Adventures in Multimodal Machine Learning

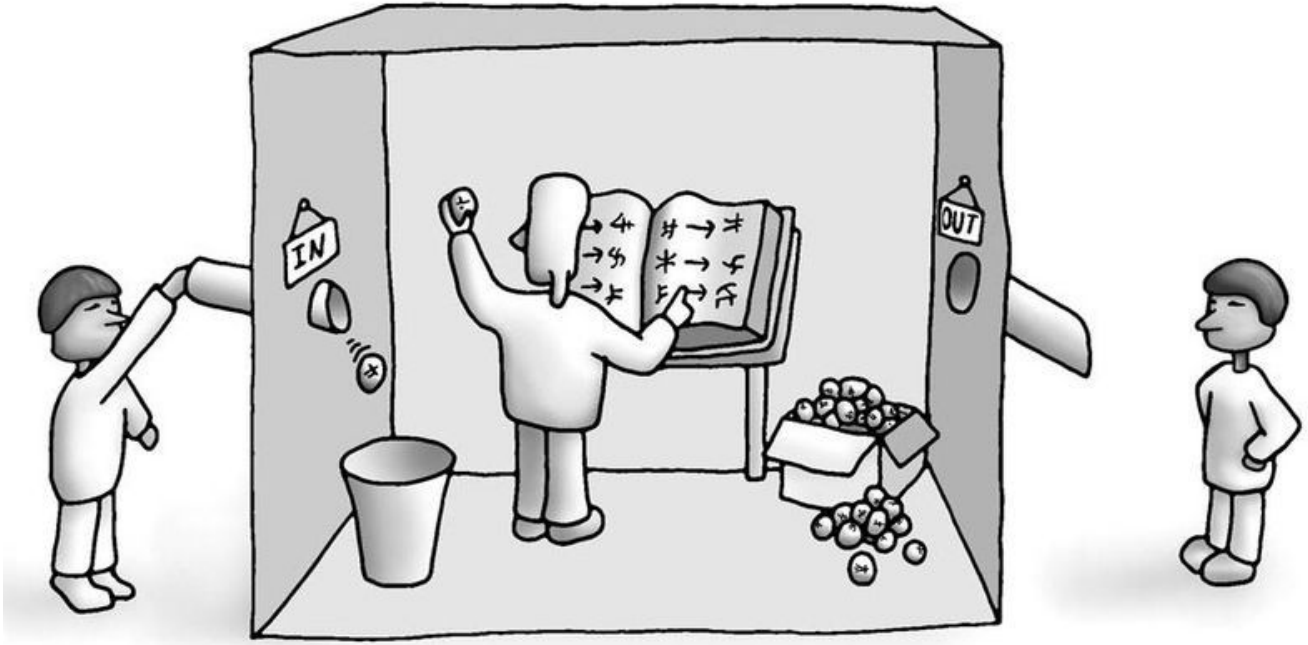
Grounding, Meaning and Foundation Models

Douwe Kiela

Hugging Face

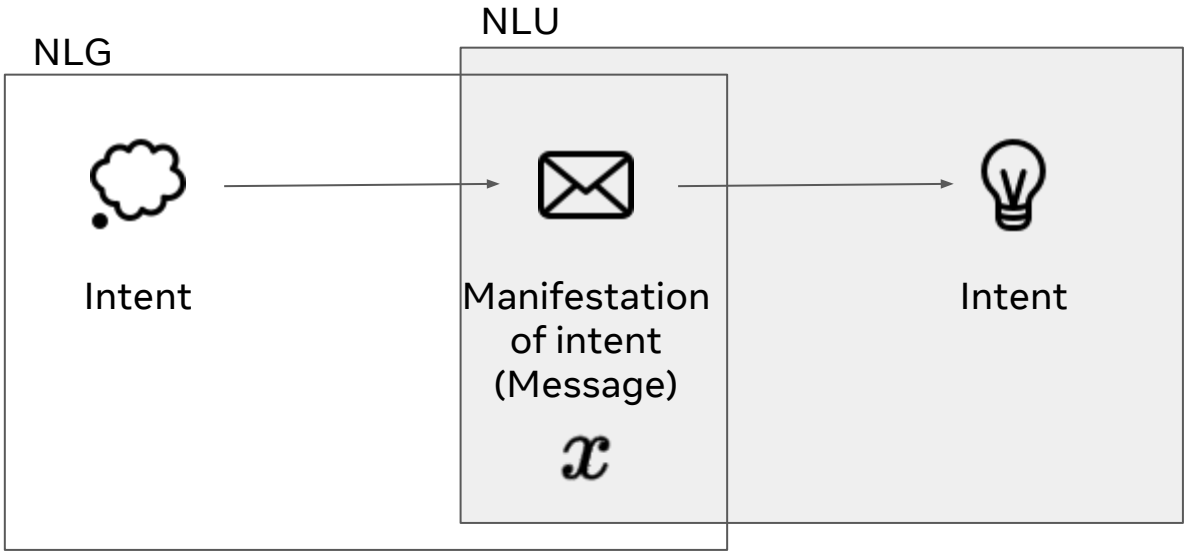
Stanford University

How do we get (true, human-like) meaning into machines?



Source: WikiCommons

What is meaning?


NLG
NLU
Intent
Manifestation of intent (Message)
x
Intent

Thinking about language learning..

(excuse notational shorthand)

$$arg\,max_\theta P(\hat{y} = correct \mid x; \theta)$$

Manifestation of intent (Message)

x

Inferred intent

$\hat{y}$

Additional assumptions: i.i.d. train/test data; MLE is good enough; etc.

$$arg\,max_\theta P_{LM}(\tilde{x} \mid x; \theta)$$

$$arg\,max_\theta P(y = \hat{y} \mid x; \theta)$$

True intent

y

Manifestation of intent (Message)

x

Inferred intent

$\hat{y}$

$$argmax_{(\theta_S,\theta_L)} P(g_{\theta_L}(f_{\theta_S}(y)) = y \mid \theta_S, \theta_L)$$

True intent

y

Speaker

$f : y \rightarrow m$

Manifestation of intent (Message)

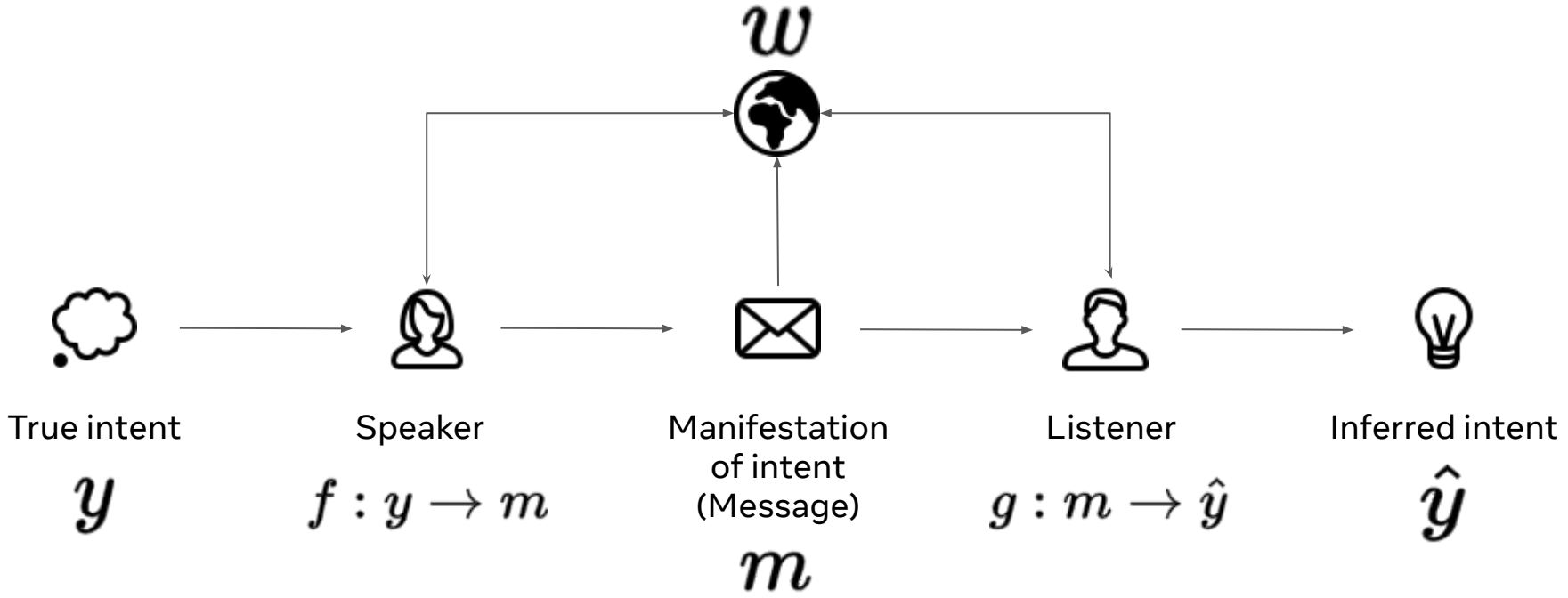
m

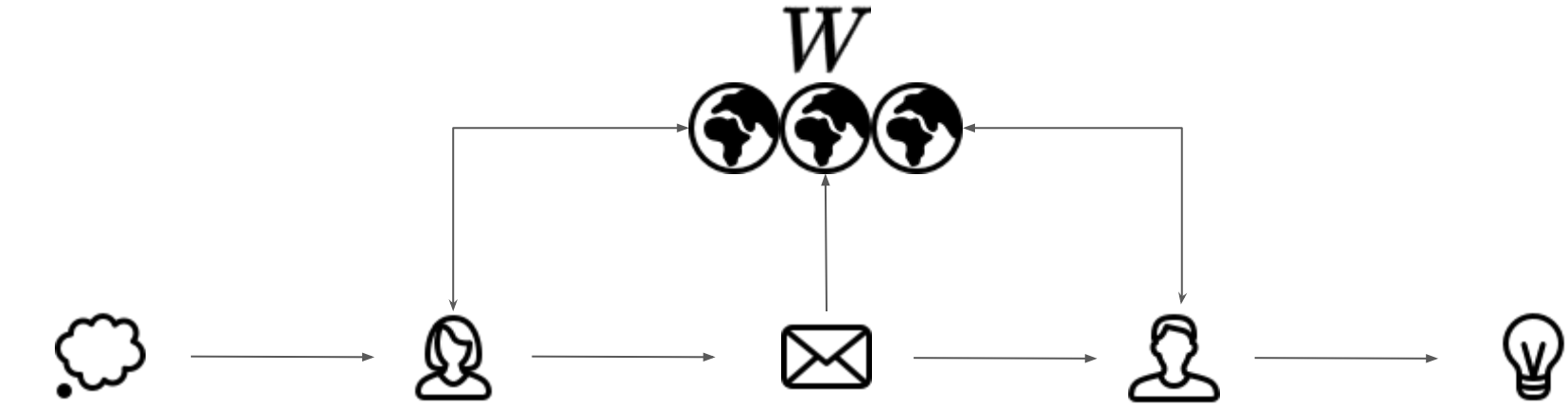
Listener

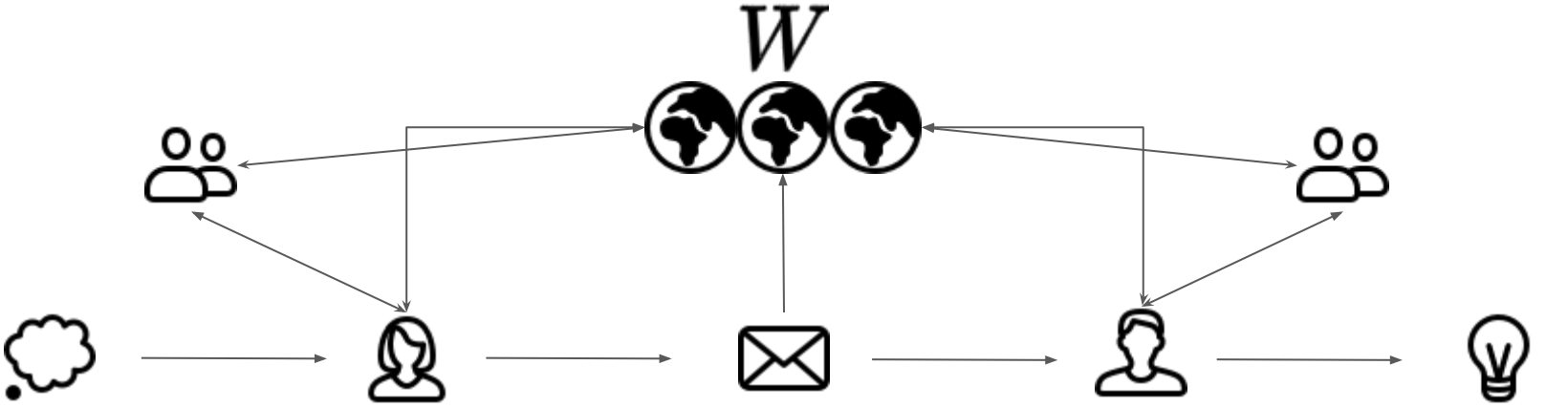
$g : m \rightarrow \hat{y}$

Inferred intent

$\hat{y}$


w
True intent
y
Speaker
$f : y \rightarrow m$
Manifestation of intent (Message)
m
Listener
$g : m \rightarrow \hat{y}$
Inferred intent
$\hat{y}$


W
True intent
y
Speaker
$f : y \rightarrow m$
Manifestation of intent (Message)
m
Listener
$g : m \rightarrow \hat{y}$
Inferred intent
$\hat{y}$



True intent

y

Speaker

$f : y \rightarrow m$

Manifestation of intent (Message)

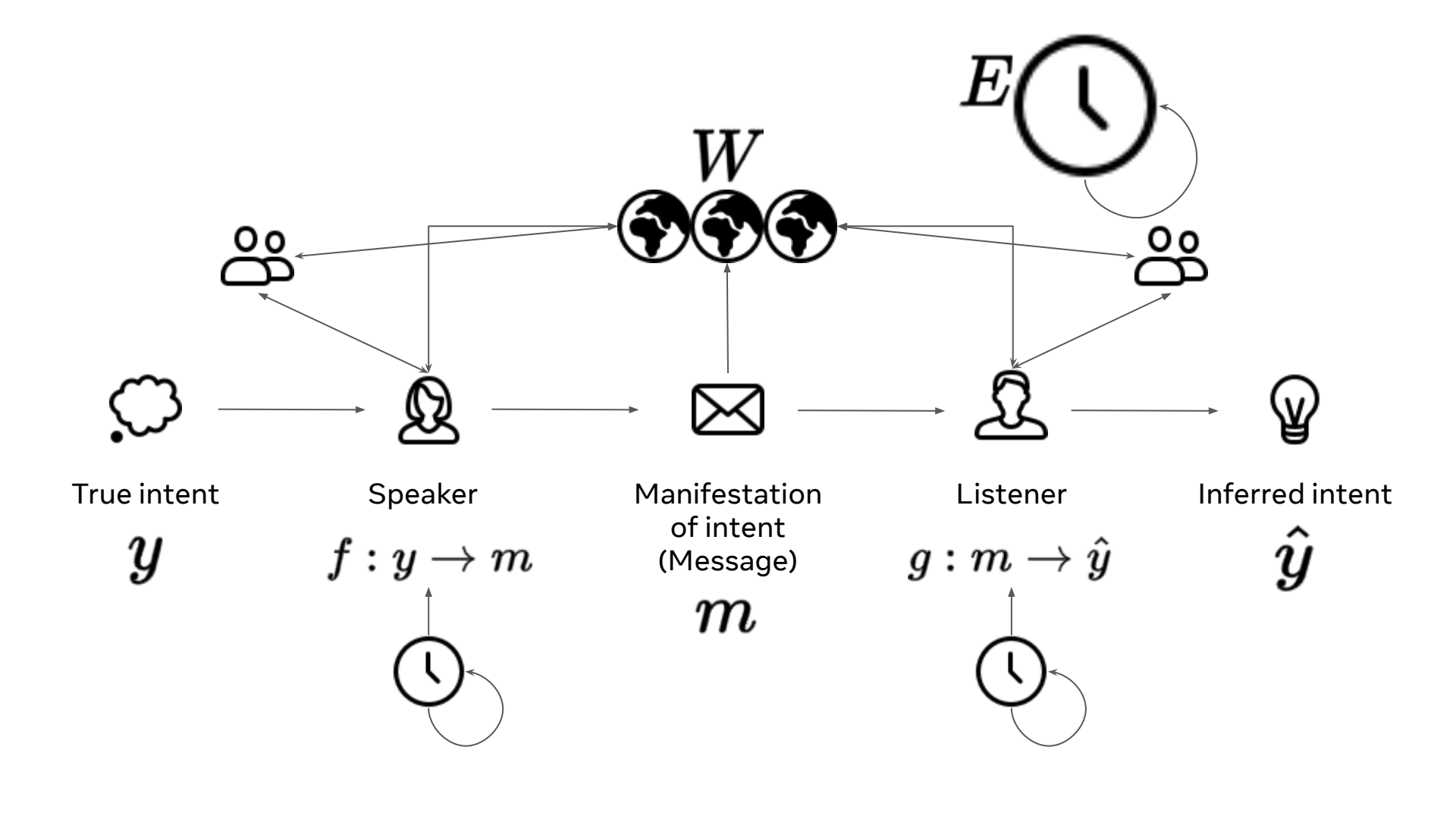
m

Listener

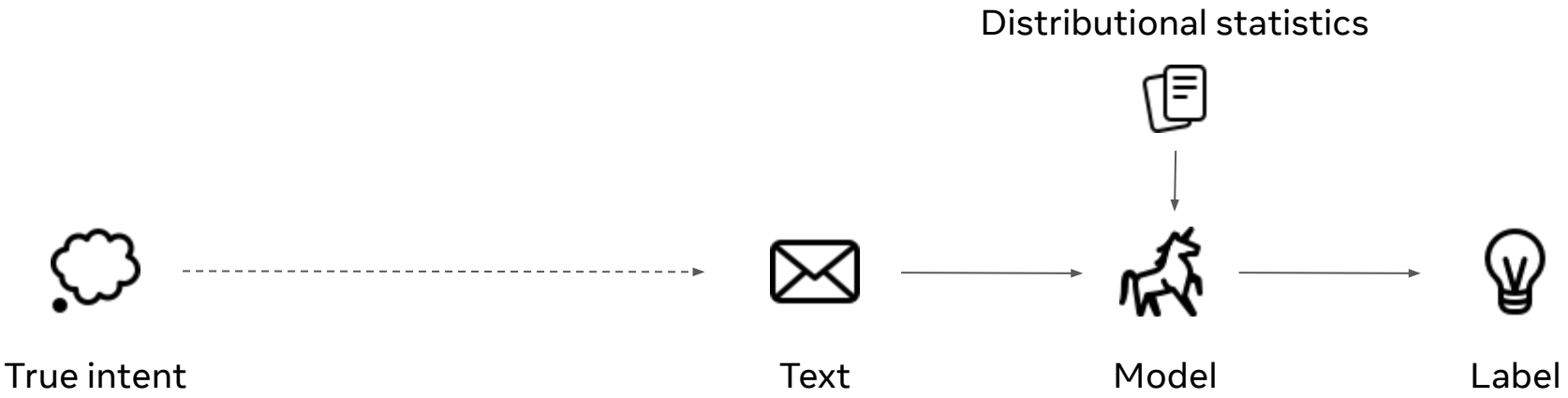
$g : m \rightarrow \hat{y}$

Inferred intent

$\hat{y}$


E
W
True intent
y
Speaker
$f : y \to m$
Manifestation of intent (Message)
m
Listener
$g : m \to \hat{y}$
Inferred intent
$\hat{y}$

What we are doing..



This feels wrong! Can we do better?

This talk: Outline

- **The Past.** Multimodal Semantics & Language Games
- **The Present:** Multimodal Evaluation
- **The Future?** Multimodal Universal “Foundation” Models

Multimodal theory of meaning

Surface form / Syntax
Proof theory

Grounding / Semantics
Embodiment
Model theory

Pragmatics
Multi-agent (emergent) communication
Theory of mind

Language acquisition
Evolution

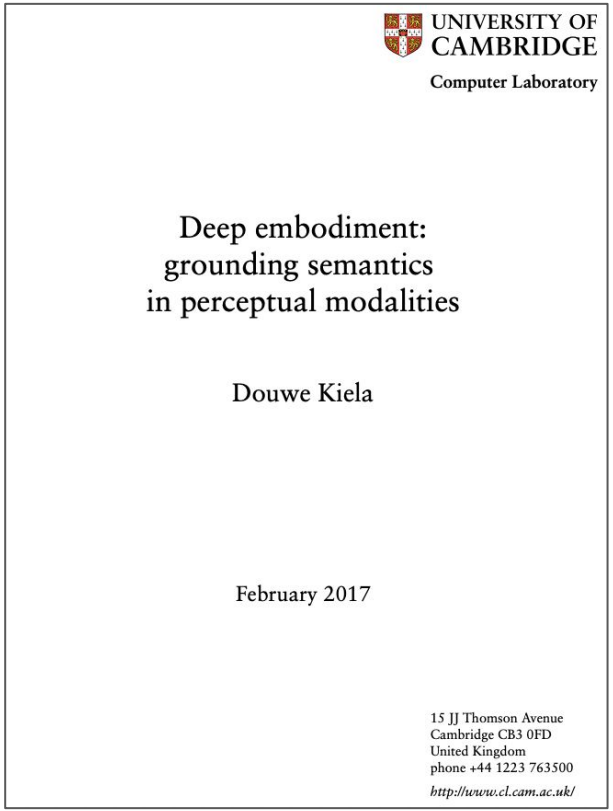
Grounding

Surface form / Syntax
Proof theory

Grounding / Semantics
Embodiment
Model theory

Pragmatics
Multi-agent (emergent) communication
Theory of mind

Language acquisition
Evolution


UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Deep embodiment:
grounding semantics
in perceptual modalities

Douwe Kiela

February 2017

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500
http://www.cl.cam.ac.uk/

Grounding semantics in perceptual modalities

Grounding semantics in perceptual modalities

Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics

Douwe Kiela*
University of Cambridge
Computer Laboratory
douwe.kiela@cl.cam.ac.uk

Léon Bottou
Microsoft Research
New York
leon@bottou.org

Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More

Douwe Kiela*, Felix Hill*, Anna Korhonen and Stephen Clark
University of Cambridge
Computer Laboratory

Exploiting Image Generality for Lexical Entailment Detection

Douwe Kiela
Computer Laboratory
University of Cambridge
douwe.kiela@cl.cam.ac.uk

Laura Rimell
Computer Laboratory
University of Cambridge
laura.rimell@cl.cam.ac.uk

Ivan Vulić
Department of Computer Science
KU Leuven
ivan.vulic@cs.kuleuven.be

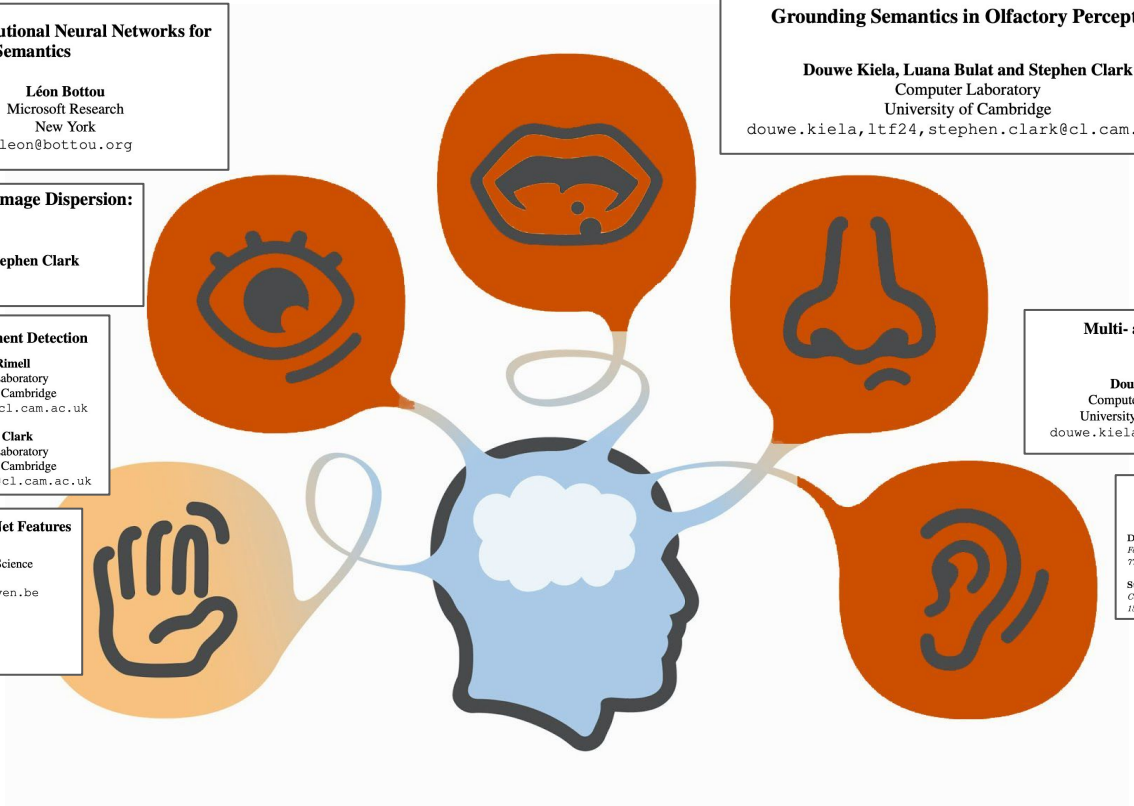
Stephen Clark
Computer Laboratory
University of Cambridge
stephen.clark@cl.cam.ac.uk

Visual Bilingual Lexicon Induction with Transferred ConvNet Features

Douwe Kiela
Computer Laboratory
University of Cambridge
douwe.kiela@cl.cam.ac.uk

Ivan Vulić
Department of Computer Science
KU Leuven
ivan.vulic@cs.kuleuven.be

Stephen Clark
Computer Laboratory
University of Cambridge
stephen.clark@cl.cam.ac.uk



Grounding Semantics in Olfactory Perception

Douwe Kiela, Luana Bulat and Stephen Clark
Computer Laboratory
University of Cambridge
douwe.kiela,ltf24,stephen.clark@cl.cam.ac.uk

Multi- and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception

Douwe Kiela
Computer Laboratory
University of Cambridge
douwe.kiela@cl.cam.ac.uk

Stephen Clark
Computer Laboratory
University of Cambridge
stephen.clark@cl.cam.ac.uk

Learning Neural Audio Embeddings for Grounding Semantics in Auditory Perception

Douwe Kiela DKIELA@FB.COM
Facebook Artificial Intelligence Research
770 Broadway, New York, NY 10003, USA

Stephen Clark STEPHEN.CLARK@CL.CAM.AC.UK
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

ConvNets for Visual Semantics

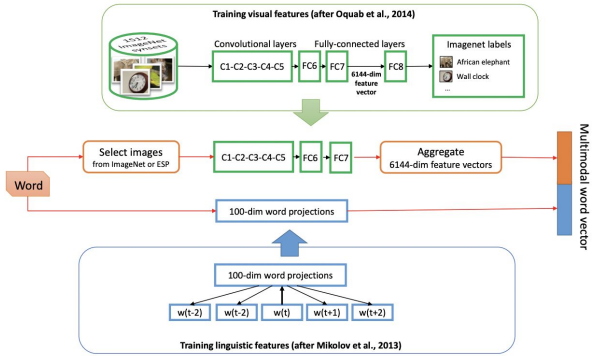
Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics

Douwe Kiela*
University of Cambridge
Computer Laboratory
douwe.kiela@cl.cam.ac.uk

Léon Bottou
Microsoft Research
New York
leon@bottou.org

- Deep learning was becoming really cool.
- Bag-of-visual words was used for multi-modal semantics (Bruni et al)

- Leon Bottou, intern supervisor at MSR in 2014 - here's some cool projects:
  1. Try out LSTMs on NLP problems like language modelling
  2. Build a Python framework for auto-differentiable neural networks
  3. … and some other brilliant idea.
- Me, wanting to work on my own idea:
  * Convolutional networks for multimodal semantics



(Later applied similar ideas to audio,
and similar features to other problems)

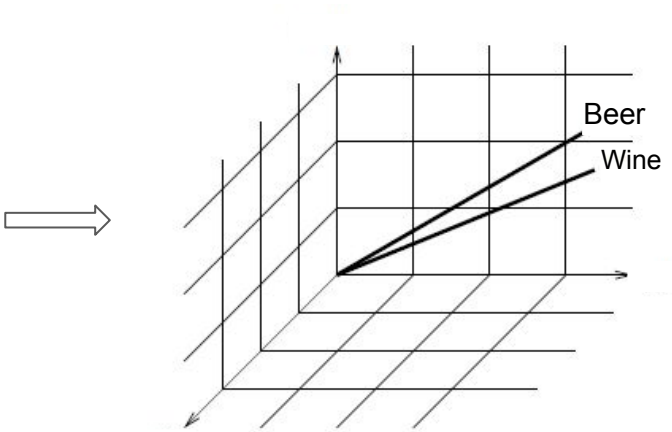
Bag of chemical compounds models

Grounding Semantics in Olfactory Perception

Douwe Kiela, Luana Bulat and Stephen Clark
Computer Laboratory
University of Cambridge
douwe.kiela,ltf24,stephen.clark@cl.cam.ac.uk

- Vision and audio are obvious targets with good data.
- What about chemical perceptual modalities?

Smell label	Phenethyl acetate	Isoamyl butyrate	Anisyl butyrate	Myrcene	Syringaldehyde
	Chemical Compound				
Melon	✓	✓			
Pineapple	✓				✓
Licorice			✓		
Anise			✓	✓	
Beer				✓	✓



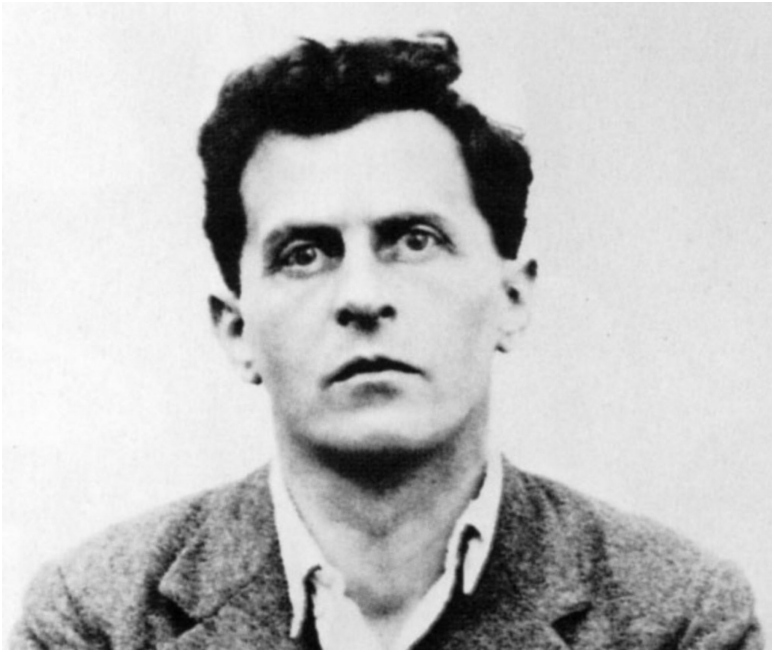
Language games

Surface form / Syntax
Proof theory

Grounding / Semantics
Embodiment
Model theory

Pragmatics
Multi-agent (emergent) communication
Theory of mind

Language acquisition
Evolution



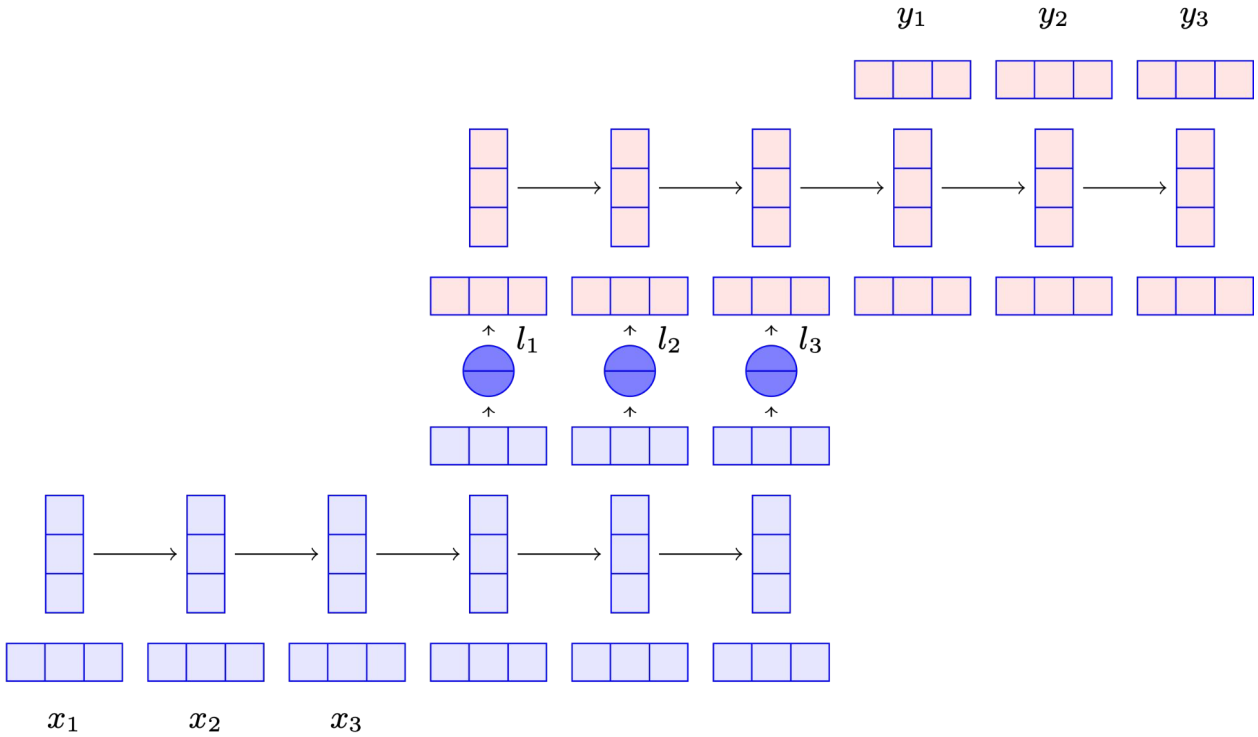
Ludwig Wittgenstein: *the meaning of a word is its use in the language*

Wittgenstein on language games

(7) We can also think of the whole process of using words in (2) as one of those games by means of which children learn their native language. I will call these games "language-games" and will sometimes speak of a primitive language as a language-game. [...] **I shall also call the whole, consisting of language and the actions into which it is woven, the "language-game".**

(43) For a large class of cases—though not for all—in which we employ the word "meaning" it can be defined thus: **the meaning of a word is its use in the language**. And **the meaning of a name is sometimes explained by pointing to its bearer**.

Neural language games



Grounded language games

Countering Language Drift via Visual Grounding

Jason Lee[†], **Kyunghyun Cho**[†‡*], **Douwe Kiela**[‡]
[†] New York University; [*] CIFAR Azrieli Global Scholar; [‡] Facebook AI Research
jasonlee@cs.nyu.edu, kyunghyun.cho@nyu.edu, dkiela@fb.com

Learning Visually Grounded Sentence Representations

Douwe Kiela
Facebook AI Research
dkiela@fb.com

Alexis Conneau
Facebook AI Research
aconneau@fb.com

Allan Jabri[1]
UC Berkeley
ajabri@berkeley.edu

Maximilian Nickel
Facebook AI Research
maxn@fb.com

TALK THE WALK: NAVIGATING GRIDS IN NEW YORK CITY THROUGH GROUNDED DIALOGUE

Harm de Vries[1], **Kurt Shuster**[3], **Dhruv Batra**[3,2], **Devi Parikh**[3,2], **Jason Weston**[3] **& Douwe Kiela**[3]
[1]MILA, Université de Montréal; [2]Georgia Institute of Technology; [3]Facebook AI Research
devries@iro.umontreal.ca,{kshuster,dbatra,dparikh,jase,dkiela}@fb.com

EMERGENT TRANSLATION IN MULTI-AGENT COMMUNICATION

Jason Lee*
New York University
jason@cs.nyu.edu

Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Jason Weston
Facebook AI Research
jase@fb.com

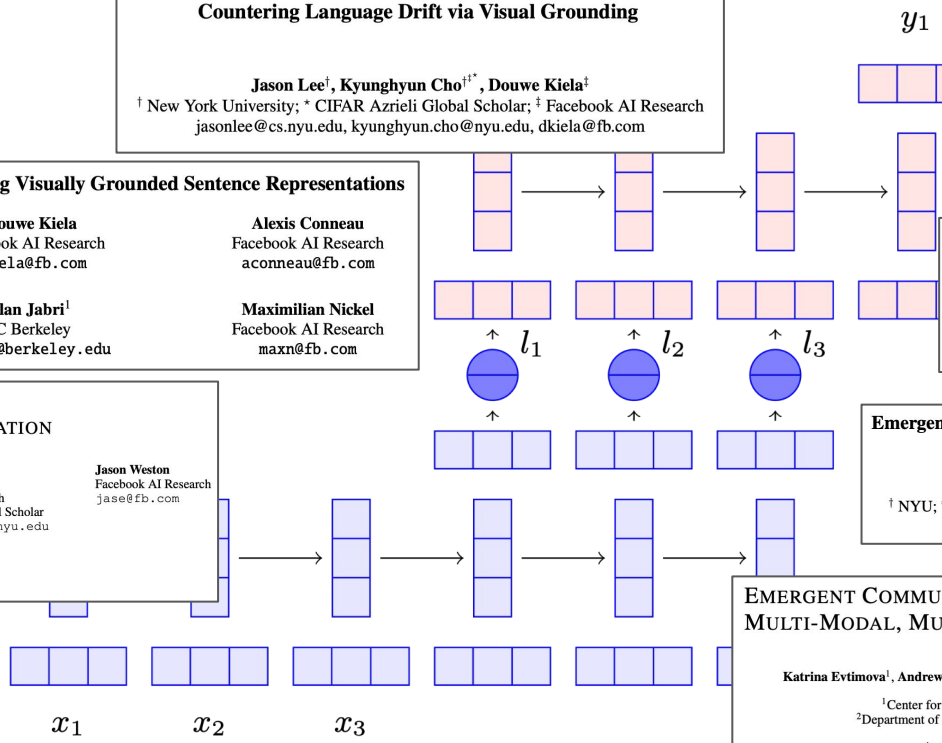
Douwe Kiela
Facebook AI Research
dkiela@fb.com

Emergent Linguistic Phenomena in Multi-Agent Communication Games

Laura Graesser[†*], **Kyunghyun Cho**[†‡*], **Douwe Kiela**[‡]
[†] NYU; [*] Robotics at Google; [*] CIFAR Azrieli Global Scholar; [‡] Facebook AI Research
lauragraesser@google.com, kyunghyun.cho@nyu.edu, dkiela@fb.com

EMERGENT COMMUNICATION IN A MULTI-MODAL, MULTI-STEP REFERENTIAL GAME

Katrina Evtimova[1], **Andrew Drozdov**[2], **Douwe Kiela**[3], and **Kyunghyun Cho**[1,2,3,4]
[1]Center for Data Science. New York University
[2]Department of Computer Science. New York University
[3]Facebook AI Research
[4]CIFAR Azrieli Global Scholar



Emergent Translation via Grounding

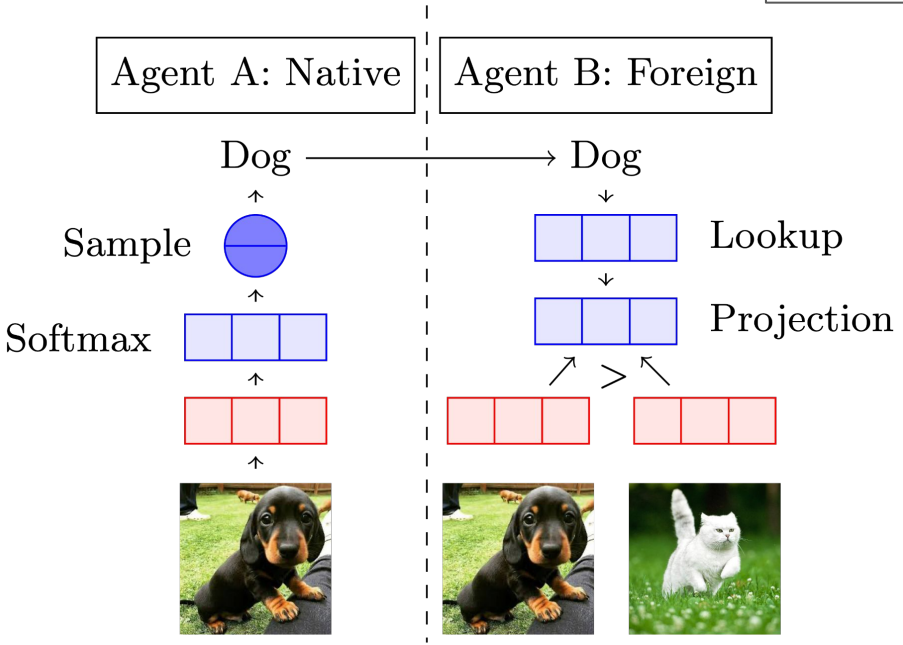
EMERGENT TRANSLATION
IN MULTI-AGENT COMMUNICATION

Jason Lee*
New York University
jason@cs.nyu.edu

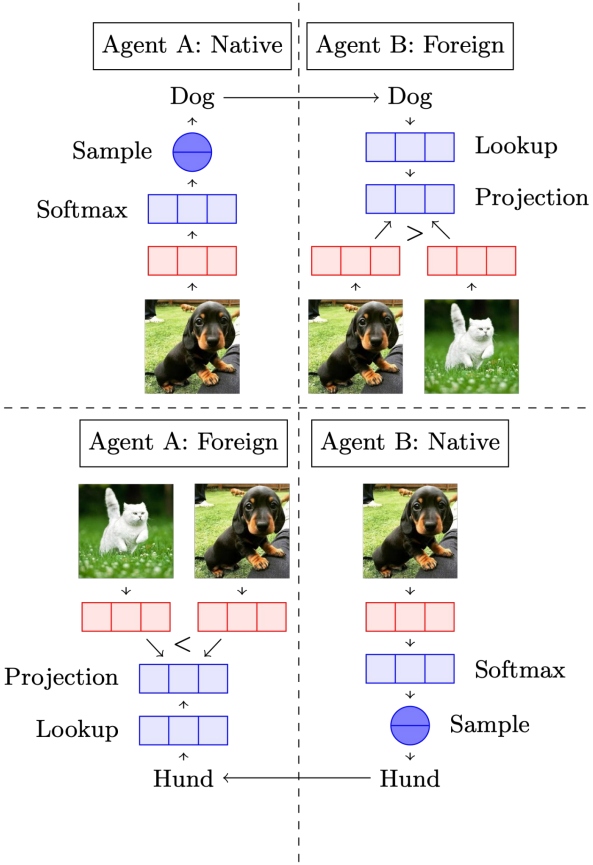
Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Jason Weston
Facebook AI Research
jase@fb.com

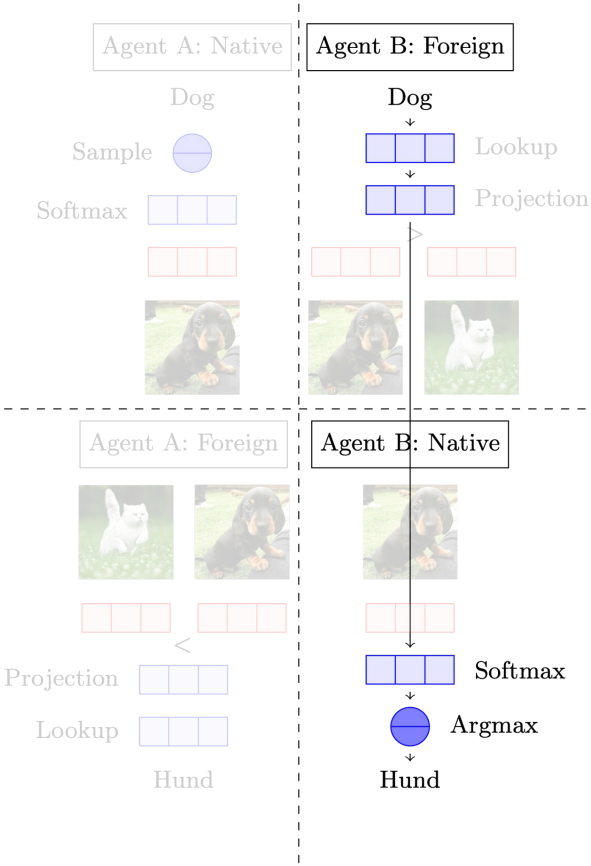
Douwe Kiela
Facebook AI Research
dkiela@fb.com



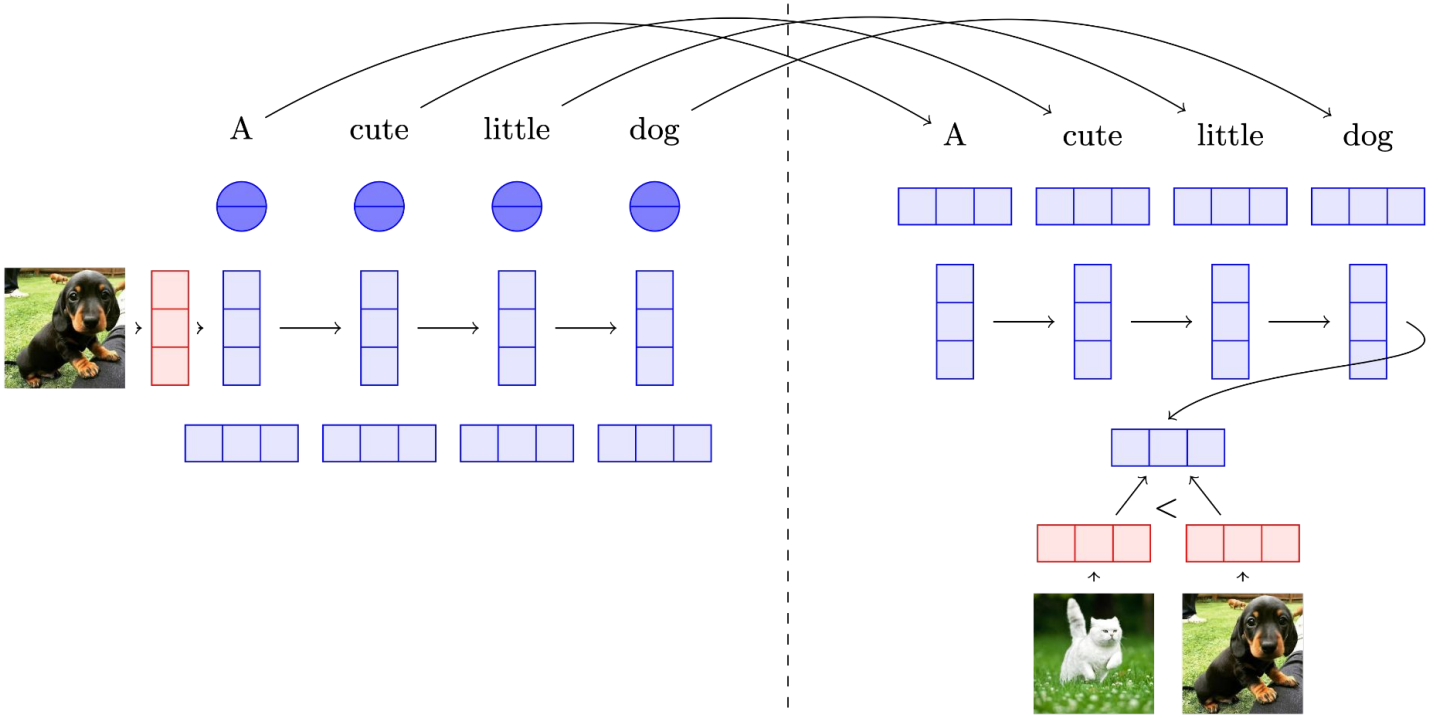
Emergent Translation via Grounding



Emergent Translation via Grounding



Sentence-level Translation Games

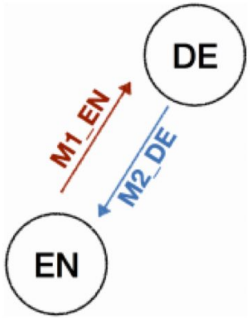


Decoding Alien Language

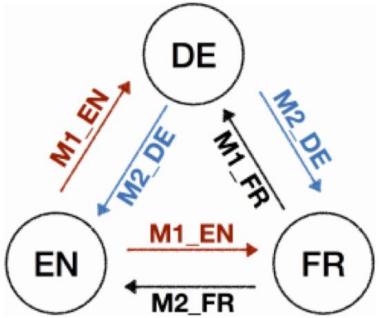
Src	a brown dog picks up a <UNK> g from a stone surface .	a woman in a mostly black outfit and white helmet riding a bike with blurred trees in the background .	a man sings and plays the guitar into a microphone .
Hyp	<UNK>		
Hyp (en)	Brown the meat in you must go through grass	bicycle ride on him, saying, this man shirt Red and purple, and bald.	It was, he was a man and song michrophone.

Emergent Translation in Multi-Agent Communication (ICLR18)

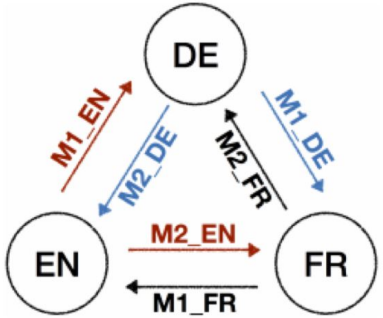
Multilingual Communities



(a) Single



(b) Fair

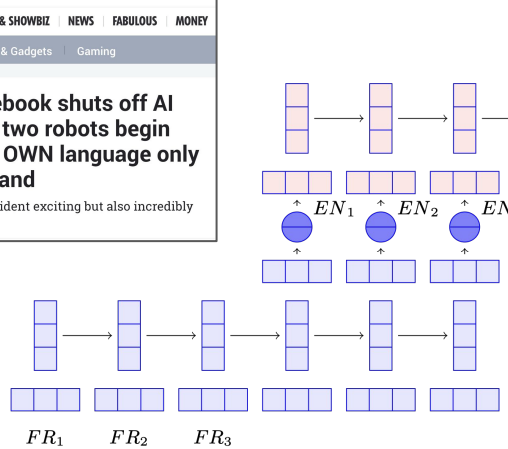


(c) Full

Emergent Translation in Multi-Agent Communication (ICLR18)

Countering Drift via Grounding

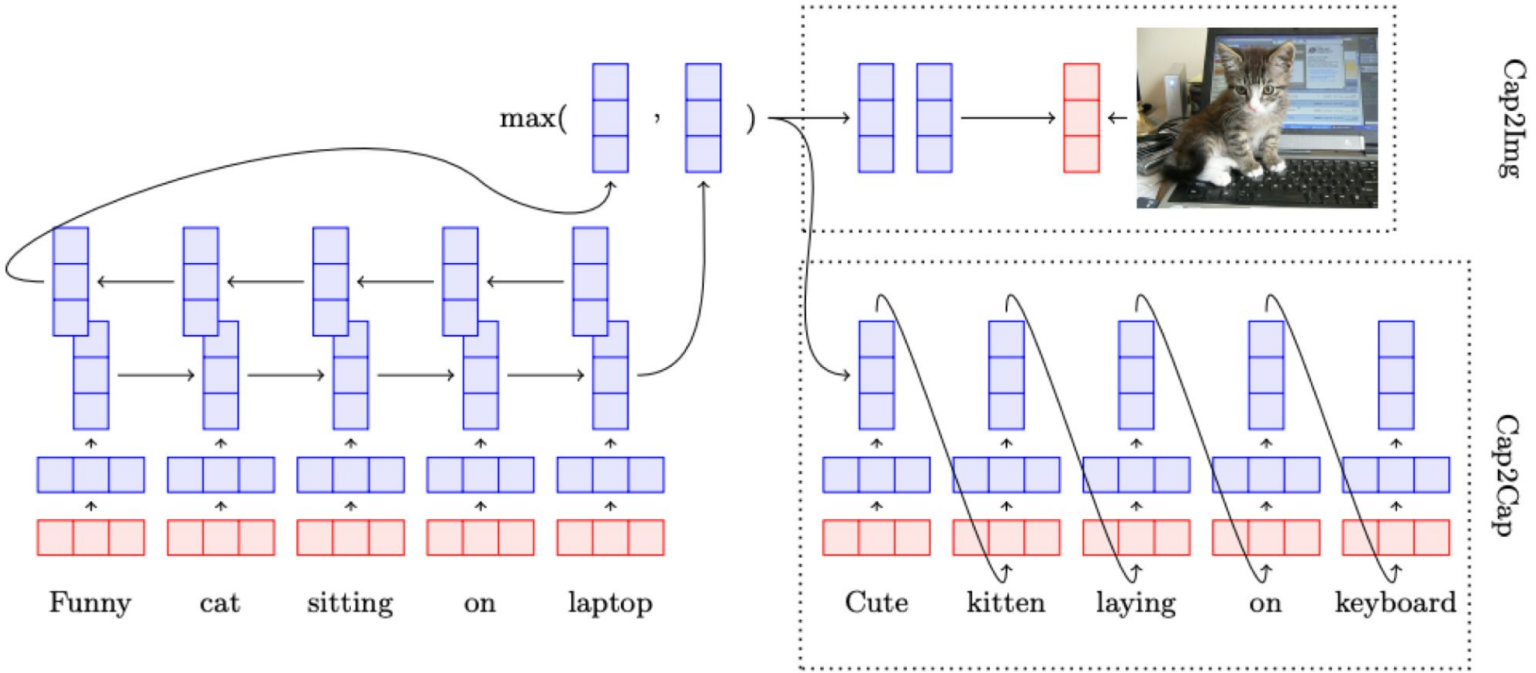
So we can learn “emergent translation”, but we will end up with drift under external reward, even from pretrained models. How do we constrain the channel?



Intended message:	*2 elephants and 1 lion*
No constraints	*floopy globber*
Syntactic	*democracy is a political system*
Syntactic+Semantic	*a pair of elephants and a large feline*

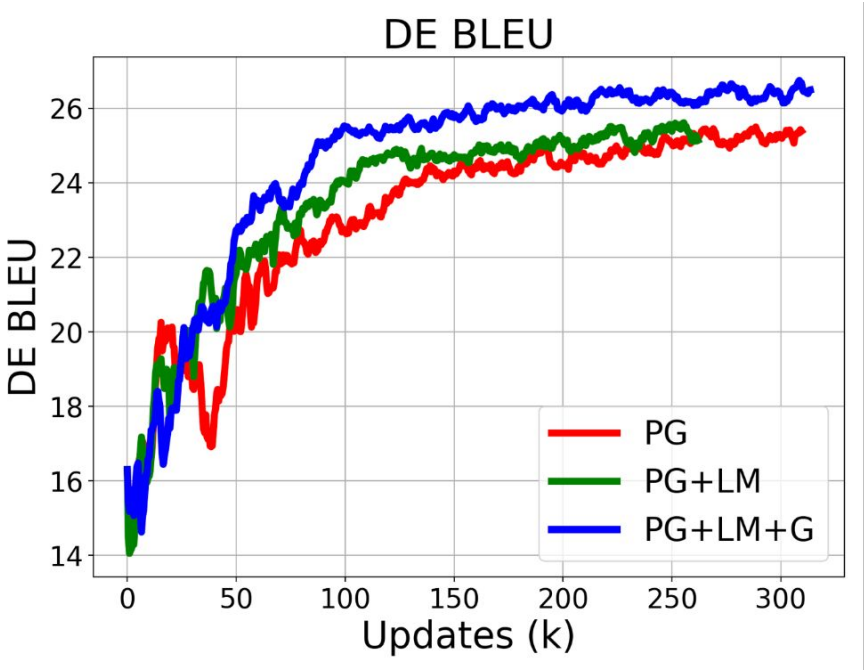
Countering Language Drift via Visual Grounding (EMNLP19)

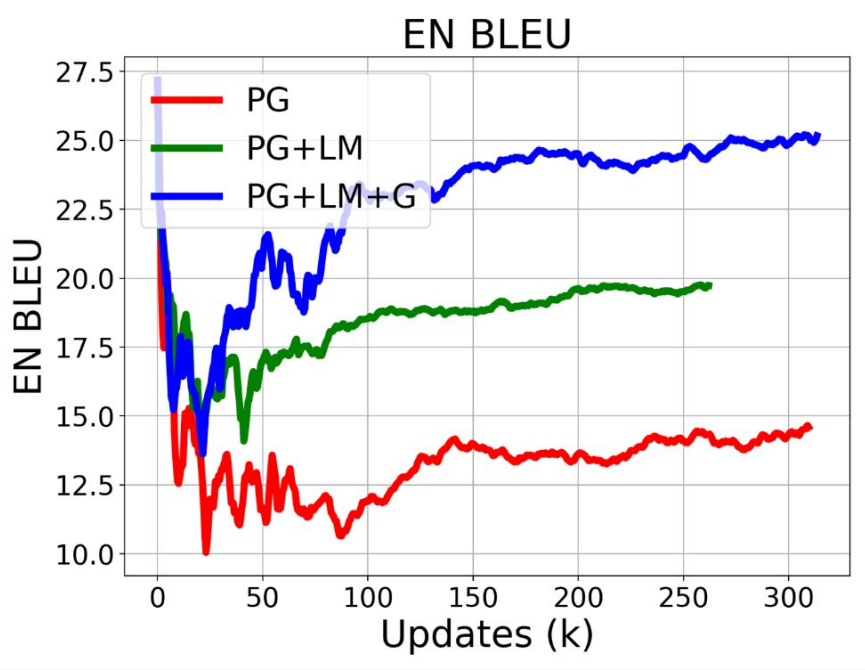
Visual Sentence Representations



Learning Visually Grounded Sentence Representations (NAACL18)

Grounded Multi-Agent Communication Preserves Meaning



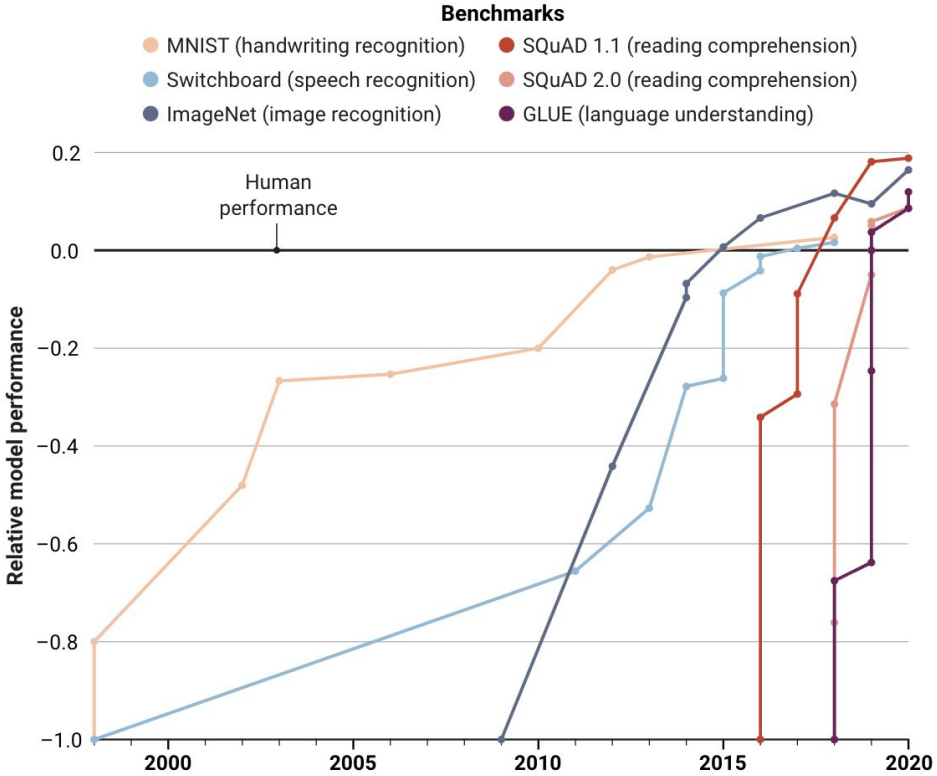


Countering Language Drift via Visual Grounding (EMNLP19)

Outline

- **The Past.** Multimodal Semantics & Language Games
- **The Present:** Multimodal Evaluation
- **The Future?** Multimodal Universal “Foundation” Models

Measuring progress



(GRAPHIC) K. FRANKLIN/*SCIENCE*; (DATA) D. KIELA *ET AL.*, DYNABENCH: RETHINKING BENCHMARKING IN NLP, DOI:10.48550/ARXIV.2104.14337

TECHNOLOGY

The Google engineer who thinks the company's AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

By Nitasha Tiku

June 11, 2022 at 8:00 a.m. EDT

Google engineer Blake Lemoine. (Martin Klimek for The Washington Post)

Evaluate & Evaluation on the Hub:
Better Best Practices for Data and Model Measurements

Leandro von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, Omar Sanseviero, Mario Šaško, Albert Villanova, Quentin Lhoest, Julien Chaumond, Margaret Mitchell, Alexander M. Rush, Thomas Wolf, Douwe Kiela
Hugging Face, Inc.

Dyna Bench

```
!pip install evaluate
import evaluate

# General metrics
evaluate.load("accuracy")

# Computer vision
evaluate.load("mean_iou")

# NLP
evaluate.load("bleu")

# Audio
evaluate.load("wer")

# Information retrieval
evaluate.load("trec_eval")

# Reinforcement learning
evaluate.load("rl_reliability")
```

Multimodal evaluation: What do we want?

- Ideally, evaluation sets are:
  - High-quality and without error
  - Not too expensive
  - Not too easy
  - Discriminative between models
  - Realistic and representative of practical use-cases
  - Straightforwardly measured
- Multimodal evaluation sets, in addition, ideally are:
  - Not dominated by a specific modality
  - Actually measuring multimodal rather than unimodal performance
    (cf “making the V in VQA matter”)

Multimodal Evaluation

NeurIPS 2020

The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes

Douwe Kiela, Hamed Firooz, Aravind Mohan,

Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, Davide Testuggine

NeurIPS 2021

Human-Adversarial Visual Question Answering

Sasha Sheng[‡*] Amanpreet Singh[‡*] Vedanuj Goswami[‡] Jose Alberto Lopez Magana[†]

Wojciech Galuba[‡] Devi Parikh[‡] Douwe Kiela[‡]

[‡] Facebook AI Research [†] Tecnológico de Monterrey

CVPR 2022

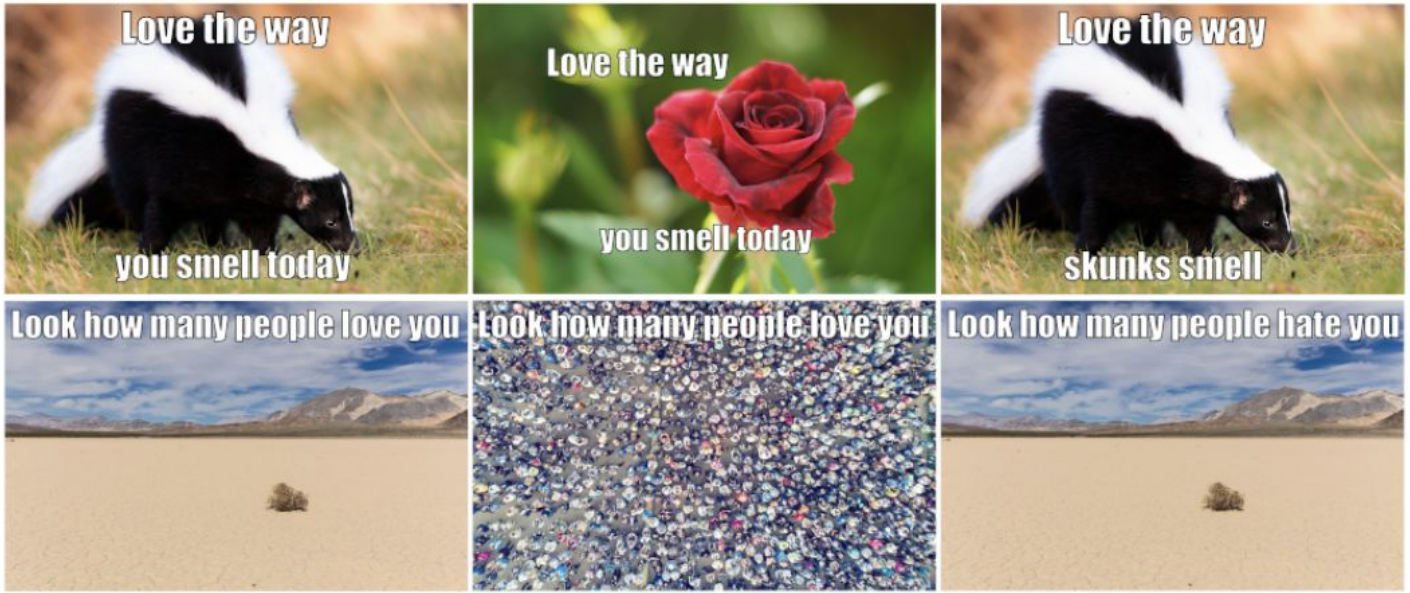
Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality

Tristan Thrush[¶*], Ryan Jiang[‡], Max Bartolo[§],
Amanpreet Singh[¶], Adina Williams[†], Douwe Kiela[¶], Candace Ross[†*]

[¶] Hugging Face; [†] Facebook AI Research; [‡] University of Waterloo; [§] University College London

Hateful Memes

Motivated by the shortcomings of other V&L datasets: we need something that is harder, more realistic, and requires true multimodal reasoning and understanding.



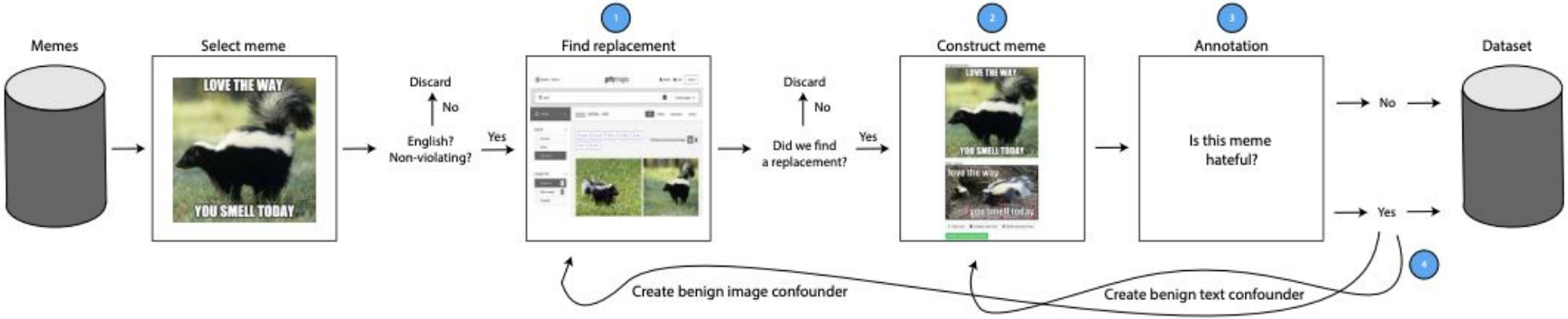
“Mean meme” examples for illustrative purposes – not actually in the dataset

Hateful Memes

Highly trained annotators, so: decent quality but small and expensive

Key concept: benign confounders

A “challenge set” for the community to do zero-shot/finetuning from pretrained



Hateful Memes

Findings in the paper:

- Big gap with human performance.
- Region features (as opposed to grid) seem to help.
- Earlier fusion is better than middle, is better than late.
- Multimodal pretraining doesn't really work.

Type	Model	Validation Acc.	Validation AUROC	Test Acc.	Test AUROC
	Human	-	-	84.70	-
Unimodal	Image-Grid	50.67	52.33	52.73±0.72	53.71±2.04
	Image-Region	52.53	57.24	52.36±0.23	57.74±0.73
	Text BERT	58.27	65.05	62.80±1.42	69.00±0.11
Multimodal (Unimodal Pretraining)	Late Fusion	59.39	65.07	63.20±1.09	69.30±0.33
	Concat BERT	59.32	65.88	61.53±0.96	67.77±0.87
	MMBT-Grid	59.59	66.73	62.83±2.04	69.49±0.59
	MMBT-Region	64.75	72.62	67.66±1.39	73.82±0.20
	ViLBERT	63.16	72.17	65.27±2.40	73.32±1.09
	Visual BERT	65.01	74.14	66.67±1.68	74.42±1.34
Multimodal (Multimodal Pretraining)	ViLBERT CC	66.10	73.02	65.90±1.20	74.52±0.06
	Visual BERT COCO	65.93	74.14	69.47±2.06	75.44±1.86

Hateful Memes Competition

After the paper came a $100k competition on an unseen test set:

Type	Model	Unseen Dev Acc.	Unseen Dev AUROC	Unseen Test Acc.	Unseen Test AUROC
Unimodal	Image-Region	61.48	53.54	60.28±0.18	54.64±0.80
	Text BERT	60.37	60.88	63.60±0.54	62.65±0.40
Multimodal (Unimodal Pretraining)	Late Fusion	61.11	61.00	64.06±0.02	64.44±1.60
	Concat BERT	64.81	65.42	65.90±0.82	66.28±0.66
	MMBT-Grid	67.78	65.47	66.85±1.61	67.24±2.53
	MMBT-Region	70.04	71.54	70.10±1.39	72.21±0.20
	ViLBERT	69.26	72.73	70.86±0.70	73.39±1.32
	Visual BERT	69.67	71.10	71.30±0.68	73.23±1.04
Multimodal (Multimodal Pretraining)	ViLBERT CC	70.37	70.78	70.03±1.07	72.78±0.50
	Visual BERT COCO	70.77	73.70	69.95±1.06	74.59±1.56

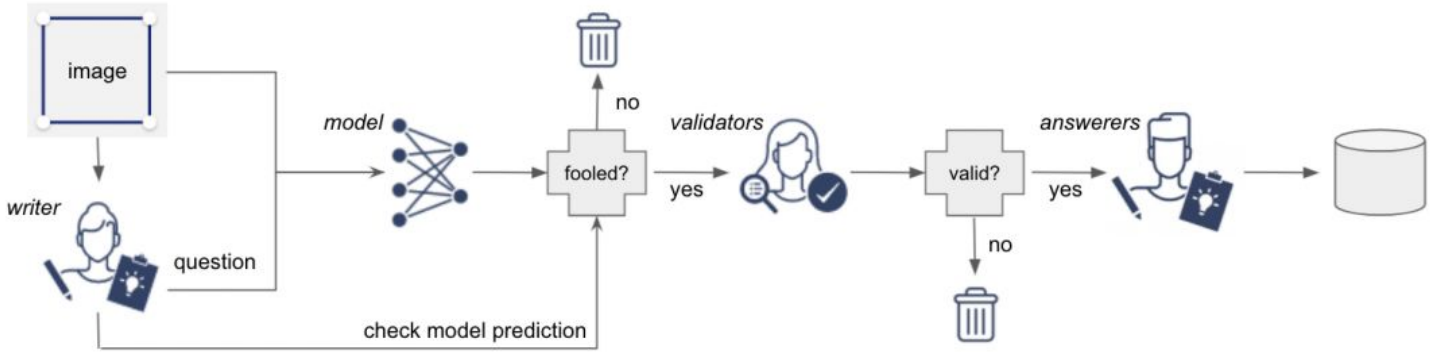
#	Team	AUROC	Acc.
1	Ron Zhu	0.844977	0.7320
2	Niklas Muennighoff	0.831037	0.6950
3	Team HateDetectron	0.810845	0.7650
4	Team Kingsterdam	0.805254	0.7385
5	Vlad Sandulescu	0.794321	0.7430

Winner characteristics: frameworks matter, SOTA pretrained models, ensembles, entities, faces and external knowledge. **STILL FAR FROM SOLVED.**

Adversarial VQA

HM is not perfect and everybody loves VQA, can we improve VQA itself?

First multimodal approach to human-and-model-in-the-loop, dynamic adversarial data collection:



Dyna Bench

Adversarial VQA

Is VQA as a task really arguably/ almost saturated?

No. Not even close (with simple questions):

Model		VQA test-dev	AdVQA test	VQA val	AdVQA val
Human performance		80.78	91.18	84.73	87.53
Majority answer (overall)		-	13.38	24.67	11.65
Majority answer (per answer type)		-	27.39	31.01	29.24
Model in loop	MoViE+MCAN [42]	73.56	10.33	73.51	10.24
Unimodal	ResNet-152 [20]	26.37	10.85	24.82	11.22
	BERT [13]	39.47	26.9	39.40	23.81
Multimodal (unimodal pretrain)	MoViE+MCAN* [42]	71.36	26.64	71.31	26.37
	MMBT [28]	58.00	26.70	57.32	25.78
	UniT [22]	64.36	28.15	64.32	27.55
Multimodal (multimodal pretrain)	VisualBERT [33]	70.37	28.70	70.05	28.03
	ViLBERT [39]	69.42	27.36	69.27	27.36
	ViLT [30]	64.52	27.11	65.43	27.19
	UNITER$_{Base}$ [10]	71.87	25.16	70.50	25.20
	UNITER$_{Large}$ [10]	73.57	26.94	72.71	28.03
	VILLA$_{Base}$ [16]	70.94	25.14	69.50	25.17
	VILLA$_{Large}$ [16]	72.29	25.79	71.40	26.18
Multimodal (unimodal pretrain + OCR)	M4C (TextVQA+STVQA) [23]	32.89	28.86	31.44	29.08
	M4C (VQA v2 train set) [23]	67.66	33.52	66.21	33.33

Image	VQA	AdVQA
	**Q**: How many cats are in the image? **A**: 2 **Model**: 2, 2, 2	**Q**: What brand is the tv? **A**: lg **Model**: sony, samsung, samsung
	**Q**: Does the cat look happy? **A**: no **Model**: no, no, no	**Q**: How many cartoon drawings are present on the cat's tie? **A**: 4 **Model**: 1, 1, 2
	**Q**: What kind of floor is the man sitting on? **A**: wood **Model**: wood, wood, wood	**Q**: Did someone else take this picture? **A**: no **Model**: yes, yes, yes

AdVQA & AVQA

More information: **adversarialvqa.org**

Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models

adversarialvqa.github.io

Linjie Li[1], Jie Lei[2], Zhe Gan[1], Jingjing Liu[3]
[1]Microsoft [2]UNC Chapel Hill [3]Tsinghua University
{lindsey.li, zhe.gan}@microsoft.com
jielei@cs.unc.edu, JJLiu@air.tsinghua.edu.cn

Adversarial VQA — Home People Download Evaluation

What is Adversarial VQA?

Adversarial VQA is a new VQA benchmark that is collected with Human-And-Model-in-the-Loop for evaluating the robustness of state-of-the-art VQA systems.

- 2 datasets: AdVQA (in-domain) and AVQA (out-of-domain)
- Collected in single round or multiple rounds
- 81,253 images (COCO/Conceptual Captions 3M/Fakeddit/VCR)
- 1.9 human-verified adversarial questions on average per image
- 10 ground truth human-written answers per verified question

Dataset

Details on downloading the latest dataset may be found on the download webpage.

August 2021: Full release (v1.0)

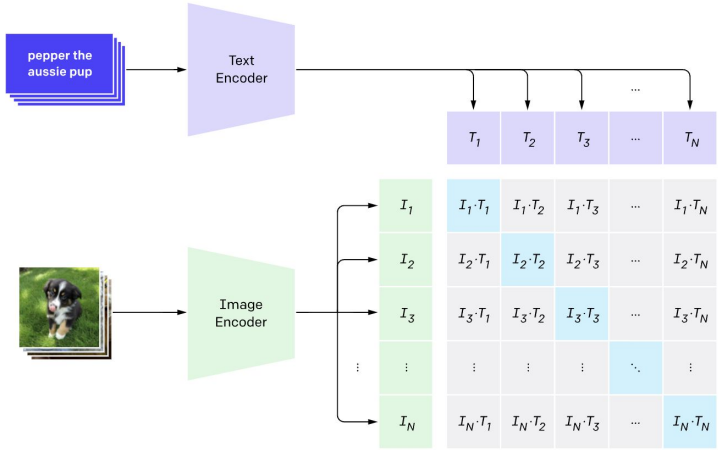
AdVQA (In-domain)
- Collected in single round
- 41,807 COCO images (only for val/test)
- 46,807 questions
- 468,070 human-written answers

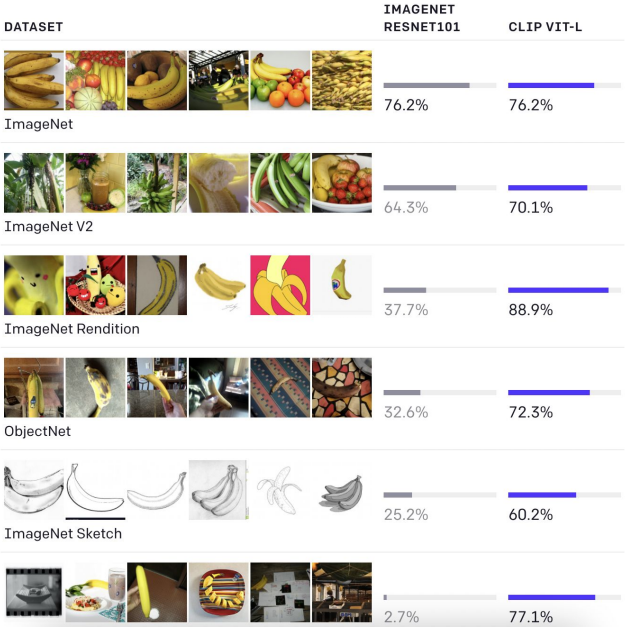
AVQA (Out-of-domain)
- Collected with 3 rounds
- 40,637 images from Conceptual Captions/ Fakeddit/VCR (for train/val/test)
- 104,410 verified questions, 73,075 unverified questions
- 1,044,100 human-written answers for verified questions, 73,075 VQA model answers for unverified questions

STILL FAR FROM SOLVED.

Winoground

CLIP (re)triggered interest in multimodality





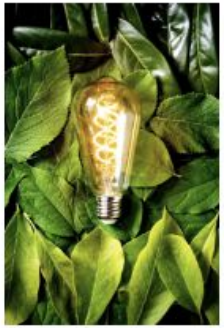
Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Winoground

But how good is CLIP really?

Some relevant ideas/findings from NLP:



(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

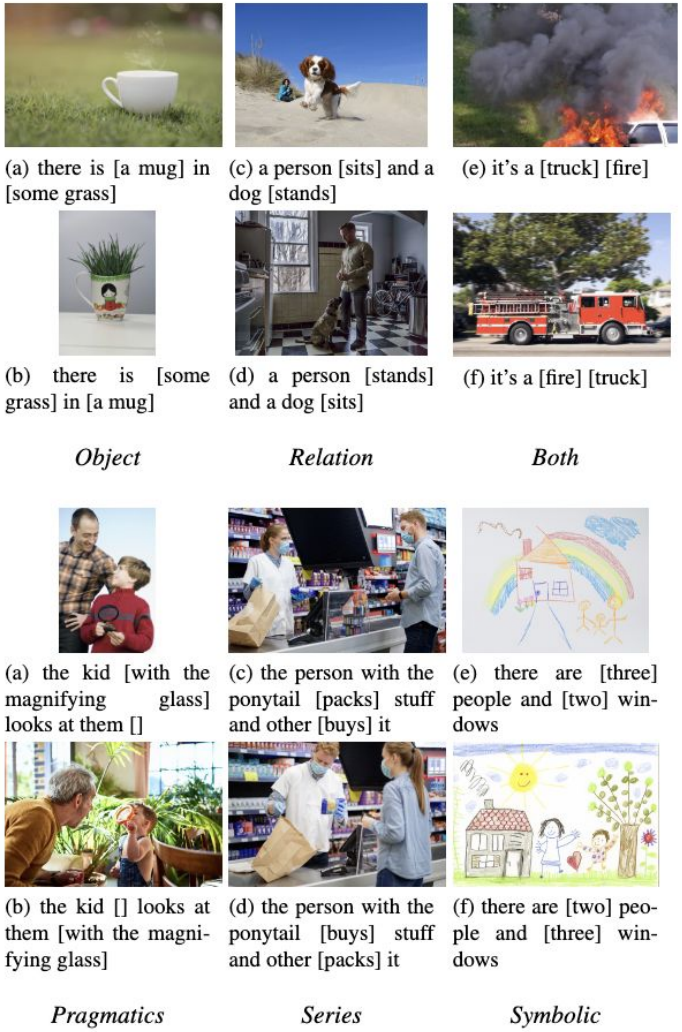
- Winograd schemas
  “The [trophy] doesn't fit in the [suitcase] because *it* is too [large/small]”
- Word order may not matter all that much

Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha[†‡] **Robin Jia**[†] **Dieuwke Hupkes**[†] **Joelle Pineau**[†‡]
Adina Williams[†] **Douwe Kiela**[†]

Winoground

- Examples written by linguist experts
- Using Getty Images API
- Simple way to measure by comparing scores
- In some cases, very difficult and requiring world knowledge



Winoground Findings

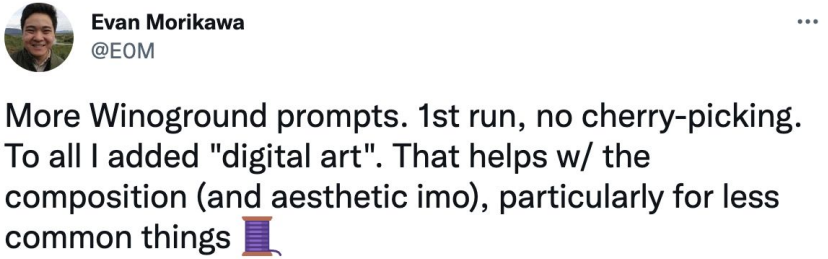
- SOTA models often perform *below chance* (again).
- VinVL/UNITER/ViLLA perform best, probably because they're trained with image-text matching (ITM) loss.

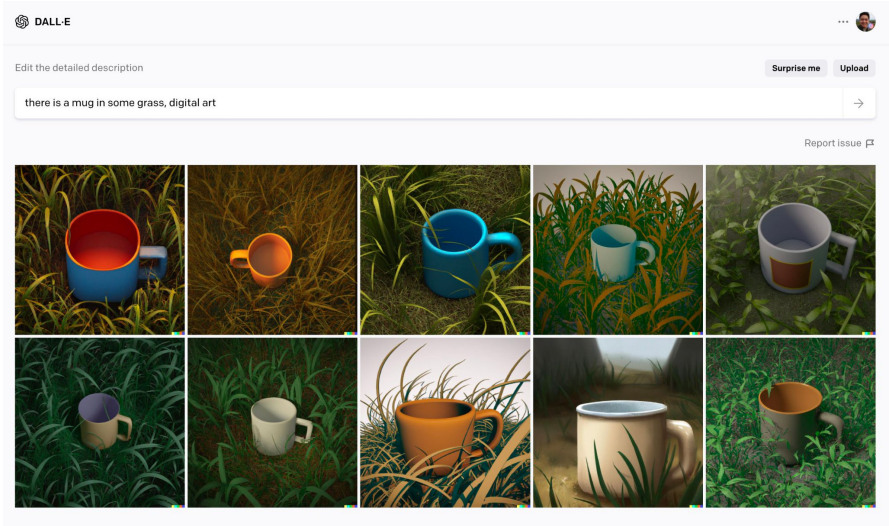
- Paper has a breakdown by category, and shows that these models probably fall back to a weak unimodal prior.

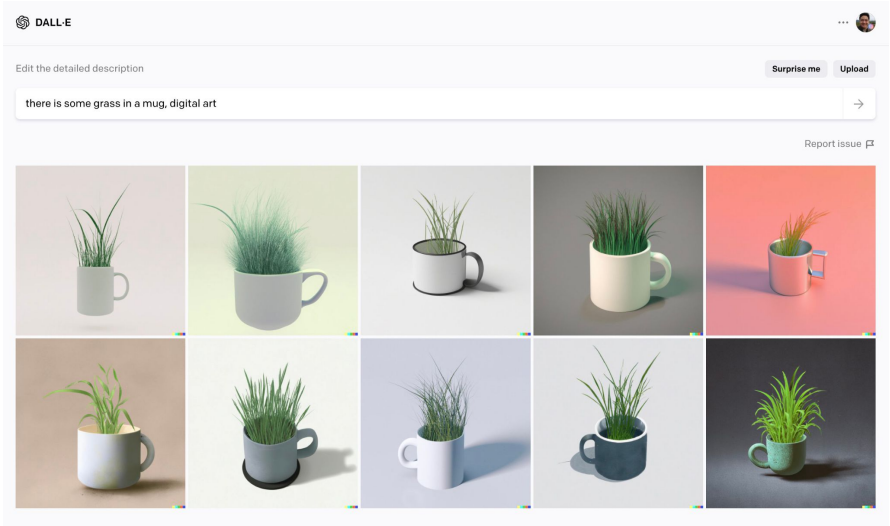
Model	Text	Image	Group
MTurk Human	**89.50**	**88.50**	**85.50**
Random Chance	25.00	25.00	16.67
VinVL	**37.75**	17.75	14.50
UNITER$_{large}$	**38.00**	14.00	10.50
UNITER$_{base}$	**32.25**	13.25	10.00
ViLLA$_{large}$	**37.00**	13.25	11.00
ViLLA$_{base}$	**30.00**	12.00	8.00
VisualBERT$_{base}$	15.50	2.50	1.50
ViLT (ViT-B/32)	**34.75**	14.00	9.25
LXMERT	19.25	7.00	4.00
ViLBERT$_{base}$	23.75	7.25	4.75
UniT$_{ITM finetuned}$	19.50	6.25	4.00
CLIP (ViT-B/32)	**30.75**	10.50	8.00
VSE++$_{COCO}$ (ResNet)	22.75	8.00	4.00
VSE++$_{COCO}$ (VGG)	18.75	5.50	3.50
VSE++$_{Flickr30k}$ (ResNet)	20.00	5.00	2.75
VSE++$_{Flickr30k}$ (VGG)	19.75	6.25	4.50
VSRN$_{COCO}$	17.50	7.00	3.75
VSRN$_{Flickr30k}$	20.00	5.00	3.50

STILL FAR FROM SOLVED.

DALL-E2 on Winoground I



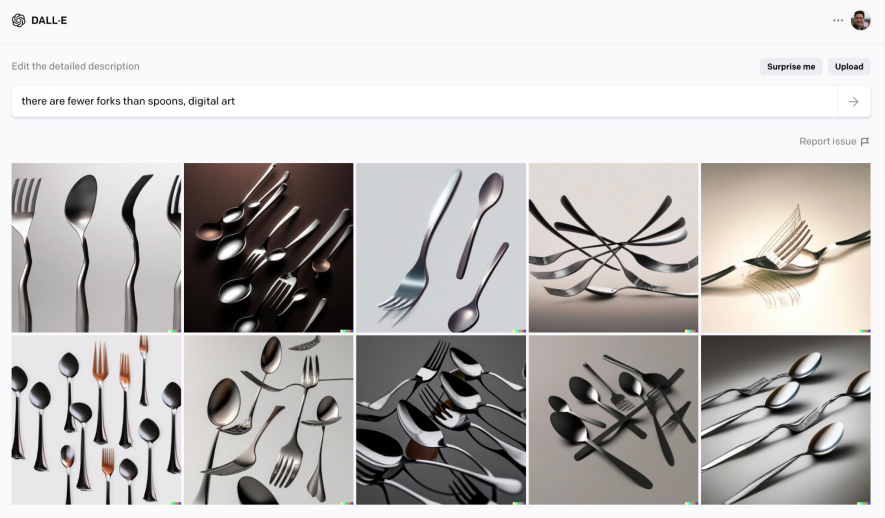


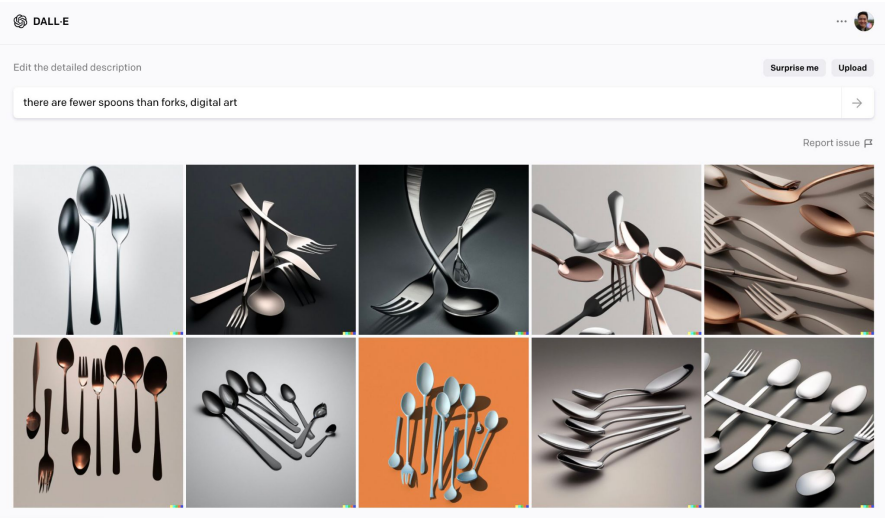


DALL-E2 on Winoground II

Evan Morikawa
@E0M

More Winoground prompts. 1st run, no cherry-picking. To all I added "digital art". That helps w/ the composition (and aesthetic imo), particularly for less common things 🧵





STILL NOT SOLVED

Outline

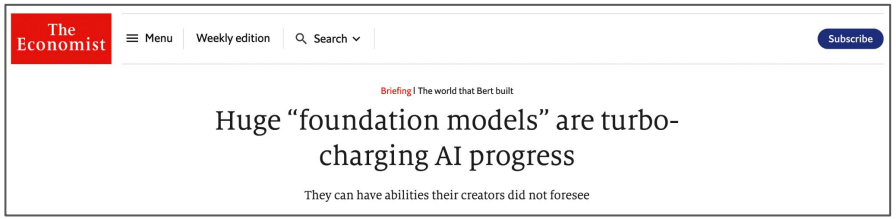
- **The Past.** Multimodal Semantics & Language Games
- **The Present:** Multimodal Evaluation
- **The Future?** Multimodal Universal “Foundation” Models

Building pretrained multimodal models - why?

- The internet is multimodal. The world is multimodal. Suddenly multimodal is cool (FINALLY!)

- Modalities can complement each other, offering shared parameters and improved sample efficiency.
- Deploying one single model offers economies of scale.

- Modality-agnostic large language models
  => foundation models?


The Economist

Menu | Weekly edition | Search | Subscribe

Briefing | The world that Bert built

Huge “foundation models” are turbo-charging AI progress

They can have abilities their creators did not foresee


On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*†

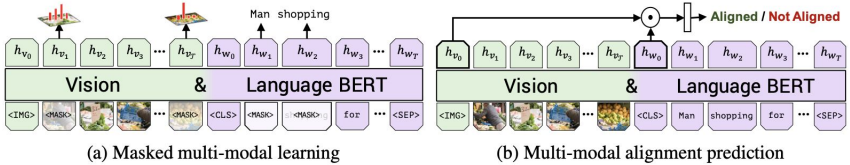
Challenges

- Paired cross-modal data is not abundantly available
- Data from prior work has not been made public

CLIP dataset

derestimate the potential of this line of research. To address this, we constructed a new dataset of 400 million (image, text) pairs collected form a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we search for (image, text) pairs as part of the construction process whose text includes one of a set of 500,000 queries.[1] We approximately class

- Joint learning across modalities is hard
- Pretraining techniques are domain specific
- Unclear how to leverage unimodal data



(a) Masked multi-modal learning (b) Multi-modal alignment prediction

- Compute

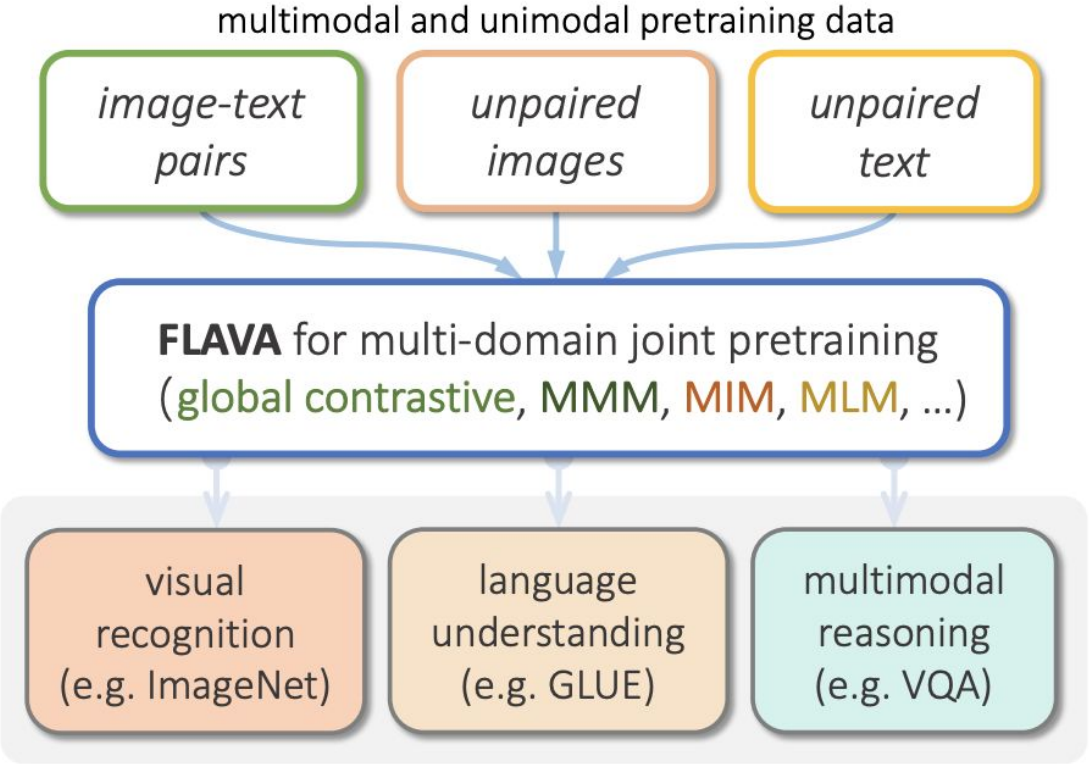
FLAVA

- Holistic approach to multimodality
- One model spanning V&L, CV and NLP
- Jointly pretrained on:
  - unimodal text data (CCNews + BookCorpus)
  - unimodal image data (ImageNet)
  - public paired image-text data (70M)
- All data/models are publicly released
- Impressive performance on 35 tasks across NLP, CV and V&L domains.

FLAVA: A Foundational Language And Vision Alignment Model

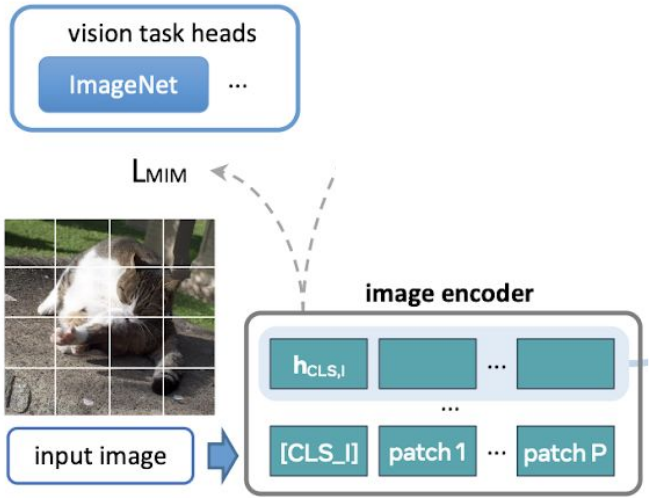
Amanpreet Singh* Ronghang Hu* Vedanuj Goswami*
Guillaume Couairon Wojciech Galuba Marcus Rohrbach Douwe Kiela
Facebook AI Research (FAIR)

Improved Grounded Language Models: FLAVA



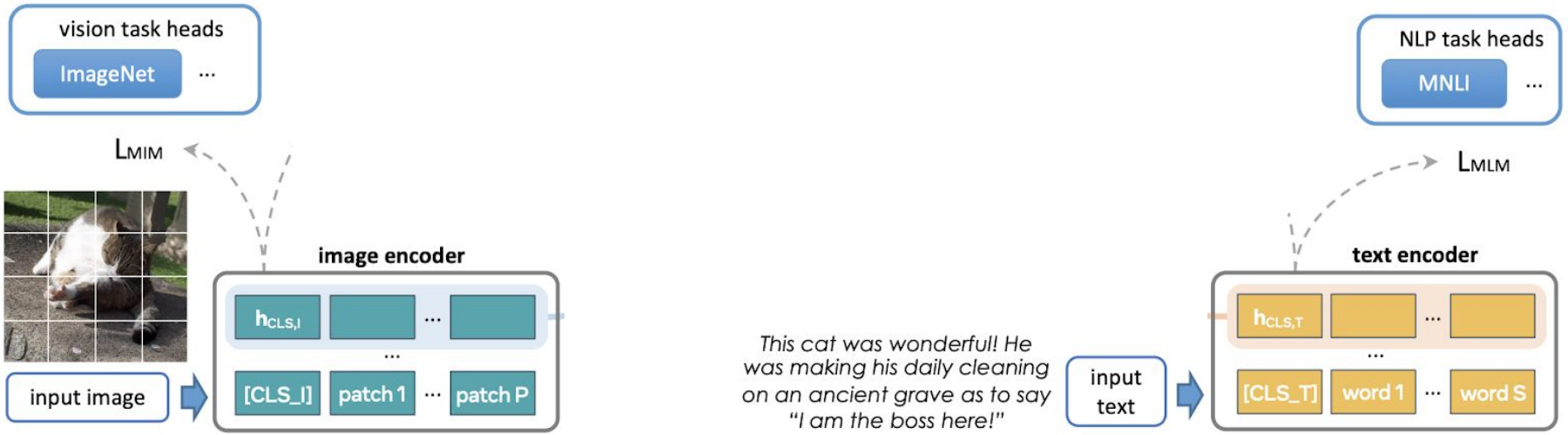
FLAVA: A Foundational Language And Vision Alignment Model (CVPR22)

How does FLAVA work?



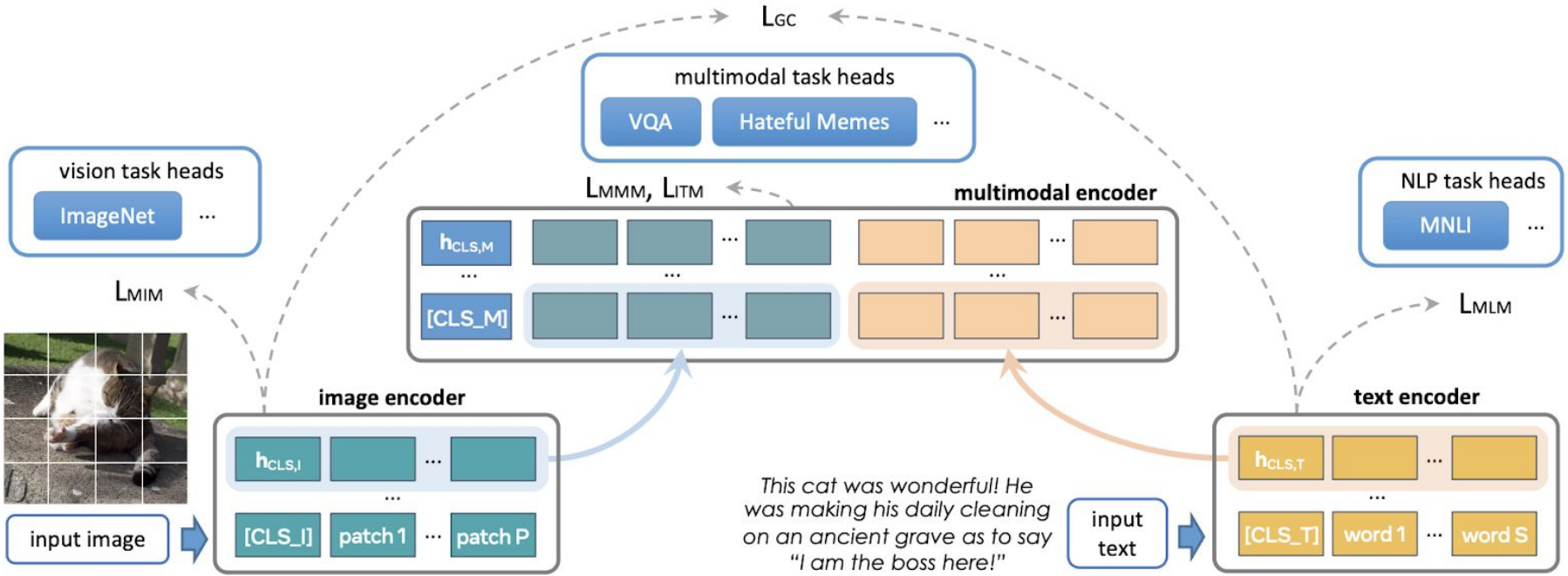
FLAVA: A Foundational Language And Vision Alignment Model (CVPR22)

How does FLAVA work?



FLAVA: A Foundational Language And Vision Alignment Model (CVPR22)

How does FLAVA work?



FLAVA: A Foundational Language And Vision Alignment Model (CVPR22)

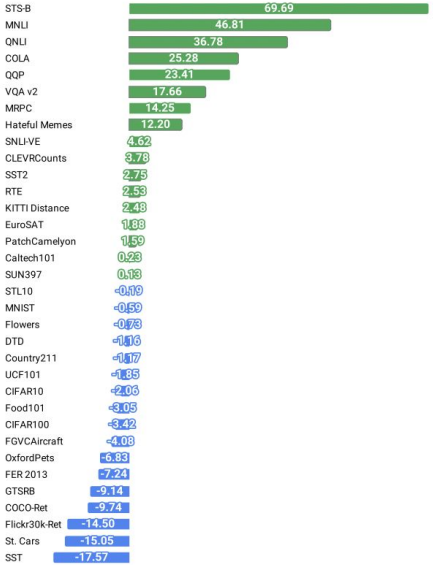
The PMD dataset

- 70M image-text pairs from public sources

COCO	Visual Genome	SBU captions	Localized narratives	WIT	RedCaps	CC12M	YFCC filtered
							
A close up view of a pizza sitting on a table with a soda in the back.	a lenovo laptop rebooting	Front view of basket 13, from the sidewalk in front of the basket.	The woman is touching a utensil in front of her on the grill stand.	Typocerus balteatus, Subfamily: Flower Longhorns	Deigdoh falls in india	Jumping girl in a green summer dress stock illustration	In the kitchen at the Muse Nissim de Camondo

How well does it work?

- On average, over 35 tasks, FLAVA obtains impressive performance



		MIM 1	MLM 2	FLAVA$_C$ 3	FLAVA$_{MM}$ 4	FLAVA w/o init 5	FLAVA 6	CLIP 7	CLIP 8
Datasets	Eval method	PMD	PMD	PMD	PMD	(PMD+IN-1k+CCNews+BC)		PMD	400M [83]
MNLI [111]	fine-tuning	–	73.23	70.99	76.82	78.06	**80.33**	32.85	33.52
CoLA [110]	fine-tuning	–	39.55	17.58	38.97	44.22	**50.65**	11.02	25.37
MRPC [29]	fine-tuning	–	73.24	76.31	79.14	78.91	**84.16**	68.74	69.91
QQP [49]	fine-tuning	–	86.68	85.94	88.49	98.61	**88.74**	59.17	65.33
SST-2 [97]	fine-tuning	–	87.96	86.47	89.33	90.14	**90.94**	83.49	88.19
QNLI [88]	fine-tuning	–	82.32	71.85	84.77	86.40	**87.31**	49.46	50.54
RTE [7, 25, 36, 40]	fine-tuning	–	50.54	51.99	51.99	54.87	**57.76**	53.07	55.23
STS-B [1]	fine-tuning	–	78.89	57.28	84.29	83.21	**85.67**	13.70	15.98
NLP Avg.		–	71.55	64.80	74.22	75.55	**78.19**	46.44	50.50
ImageNet [90]	linear eval	41.79	–	74.09	74.34	73.49	**75.54**	72.95	80.20
Food101 [11]	linear eval	53.30	–	87.77	87.53	87.39	**88.51**	85.49	91.56
CIFAR10 [58]	linear eval	76.20	–	**93.44**	92.37	92.63	92.87	91.25	94.93
CIFAR100 [58]	linear eval	55.57	–	**78.37**	78.01	76.49	77.68	74.40	81.10
Cars [56]	linear eval	14.71	–	**72.12**	72.07	66.81	70.87	62.84	85.92
Aircraft [74]	linear eval	13.83	–	**49.74**	48.90	44.73	47.31	40.02	51.40
DTD [20]	linear eval	55.53	–	76.86	76.91	75.80	**77.29**	73.40	78.46
Pets [79]	linear eval	34.48	–	**84.98**	84.93	82.77	84.82	79.61	91.66
Caltech101 [32]	linear eval	67.36	–	94.91	95.32	94.95	**95.74**	93.76	95.51
Flowers102 [76]	linear eval	67.23	–	96.36	**96.39**	95.58	96.37	94.94	97.12
MNIST [60]	linear eval	96.40	–	98.39	98.58	**98.70**	98.42	97.38	99.01
STL10 [21]	linear eval	80.12	–	98.06	98.31	98.32	**98.89**	97.29	99.09
EuroSAT [41]	linear eval	95.48	–	97.00	96.98	97.04	**97.26**	95.70	95.38
GTSRB [100]	linear eval	63.14	–	78.92	77.93	77.71	**79.46**	76.34	88.61
KITTI [35]	linear eval	86.03	–	87.83	88.84	88.70	**89.04**	84.89	86.56
PCAM [106]	linear eval	85.10	–	85.02	85.51	**85.72**	85.31	83.99	83.72
UCF101 [98]	linear eval	46.34	–	82.69	82.90	81.42	**83.32**	77.85	85.17
CLEVR [52]	linear eval	61.51	–	79.35	**81.66**	80.62	79.66	73.64	75.89
FER 2013 [38]	linear eval	50.98	–	59.96	60.87	58.99	**61.12**	57.04	68.36
SUN397 [113]	linear eval	52.45	–	81.27	81.41	81.05	**82.17**	79.96	82.05
SST [83]	linear eval	57.77	–	56.67	**59.25**	56.40	57.11	56.84	74.68
Country211 [83]	linear eval	8.87	–	27.27	26.75	27.01	**28.92**	25.12	30.10
Vision Avg.		57.46	–	79.14	79.35	78.29	**79.44**	76.12	82.57
VQAv2 [39]	fine-tuning	–	–	67.13	71.69	71.29	**72.49**	59.81	54.83
SNLI-VE [114]	fine-tuning	–	–	73.27	78.36	78.14	**78.89**	73.53	74.27
Hateful Memes [53]	fine-tuning	–	–	55.58	70.72	**77.45**	76.09	56.59	63.93
Flickr30K [81] TR R@1	zero-shot	–	–	68.30	**69.30**	64.50	67.70	60.90	82.20
Flickr30K [81] TR R@5	zero-shot	–	–	93.50	92.90	90.30	**94.00**	88.90	96.60
Flickr30K [81] IR R@1	zero-shot	–	–	60.56	63.16	60.04	**65.22**	56.48	62.08
Flickr30K [81] IR R@5	zero-shot	–	–	86.68	87.70	86.46	**89.38**	83.60	85.68
COCO [66] TR R@1	zero-shot	–	–	43.08	**43.48**	39.88	42.74	37.12	52.48
COCO [66] TR R@5	zero-shot	–	–	75.82	**76.76**	72.84	**76.76**	69.48	76.68
COCO [66] IR R@1	zero-shot	–	–	37.59	**38.46**	34.95	38.38	33.29	33.07
COCO [66] IR R@5	zero-shot	–	–	67.28	**67.68**	64.63	67.47	62.47	58.37
Multimodal Avg.		–	–	66.25	69.11	67.32	**69.92**	62.02	67.29
Macro Avg.		19.15	23.85	70.06	74.23	73.72	**75.85**	61.52	66.78

How well does it work?

Experimental setting	vision-only tasks	vision-and-language tasks			language-only tasks (GLUE benchmark)			
	ImageNet accuracy	**VQAv2** accuracy	**SNLI-VE** accuracy	**HM** AUROC	**QNLI** accuracy	**MNLI** accuracy	**QQP** accuracy	**SST-2** accuracy
FLAVA **one pretrained model** shared between tasks	75.5	**72.8**	**79.0**	**76.7**	87.3	80.3	90.4	90.9
UniT one model shared between tasks	-	67.0	73.1	-	88.0	80.9	90.6	89.3
VisualBERT (Li et. al.) separately fine-tuned on each task	-	70.8	77.3	74.1	87.0	81.6	89.4	89.4
CLIP (Radford et. al.)	**80.2**	55.3	73.5	56.6	50.5	33.5	76.8	88.2
BERT (Devlin et. al.) separately fine-tuned on each task	-	-	-	-	**91.0**	**84.4**	**90.6**	**92.4**

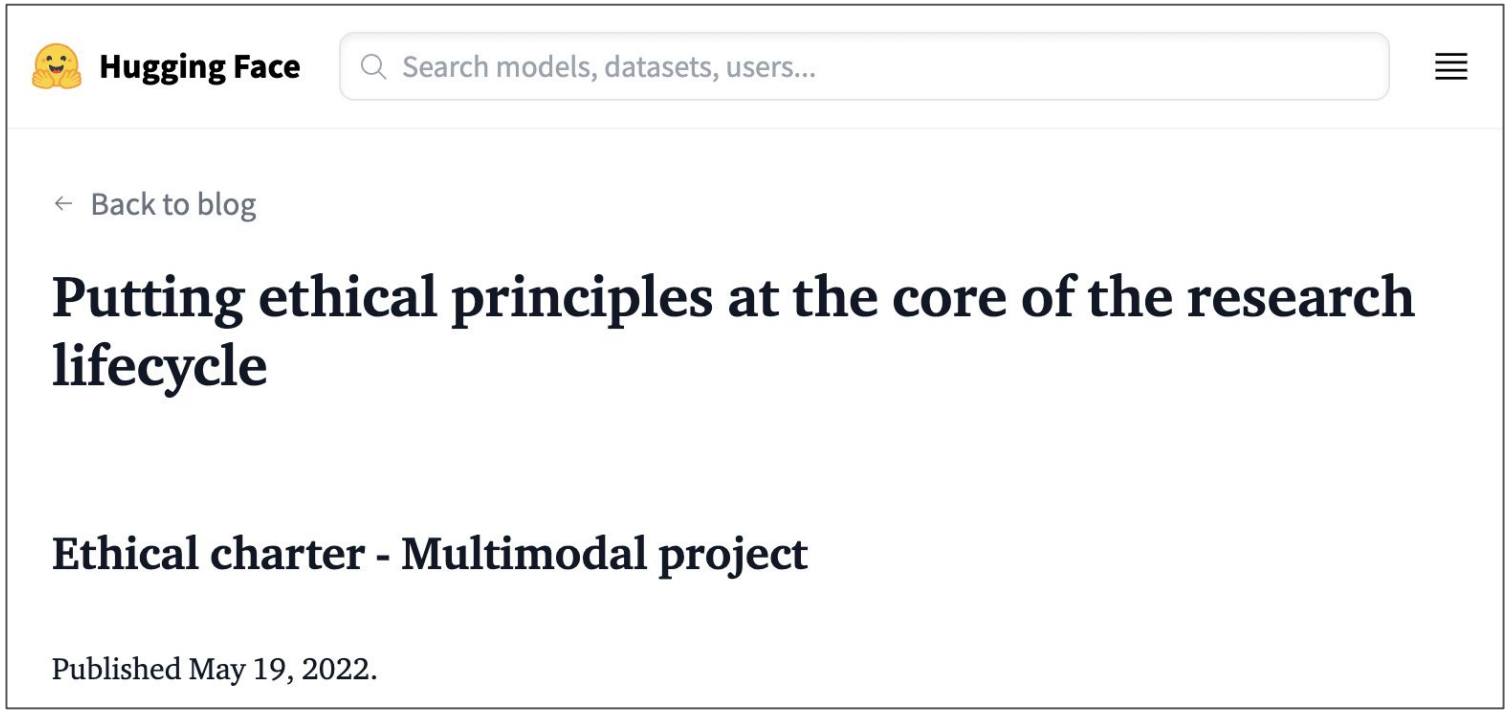
How about closing the loop?

- Results for FLAVA on Winoground (WG) and AdVQA:

	WG-Text	WG-Image	WG-Group	AdVQA
Best	**37.75**	**17.75**	**14.50**	**33.67**
FLAVA	**32.25**	**20.50**	**14.25**	**36.02**

- There is more work to be done!

Ethics first, more coming soon!



https://huggingface.co/blog/ethical-charter-multimodal

In the meantime: Of course it’s on Hugging Face

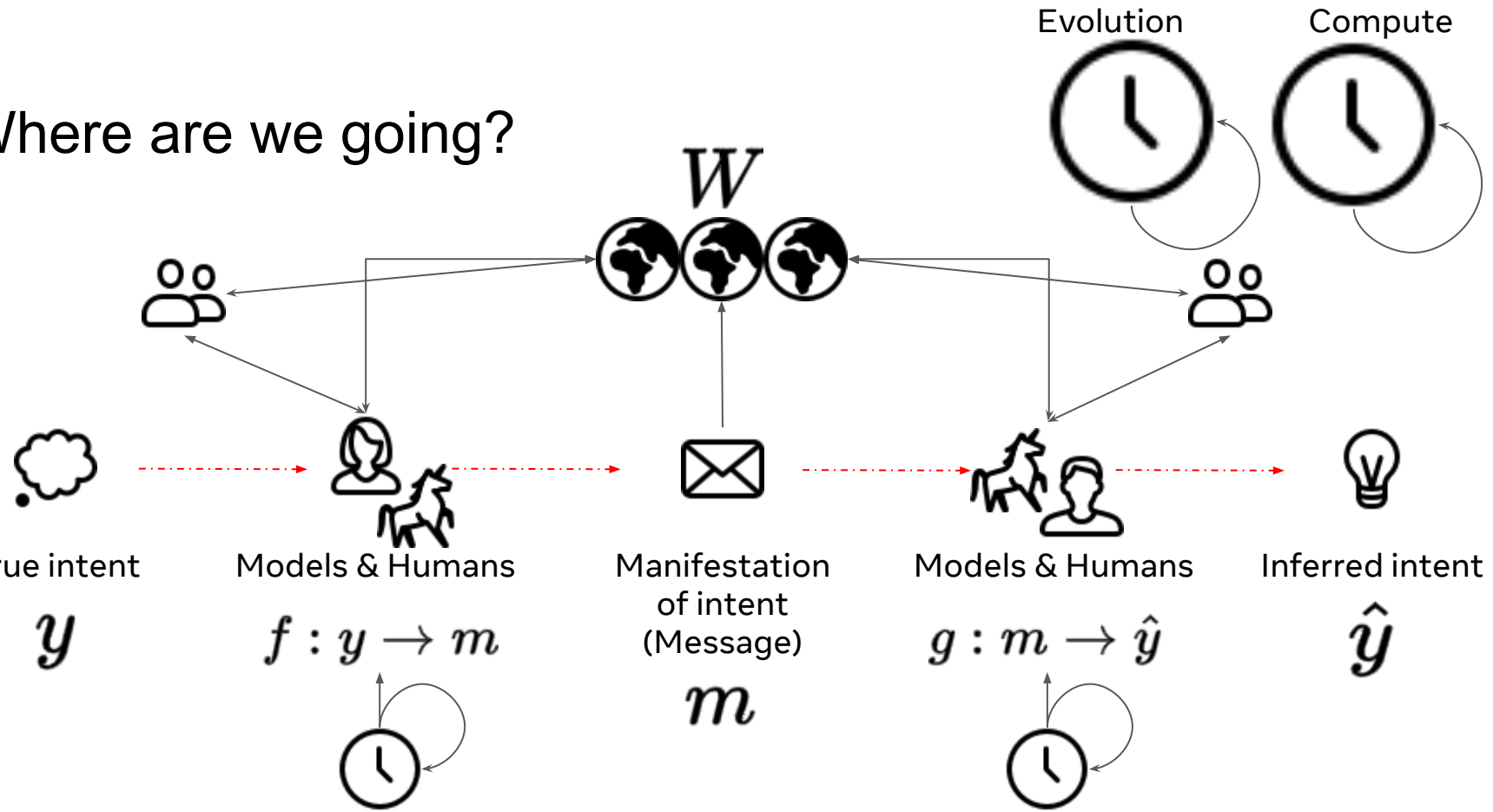
Datasets:

- https://huggingface.co/datasets/facebook/pmd
- https://huggingface.co/datasets/facebook/winoground

FLAVA:

- https://huggingface.co/facebook/flava-full
- https://huggingface.co/flava

Where are we going?



Measuring performance not in the average case, but in the **worst case**.

Things to work on in multimodal AI

What I think will become most useful in the next few years:

- Data-centric methods (filtering, selection, weighting)
- Joint optimization of modalities
- Responsible AI and safety
- Interactive learning / self-play
- Online learning
- Retrieval augmentation
- Further scaling

Thanks!

Thank you for listening!

Thank you to all my collaborators (at Cambridge, Meta/FAIR, Hugging Face and other places) on these projects!

Follow me on Twitter: @douwekiela