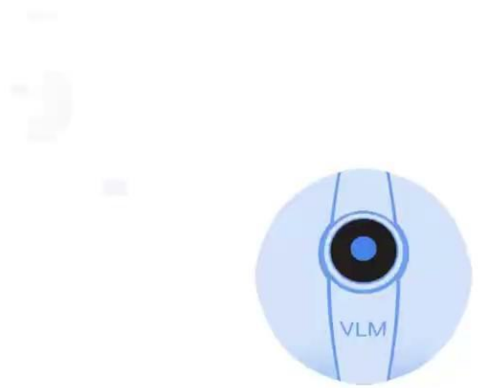
VLM


Object Localization
Is one cat behind another?
Zero-shot Segmentation
Segment: striped cat
Zero-shot Visual QA
What is the breed of these cats?
One-shot Learning with Instructions
Striped cats are called tabby cats. What is the breed of the cats in the image?
Vision Language Model
Yes, one cat is behind the other in the image. The cat in the back is facing the camera, while the cat in front it is facing away from the camera.
The cats in the image appear to be domestic shorthair cats.
The cats in the image are tabby cats. Tabby cats are a common domestic cat breed and are characterized by their distinctive coat pattern, stripes on the body, and a ringed tail.

RT-2: Vision-Language-Action Models

Transfer Web Knowledge to Robotic Control

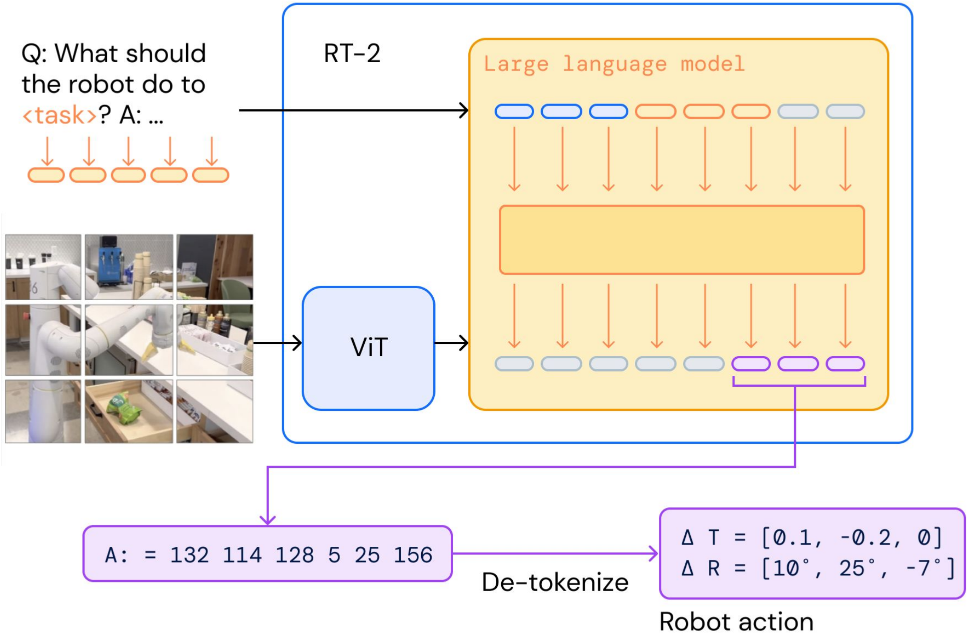
Anthony Brohan Noah Brown Justice Carbajal Yevgen Chebotar Xi Chen Krzysztof Choromanski Tianli Ding
Danny Driess Avinava Dubey Chelsea Finn Pete Florence Chuyuan Fu Montse Gonzalez Arenas Keerthana Gopalakrishnan
Kehang Han Karol Hausman Alex Herzog Jasmine Hsu Brian Ichter Alex Irpan Nikhil Joshi Ryan Julian
Dmitry Kalashnikov Yuheng Kuang Isabel Leal Lisa Lee Tsang-Wei Edward Lee Sergey Levine Yao Lu Henryk Michalewski
Igor Mordatch Karl Pertsch Kanishka Rao Krista Reymann Michael Ryoo Grecia Salazar Pannag Sanketi Pierre Sermanet
Jaspiar Singh Anikait Singh Radu Soricut Huong Tran Vincent Vanhoucke Quan Vuong Ayzaan Wahid Stefan Welker
Paul Wohlhart Jialin Wu Fei Xia Ted Xiao Peng Xu Sichun Xu Tianhe Yu Brianna Zitkovich



RT-2: Vision-Language-Action Models

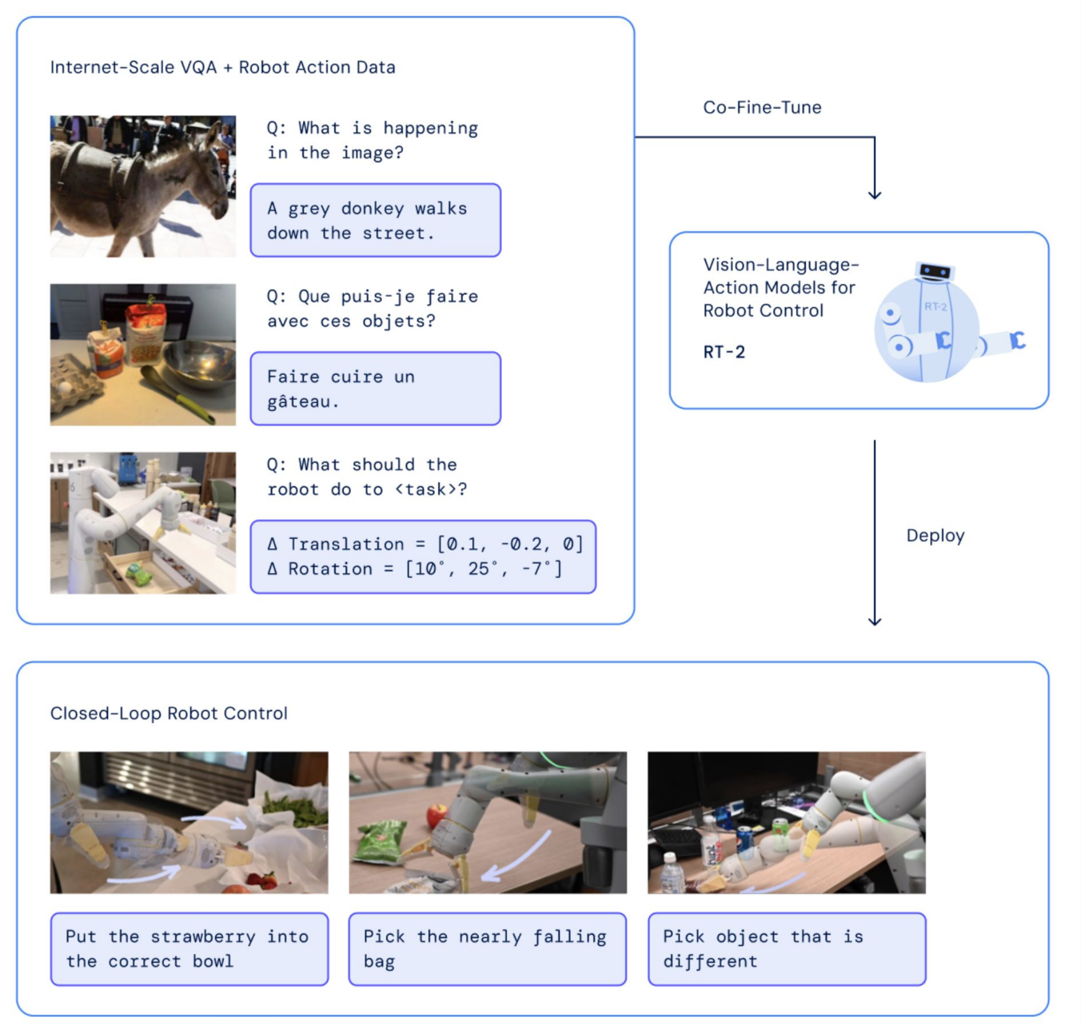
Transfer Web Knowledge to Robotic Control

Anthony Brohan Noah Brown Justice Carbajal Yevgen Chebotar Xi Chen Krzysztof Choromanski Tianli Ding
Danny Driess Avinava Dubey Chelsea Finn Pete Florence Chuyuan Fu Montse Gonzalez Arenas Keerthana Gopalakrishnan
Kehang Han Karol Hausman Alex Herzog Jasmine Hsu Brian Ichter Alex Irpan Nikhil Joshi Ryan Julian
Dmitry Kalashnikov Yuheng Kuang Isabel Leal Lisa Lee Tsang-Wei Edward Lee Sergey Levine Yao Lu Henryk Michalewski
Igor Mordatch Karl Pertsch Kanishka Rao Krista Reymann Michael Ryoo Grecia Salazar Pannag Sanketi Pierre Sermanet
Jaspiar Singh Anikait Singh Radu Soricut Huong Tran Vincent Vanhoucke Quan Vuong Ayzaan Wahid Stefan Welker
Paul Wohlhart Jialin Wu Fei Xia Ted Xiao Peng Xu Sichun Xu Tianhe Yu Brianna Zitkovich

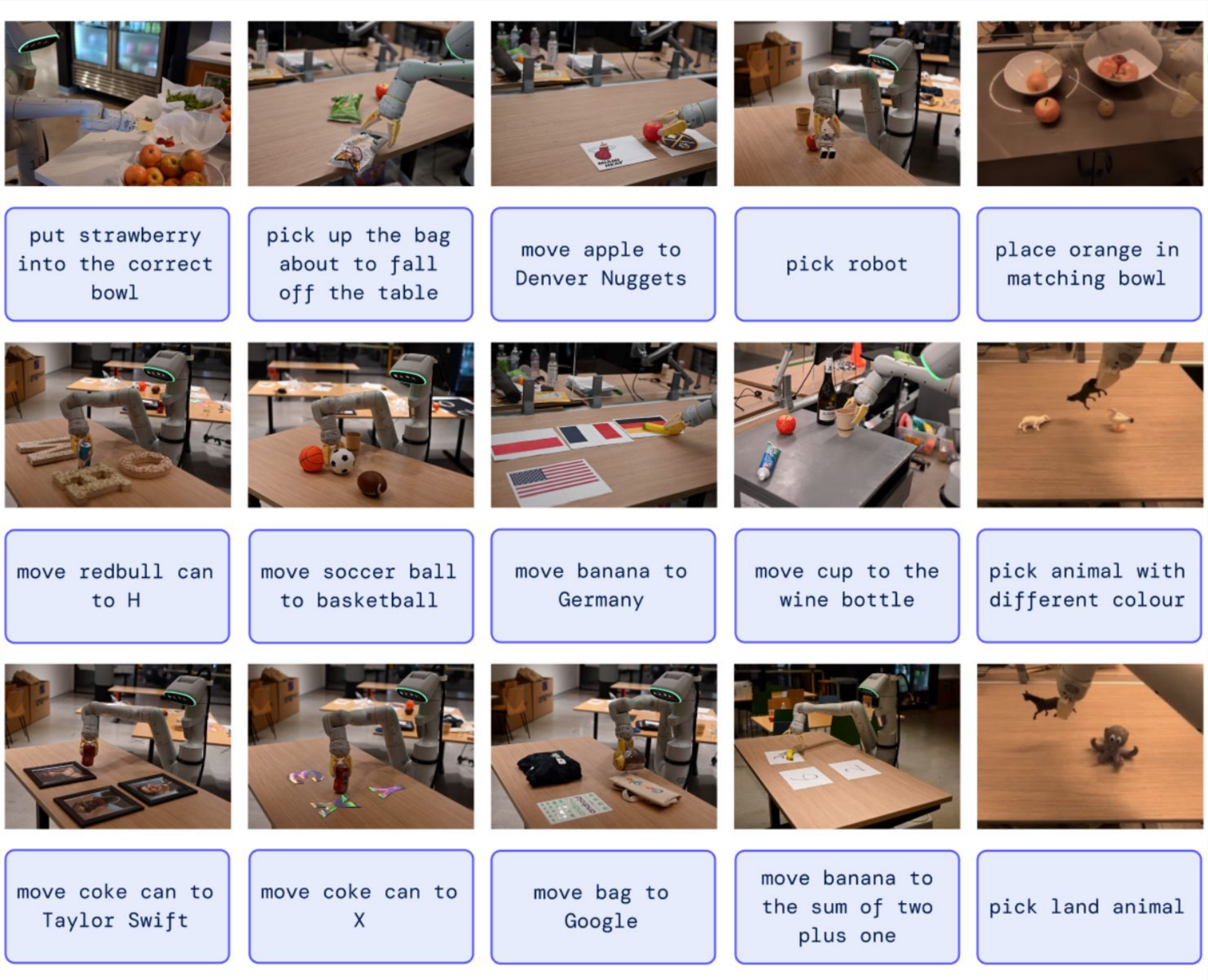


Google

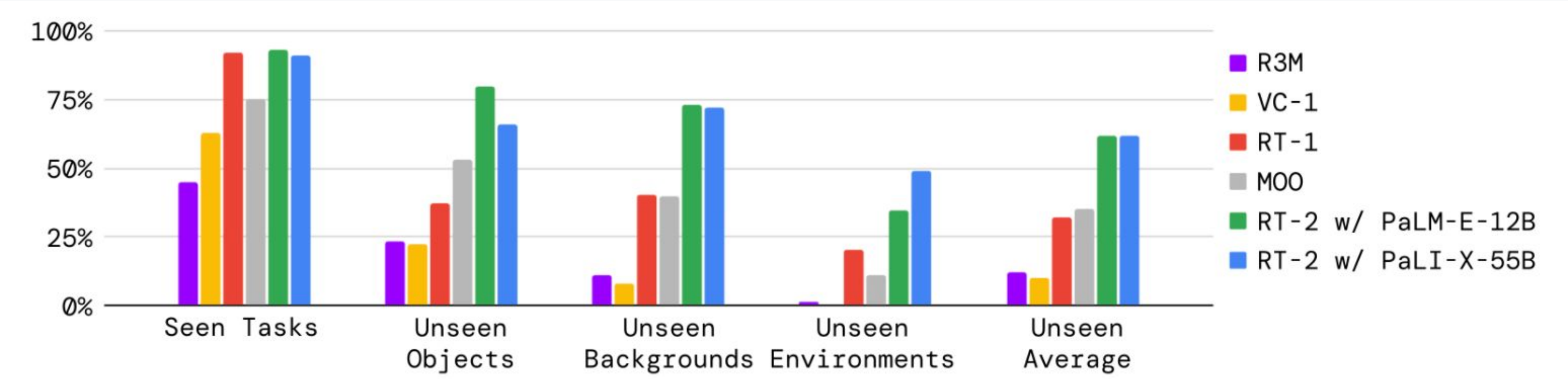
RT-2: VLA


Internet-Scale VQA + Robot Action Data
Q: What is happening in the image?
A grey donkey walks down the street.
Q: Que puis-je faire avec ces objets?
Faire cuire un gâteau.
Q: What should the robot do to <task>?
Δ Translation = [0.1, -0.2, 0]
Δ Rotation = [10°, 25°, -7°]
Co-Fine-Tune
Vision-Language-Action Models for Robot Control
RT-2
Deploy
Closed-Loop Robot Control
Put the strawberry into the correct bowl
Pick the nearly falling bag
Pick object that is different

RT-2: VLA


put strawberry into the correct bowl
pick up the bag about to fall off the table
move apple to Denver Nuggets
pick robot
place orange in matching bowl
move redbull can to H
move soccer ball to basketball
move banana to Germany
move cup to the wine bottle
pick animal with different colour
move coke can to Taylor Swift
move coke can to X
move bag to Google
move banana to the sum of two plus one
pick land animal

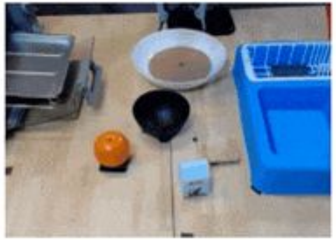
RT-2: VLA



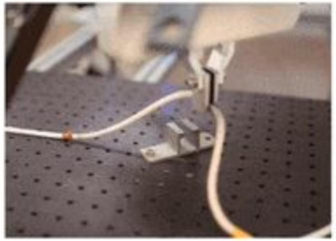
VLA were showing robotics emergent properties

Open X-Embodiment: Robotic Learning Datasets and RT-X Models

Open X-Embodiment Collaboration
(hover to display full author list)



CLVR, USC



RAIL, UC Berkeley

CILVR, NYU

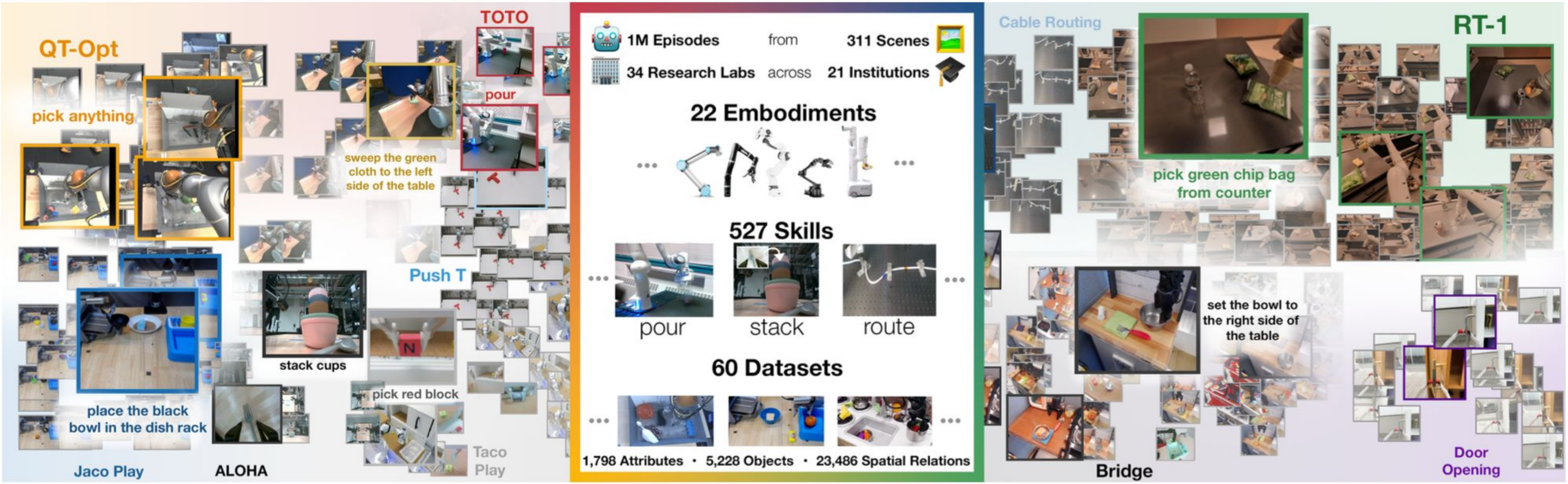
AUTOLab, UC Berkeley

AiS, University of Freiburg

Open X-Embodiment: Robotic Learning Datasets and RT-X Models

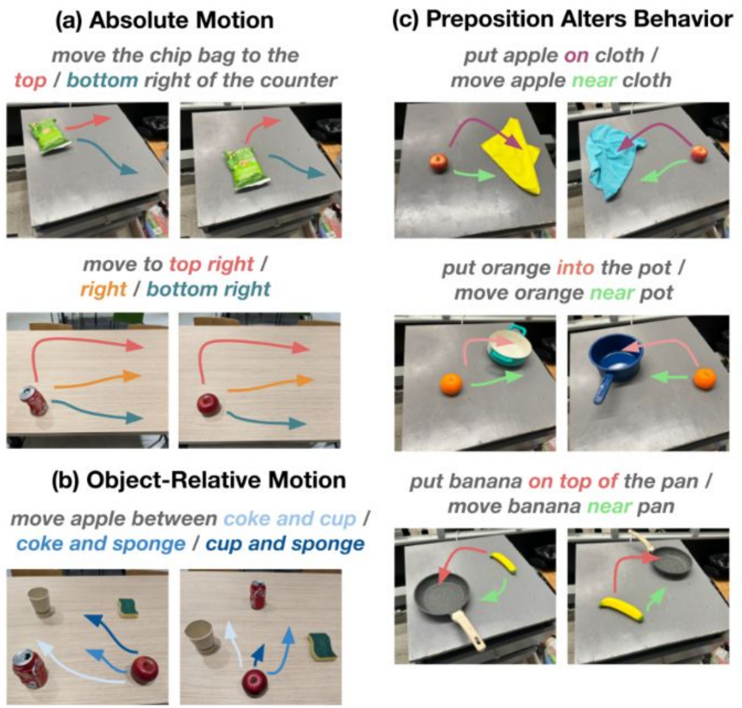
Open X-Embodiment Collaboration

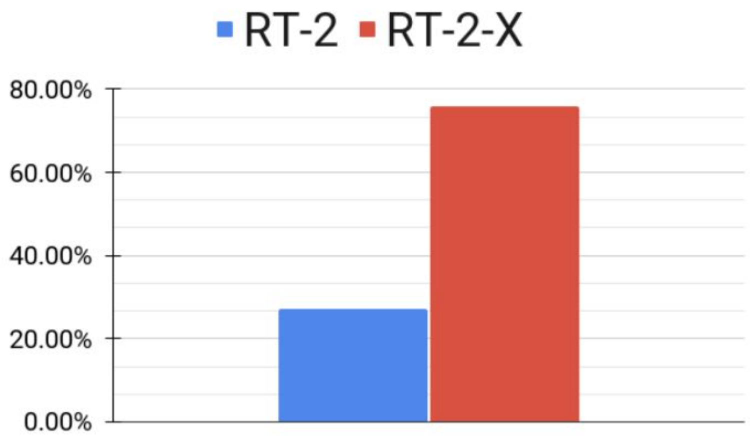
(hover to display full author list)



Open X-Embodiment: Robotic Learning Datasets and RT-X Models

Open X-Embodiment Collaboration
(hover to display full author list)

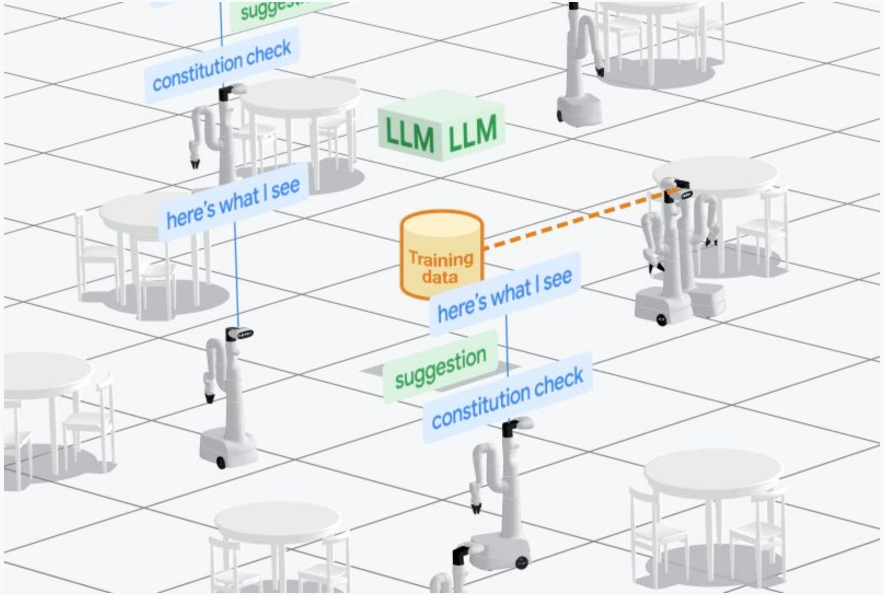


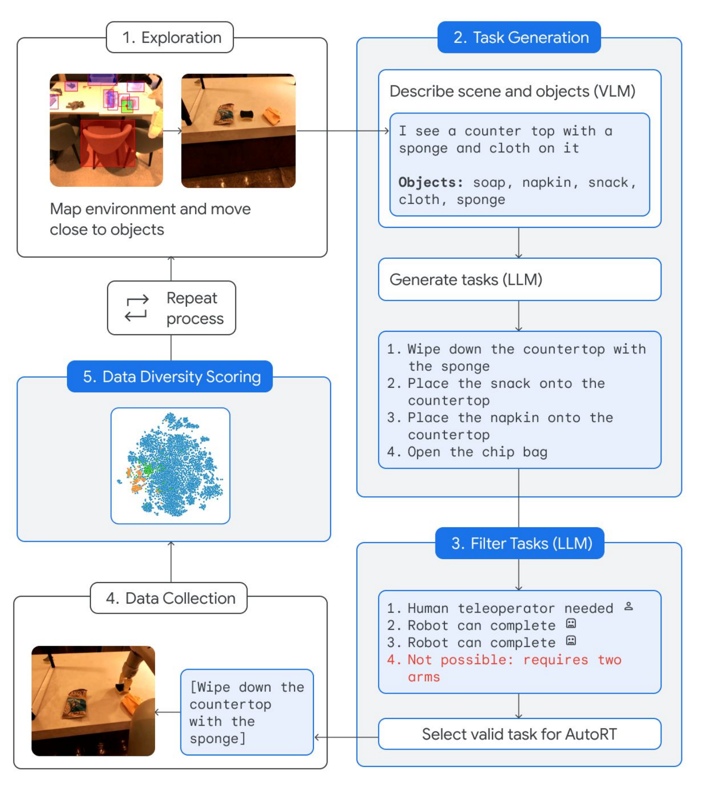


RT-2-X outperforms RT-2 by 3x in emergent skill evaluation

AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents

Michael Ahn[1], Debidatta Dwibedi[1], Chelsea Finn[1], Montse Gonzalez Arenas[1], Keerthana Gopalakrishnan[1], Karol Hausman[1], Brian Ichter[1], Alex Irpan[1], Nikhil Joshi[1], Ryan Julian[1], Sean Kirmani[1], Isabel Leal[1], Edward Lee[1], Sergey Levine[1], Yao Lu[1], Sharath Maddineni[1], Kanishka Rao[1], Dorsa Sadigh[1], Pannag Sanketi[1], Pierre Sermanet[1], Quan Vuong[1], Stefan Welker[1], Fei Xia[1], Ted Xiao[1], Peng Xu[1], Steve Xu[1], Zhuo Xu[1]
[1]Google DeepMind






Robotics @ Google
105
853

AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents

Michael Ahn[1], Debidatta Dwibedi[1], Chelsea Finn[1], Montse Gonzalez Arenas[1], Keerthana Gopalakrishnan[1], Karol Hausman[1], Brian Ichter[1], Alex Irpan[1], Nikhil Joshi[1], Ryan Julian[1], Sean Kirmani[1], Isabel Leal[1], Edward Lee[1], Sergey Levine[1], Yao Lu[1], Sharath Maddineni[1], Kanishka Rao[1], Dorsa Sadigh[1], Pannag Sanketi[1], Pierre Sermanet[1], Quan Vuong[1], Stefan Welker[1], Fei Xia[1], Ted Xiao[1], Peng Xu[1], Steve Xu[1], Zhuo Xu[1]
[1]Google DeepMind

Robot Constitution prompting

FOUNDATIONAL_RULES =
F1. A robot may not injure a human being.
F2. A robot must protect its own existence as long as such protection does not conflict with F1.
F3. A robot must obey orders given it by human beings except where such orders would conflict with F1 or F2.

SAFETY_RULES =
S1. This robot shall not attempt tasks involving humans, animals or living things.
S2. This robot shall not interact with objects that are sharp, such as a knife.
S3. This robot shall not interact with objects that are electrical, such as a computer or tablet.

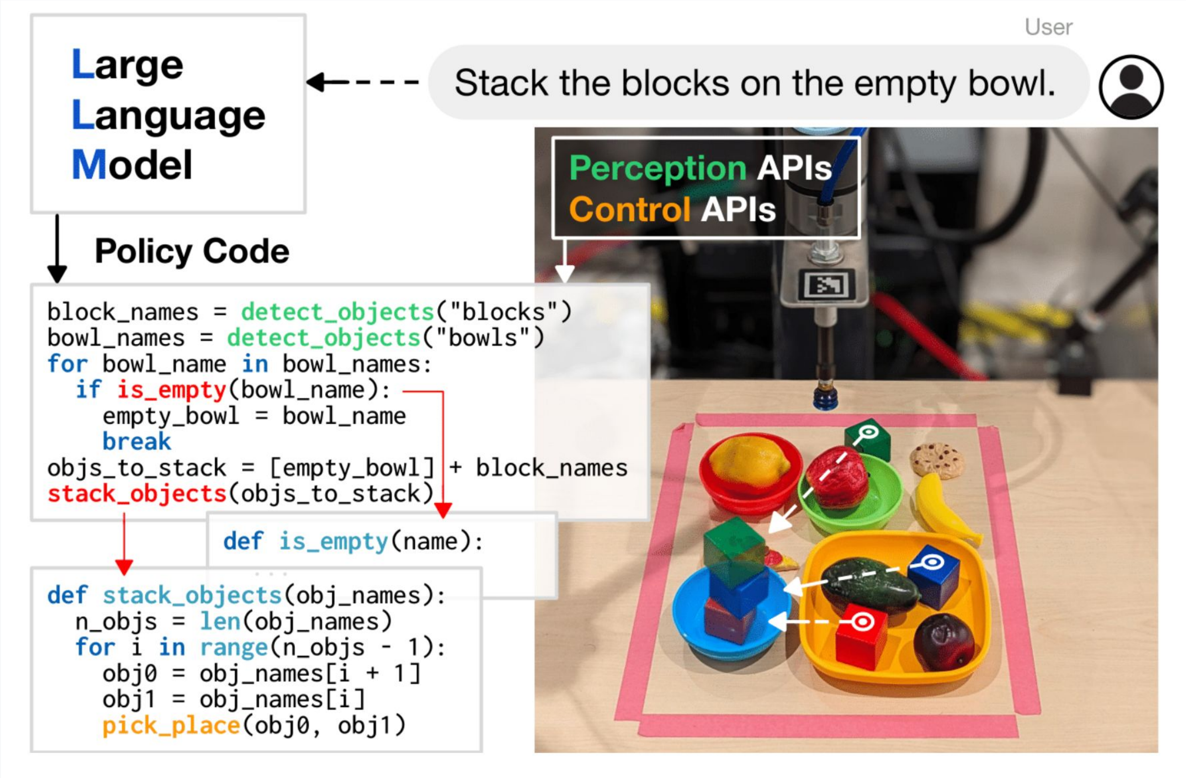
EMBODIMENT_RULES =
E1. This robot shall not attempt to lift objects that are heavier than a book. For example, it cannot move a couch but it can push plastic chairs.
E2. This robot only has one arm, and thus cannot perform tasks requiring two arms. For example, it cannot open a bottle.

GUIDANCE_RULES =
G1. The human command, which the robot should follow if given: {guidance}

Code as Policies:
Language Model Programs for Embodied Control

Jacky Liang   Wenlong Huang   Fei Xia   Peng Xu   Karol Hausman   Brian Ichter   Pete Florence   Andy Zeng

Robotics at Google



In conclusion...

Physicists in ML

Machine Learning research in industry can be very exciting and rewarding.

- Comfortable with being a noob
- Enjoys building useful things more than doing science
- How valuable is the work?
- Thrives in chaos, fast-moving, messy collaborations
- Enjoys general problem solving

Thank You