Robustness and Compositionality in ML with Insights from Cognitive Bottlenecks

PhD thesis defense of Ankit Vani

Collaborators


Aaron Courville


Aishwarya Agrawal


Alessandro Sordoni


Arian Hosseini


Bac Nguyen


Christos Tsirigotis


Dzmitry Bahdanau


Eeshan Dhekane


Frederick Tung


Gabriel L. Oliveira


Hattie Zhou


Hossein Sharifi-Noghabi


Hugo Larochelle


Jose Gallego


Max Schwarzer


Michael Noukhovitch


Ranjay Krishna


Samuel Lavoie


Sarvjeet Singh Ghotra


Simon Lacoste-Julien


Yuchen Lu

Publications

GAIT: A Geometric Approach to Information Theory.
Jose Gallego, Ankit Vani, Max Schwarzer, Simon Lacoste-Julien. AISTATS 2020.

Iterated Learning for Emergent Systematicity in VQA.
Ankit Vani, Max Schwarzer, Yuchen Lu, Eeshan Dhekane, Aaron Courville. ICLR 2021.

Fortuitous Forgetting in Connectionist Networks.
Hattie Zhou, Ankit Vani, Hugo Larochelle, Aaron Courville. ICLR 2022.

On the Compositional Generalization Gap of In-Context Learning.
Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni, Aaron Courville. BlackboxNLP Workshop 2022.

Simplicial Embeddings in Self-Supervised Learning and Downstream Classification.
Samuel Lavoie, Christos Tsirigotis, Max Schwarzer, Ankit Vani, Michael Noukhovitch, Kenji Kawaguchi, Aaron Courville. ICLR 2023.

Forget Sharpness: Perturbed Forgetting of Model Biases Within SAM Dynamics.
Ankit Vani, Frederick Tung, Gabriel L. Oliveira, Hossein Sharifi-Noghabi. ICML 2024.

SPARO: Selective Attention for Robust and Compositional Transformer Encodings for Vision.
Ankit Vani, Bac Nguyen, Samuel Lavoie, Ranjay Krishna, Aaron Courville. ECCV 2024.

Publications

GAIT: A Geometric Approach to Information Theory.
Jose Gallego, Ankit Vani, Max Schwarzer, Simon Lacoste-Julien. AISTATS 2020.

Iterated Learning for Emergent Systematicity in VQA.
Ankit Vani, Max Schwarzer, Yuchen Lu, Eeshan Dhekane, Aaron Courville. ICLR 2021. **Article 1**

Fortuitous Forgetting in Connectionist Networks.
Hattie Zhou, Ankit Vani, Hugo Larochelle, Aaron Courville. ICLR 2022.

On the Compositional Generalization Gap of In-Context Learning.
Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni, Aaron Courville. BlackboxNLP Workshop 2022.

Simplicial Embeddings in Self-Supervised Learning and Downstream Classification.
Samuel Lavoie, Christos Tsirigotis, Max Schwarzer, Ankit Vani, Michael Noukhovitch, Kenji Kawaguchi, Aaron Courville. ICLR 2023.

Forget Sharpness: Perturbed Forgetting of Model Biases Within SAM Dynamics.
Ankit Vani, Frederick Tung, Gabriel L. Oliveira, Hossein Sharifi-Noghabi. ICML 2024. **Article 2**

SPARO: Selective Attention for Robust and Compositional Transformer Encodings for Vision.
Ankit Vani, Bac Nguyen, Samuel Lavoie, Ranjay Krishna, Aaron Courville. ECCV 2024. **Article 3**

Introduction

Humans are pretty good at *Systematic Generalization*

Humans are pretty good at *Systematic Generalization*

- If you know what “a cat chases a mouse” means, no trouble understanding what “a Roomba chases an elephant” means
  - Even if you have never encountered the phrase before
- Systematic generalization[1] is the consistent application of learned rules to unseen situations

[1] Fodor and Pylyshyn. *Connectionism and cognitive architecture: A critical analysis.* Cognition, 1988.

Humans are pretty good at *Systematic Generalization*

- If you know what “a cat chases a mouse” means, no trouble understanding what “a Roomba chases an elephant” means
  - Even if you have never encountered the phrase before
- Systematic generalization[1] is the consistent application of learned rules to unseen situations

Humans are pretty good at *Compositional Generalization*

- If you understand “red sphere” and “blue cube,” you understand “red cube”

[1] Fodor and Pylyshyn. *Connectionism and cognitive architecture: A critical analysis.* Cognition, 1988.

But machine learning models can be pretty bad at behaving systematically and understanding compositionality

But machine learning models can be pretty bad at behaving systematically and understanding compositionality

- **Standard i.i.d. assumption in machine learning**: Test data comes from the same distribution as the training data
  - Possible catastrophic drops in performance when violated
- For a model capable of systematic generalization
  - Instead of performance depending on divergence between training and test distributions,
  - Performance depends on the discrepancy in the mechanistic processes generating the training and testing observations

How a model behaves on samples outside the training distribution depends on its *inductive biases*

How a model behaves on samples outside the training distribution depends on its *inductive biases*

- Inductive biases: assumptions by a model to generalize to unseen data
  - Shaped by network architecture, training objective, regularization, optimization, *etc.*

How a model behaves on samples outside the training distribution depends on its *inductive biases*

- Inductive biases: assumptions by a model to generalize to unseen data
  - Shaped by network architecture, training objective, regularization, optimization, *etc.*

What are the inductive biases behind human learning that allows us to systematically generalize?

Cognitive bottlenecks

- Overwhelming abundance of stimuli around us, and human mental resources are limited

- Our cognitive processes impose bottlenecks for what information is stored, retained, and consciously attended
  - The inductive biases imparted by these bottlenecks are crucial for systematicity

This thesis explores how inductive biases in machine learning can be informed and developed through insights from cognitive bottlenecks behind human learning and knowledge representation

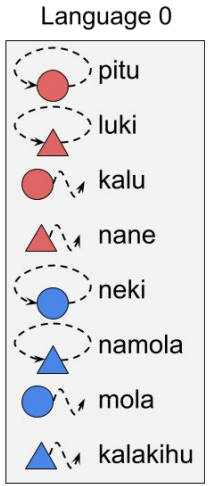
Cognitive bottlenecks we will discuss:

1. Iterated learning
2. Forgetting and relearning
3. Selective attention

1

Iterated Learning

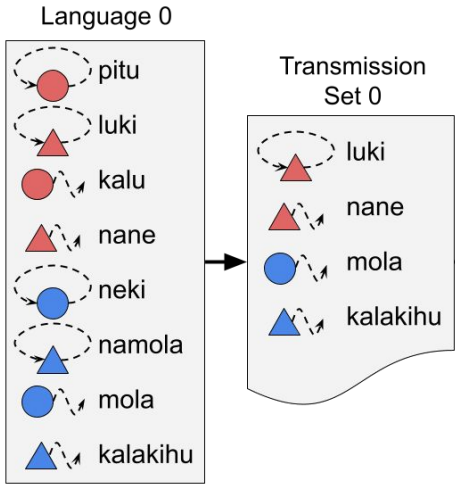
Iterated learning with humans



(Based on [1,2])

[1] Kirby et al. *Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language.* National Academy of Sciences (2008).
[2] Kirby et al. *Iterated learning and the evolution of language.* Current opinion in neurobiology 28 (2014).

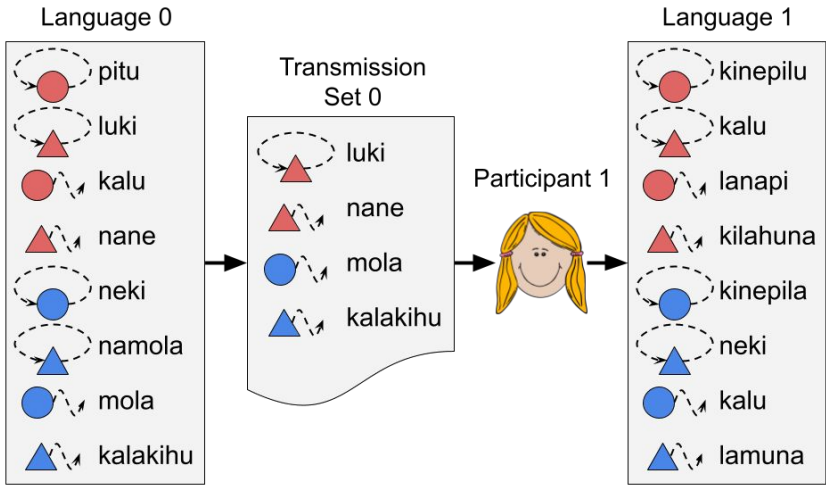
Iterated learning with humans



(Based on [1,2])

[1] Kirby et al. *Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language.* National Academy of Sciences (2008).
[2] Kirby et al. *Iterated learning and the evolution of language.* Current opinion in neurobiology 28 (2014).

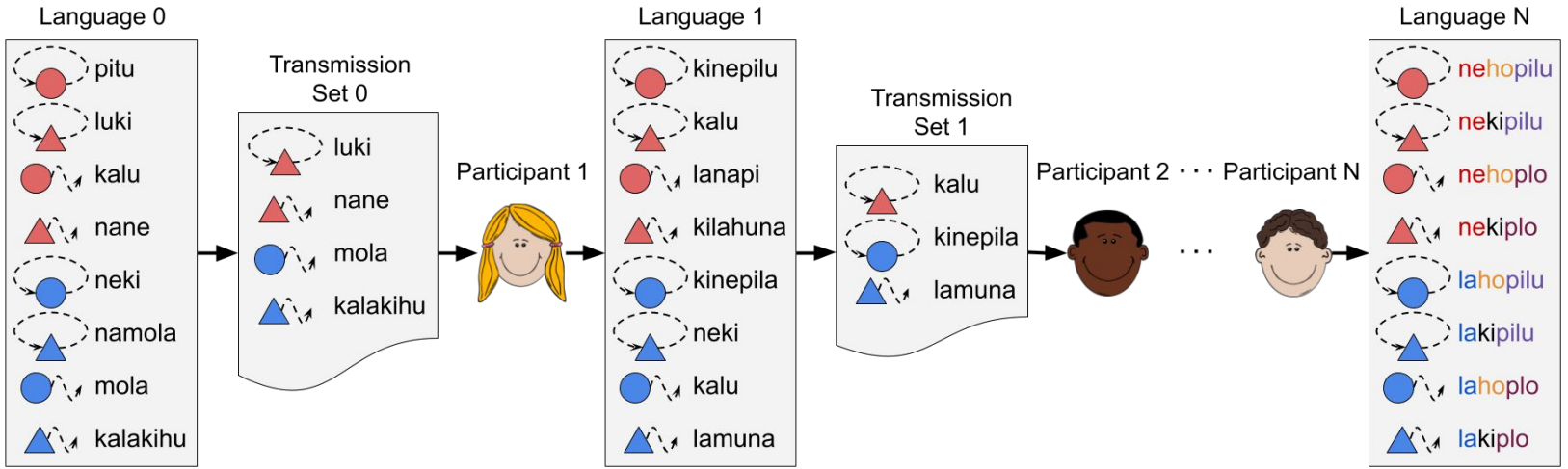
Iterated learning with humans



(Based on [1,2])

[1] Kirby et al. *Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language.* National Academy of Sciences (2008).
[2] Kirby et al. *Iterated learning and the evolution of language.* Current opinion in neurobiology 28 (2014).

Iterated learning with humans



(Based on [1,2])

[1] Kirby et al. *Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language.* National Academy of Sciences (2008).
[2] Kirby et al. *Iterated learning and the evolution of language.* Current opinion in neurobiology 28 (2014).

Learning bottleneck

- Need to learn an **highly expressive** language
  - Through **limited supervision**
- Language properties likely to pass through become universal
  - Compositional rules are more likely to survive the transmission [1]
    - Easier to learn
    - Faster to learn

[1] Kirby. *Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity.* IEEE Transactions on Evolutionary Computation (2001).

Article 1

Iterated Learning for Emergent Systematicity in VQA


Ankit Vani


Max Schwarzer


Yuchen Lu


Eeshan Dhekane


Aaron Courville

International Conference on Learning Representations (ICLR), 2021

Iterated learning in machine learning

- In 2020, iterated learning (IL) in ML stayed close to its Cog Sci roots
  - Most of the work involved agents playing very simplistic referential games [1,2,3,4,5]

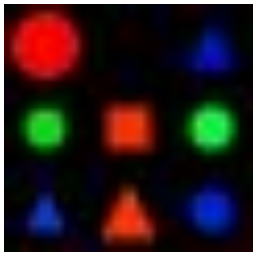
[1] Li and Bowling. *Ease-of-teaching and language structure from emergent communication.* NeurIPS 2019.
[2] Ren et al. *Compositional languages emerge in a neural iterated learning model.* ICLR 2020.
[3] Dagan et al. *Co-evolution of language and agents in referential games.* EACL 2021.
[4] Cogswell et al. *Emergence of compositional language with deep generational transmission.* arXiv:1904.09067 (2019).
[5] Guo et al. *The emergence of compositional languages for numeric concepts through iterated learning in neural agents.* arXiv:1910.05291 (2019).
[6] Zheng et al. *Iterated learning improves compositionality in large vision-language models.* CVPR 2024.

Iterated learning in machine learning

- In 2020, iterated learning (IL) in ML stayed close to its Cog Sci roots
  - Most of the work involved agents playing very simplistic referential games [1,2,3,4,5]
- Claim: Learning bottleneck is a fundamental way to recover structure
- We demonstrated broader applicability through the more complex task of visual question-answering (VQA)
  - Since then, IL has been applied to improve compositional generalization in larger and real-world settings, *e.g.*, training vision-language models like CLIP [6]

[1] Li and Bowling. *Ease-of-teaching and language structure from emergent communication.* NeurIPS 2019.
[2] Ren et al. *Compositional languages emerge in a neural iterated learning model.* ICLR 2020.
[3] Dagan et al. *Co-evolution of language and agents in referential games.* EACL 2021.
[4] Cogswell et al. *Emergence of compositional language with deep generational transmission.* arXiv:1904.09067 (2019).
[5] Guo et al. *The emergence of compositional languages for numeric concepts through iterated learning in neural agents.* arXiv:1910.05291 (2019).
[6] Zheng et al. *Iterated learning improves compositionality in large vision-language models.* CVPR 2024.

Visual question-answering (VQA)

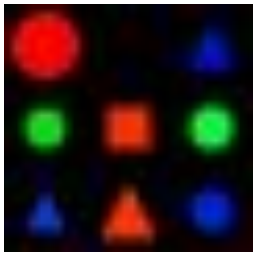


(Image taken from SHAPES [1])

Q: Is a green shape left of a square?

[1] Andreas et al. *Neural module networks*. CVPR 2016.

Visual question-answering (VQA)

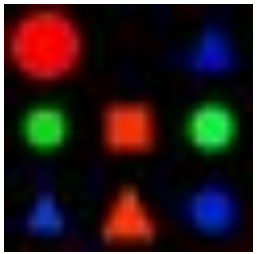


(Image taken from SHAPES [1])

Q: Is a green shape left of a square? **A:** Yes

[1] Andreas et al. *Neural module networks*. CVPR 2016.

Visual question-answering (VQA)



(Image taken from SHAPES [1])

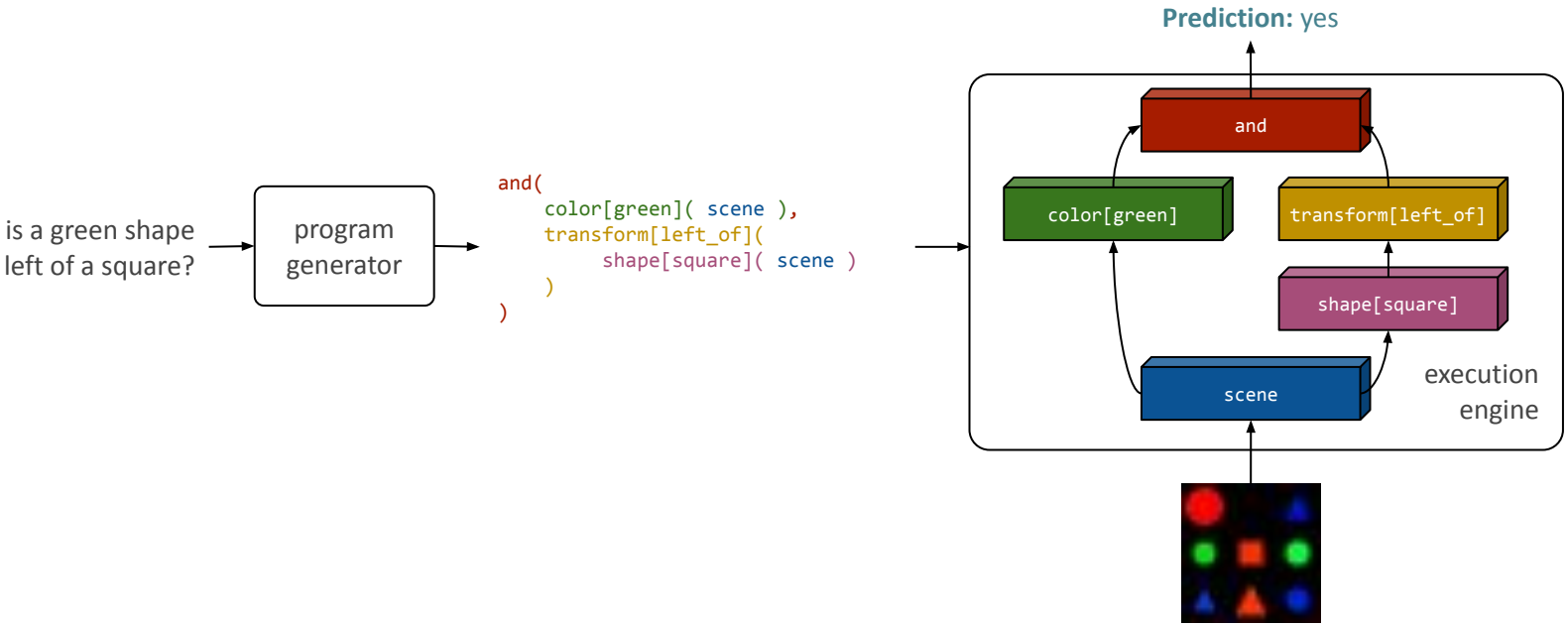
Q: Is a green shape left of a square? **A:** Yes

Program:

```
and(
    color[green]( scene ),
    transform[left_of](
        shape[square]( scene )
    )
)
```

[1] Andreas et al. *Neural module networks*. CVPR 2016.

Neural module networks (NMNs)[1,2]



[1] Andreas et al. *Neural module networks.* CVPR 2016.
[2] Johnson et al. *Inferring and executing programs for visual reasoning.* CVPR 2017.

Core idea

- With the right layout, NMNs exhibit compositionality [1]; other methods fail
- **But**: The right layout does not emerge naturally
  - Bahdanau et al., 2019[1] needed to provide correct tree-structured layouts
  - Learned layouts only converged to be robust under a strong prior for the correct structure

[1] Bahdanau et al. *Systematic generalization: what is required and can it be learned?* ICLR 2019.

Core idea

- With the right layout, NMNs exhibit compositionality [1]; other methods fail
- **But**: The right layout does not emerge naturally
  - Bahdanau et al., 2019[1] needed to provide correct tree-structured layouts
  - Learned layouts only converged to be robust under a strong prior for the correct structure

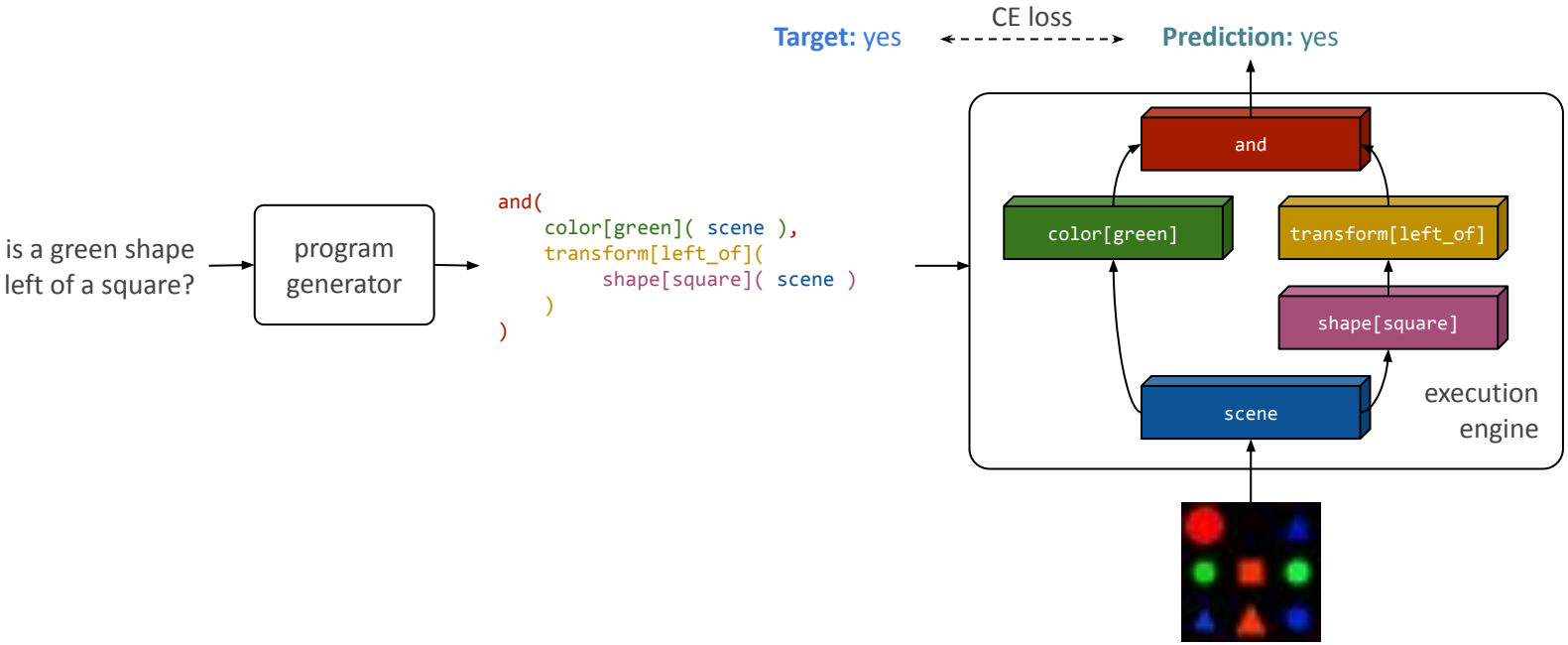
Use IL to encourage structured layouts towards systematic generalization

[1] Bahdanau et al. *Systematic generalization: what is required and can it be learned?* ICLR 2019.

Article 1

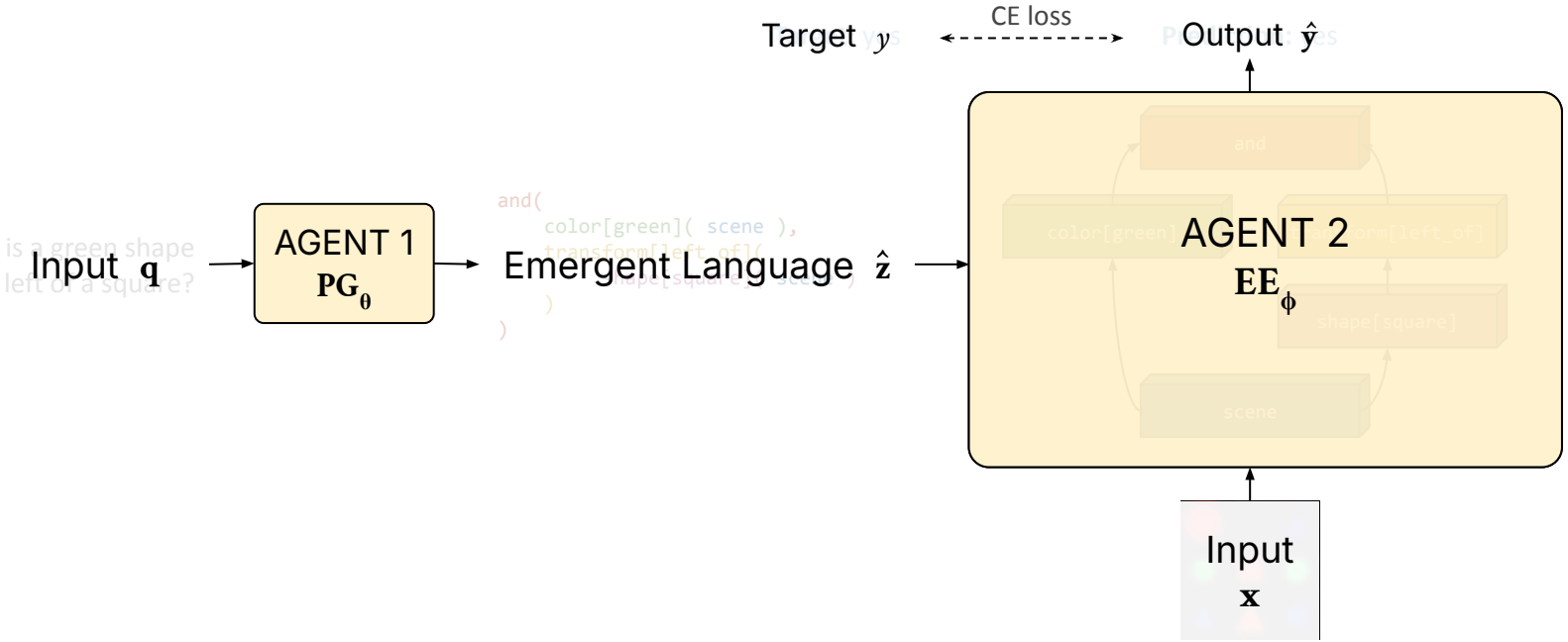
Method

Standard NMN training



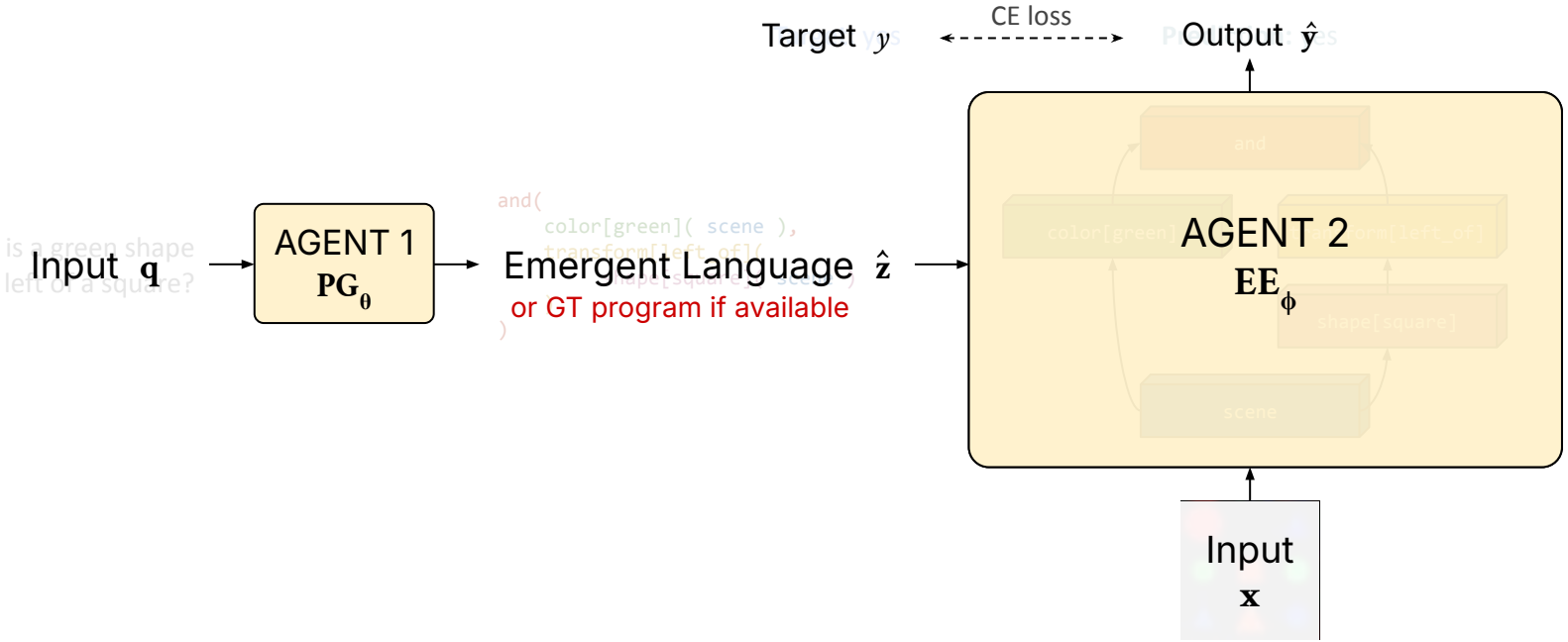
Standard NMN training

Viewed as two agents with an emergent language to solve the task



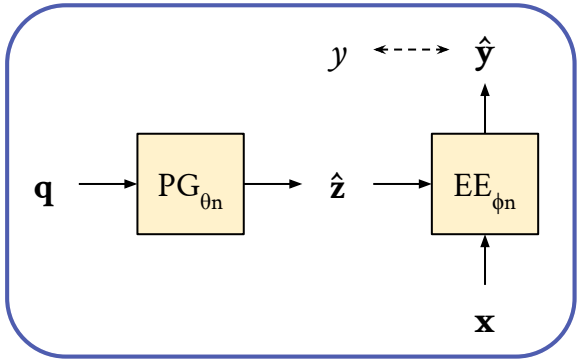
Standard NMN training

Viewed as two agents with an emergent language to solve the task



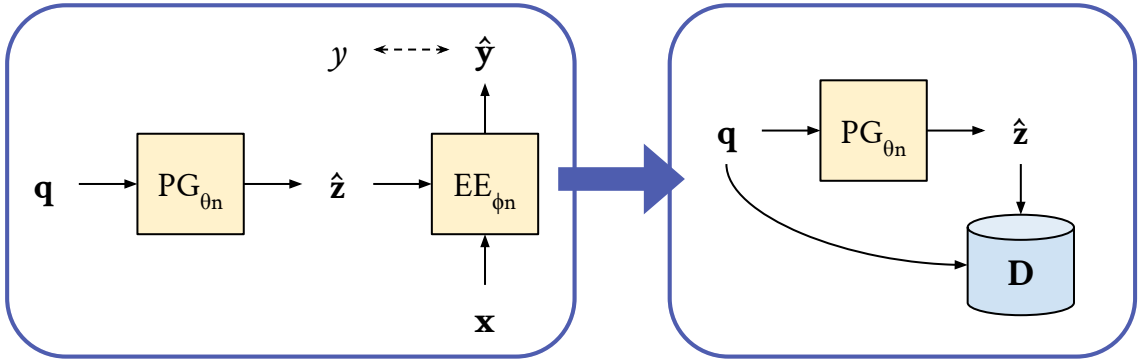
Iterated learning for NMNs

Interacting phase



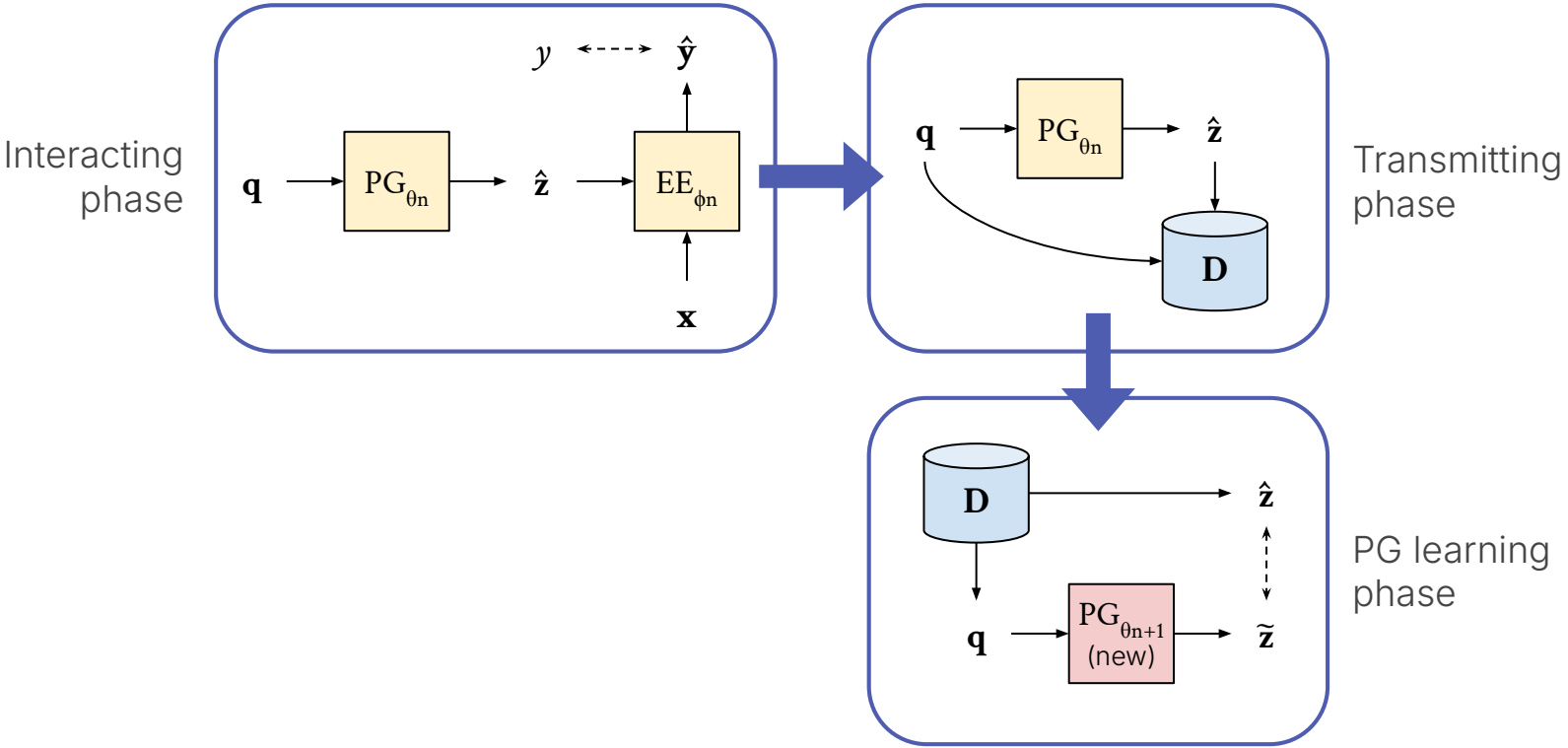
Iterated learning for NMNs

Interacting phase

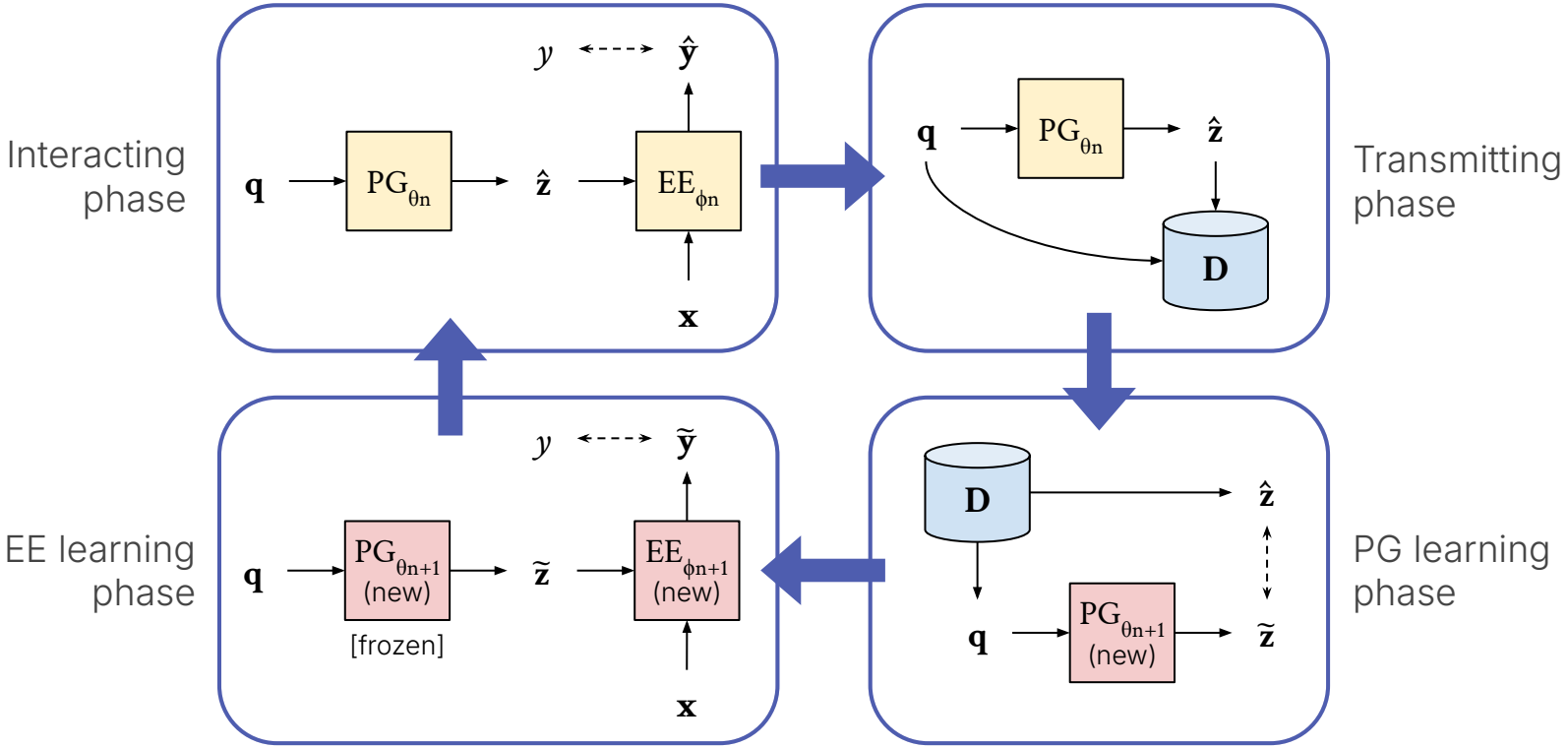


Transmitting phase

Iterated learning for NMNs



Iterated learning for NMNs



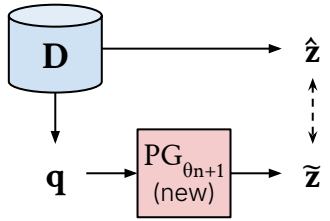
Learning bottleneck

Limit the length of the learning phases: Early stopping

- Generally a sweet spot for the number of steps

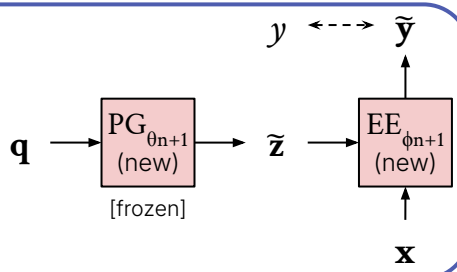
Program generator learning phase steps

- *Too few:* Low confidence in utterances, high variance
- *Too many:* Overfitting to transmitted data



Execution engine learning phase steps

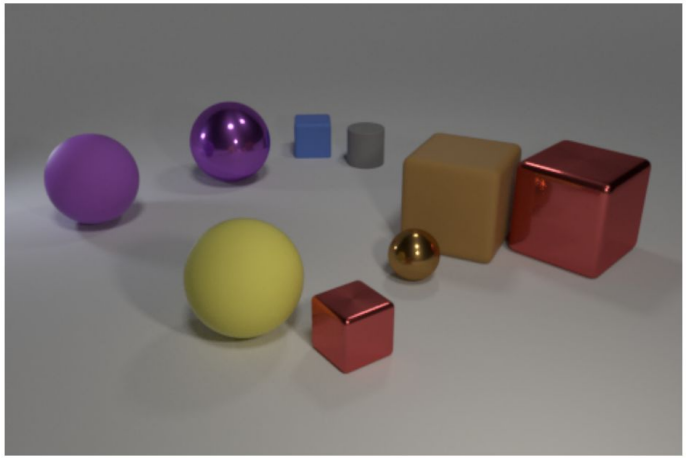
- *Too few:* High variance gradients at the start of interacting phase
- *Too many:* Overfitting to an imperfect program generator



Article 1

Results with CLEVR

CLEVR/CLOSURE example



[1] **Q1 (CLEVR):** There is another cube that is the same size as the brown cube; what is its color?

[1] **Q2 (CLEVR):** There is a thing that is in front of the yellow thing; does it have the same color as cylinder?

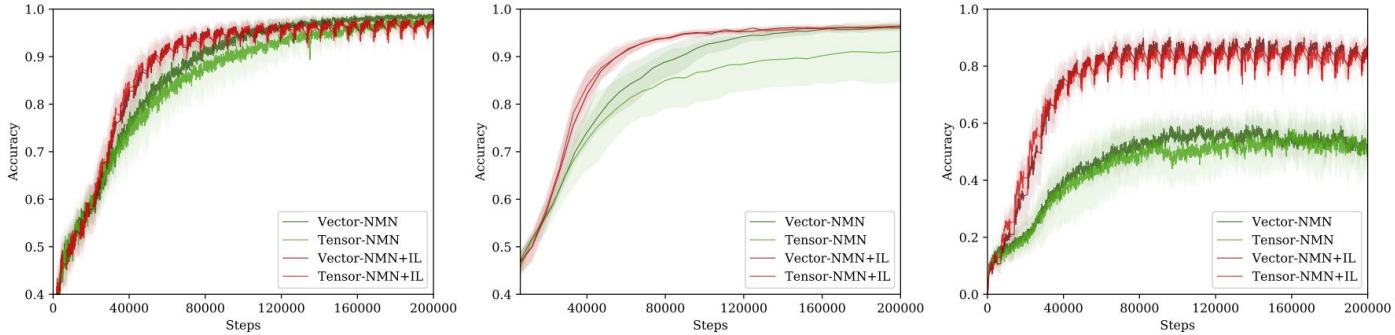
[2] **Q3 (CLOSURE):** There is another rubber object that is the same size as the gray cylinder; does it have the same color as the tiny shiny block?

(Figure taken from [2])

[1] Johnson et al. *CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning.* CVPR 2017.

[2] Bahdanau et al. *CLOSURE: Assessing Systematic Generalization of CLEVR Models.* arXiv:1912.05783 (2019).

CLEVR with 100 GT programs



(a) Task training accuracy. (b) Task validation accuracy. (c) Program accuracy.

Tensor-NMN: modules are residual blocks[1] with separate parameters, from [2]
Vector-NMN: modules share parameters with separate FiLM[3] adapters, from [4]

- All models reach similar training accuracies
- IL leads to significantly higher program accuracy
- The generalization difference is more apparent in the OOD CLOSURE dataset

[1] He et al. *Deep residual learning for image recognition.* CVPR 2016.
[2] Johnson et al. *Inferring and executing programs for visual reasoning.* CVPR 2017.
[3] Perez et al. *FiLM: Visual Reasoning with a General Conditioning Layer.* AAAI 2018.
[4] Bahdanau et al. *CLOSURE: Assessing Systematic Generalization of CLEVR Models.* arXiv:1912.05783 (2019).

CLOSURE categories: 100 GT programs

Evaluation set	Tensor-NMN		Vector-NMN	
	Without IL	With IL	Without IL	With IL
CLEVR-Val	0.912 ± 0.07	$\mathbf{0.964 \pm 0.01}$	0.960 ± 0.01	$\mathbf{0.964 \pm 0.00}$
and_mat_spa	$\mathbf{0.278 \pm 0.17}$	0.264 ± 0.16	$\mathbf{0.400 \pm 0.13}$	0.335 ± 0.18
or_mat	0.327 ± 0.11	$\mathbf{0.481 \pm 0.24}$	0.367 ± 0.11	$\mathbf{0.563 \pm 0.23}$
or_mat_spa	0.286 ± 0.13	$\mathbf{0.405 \pm 0.22}$	0.330 ± 0.11	$\mathbf{0.444 \pm 0.24}$
compare_mat	0.793 ± 0.11	$\mathbf{0.851 \pm 0.17}$	0.660 ± 0.16	$\mathbf{0.873 \pm 0.12}$
compare_mat_spa	0.746 ± 0.13	$\mathbf{0.853 \pm 0.15}$	0.677 ± 0.14	$\mathbf{0.871 \pm 0.12}$
embed_spa_mat	0.824 ± 0.07	$\mathbf{0.947 \pm 0.03}$	0.863 ± 0.07	$\mathbf{0.900 \pm 0.08}$
embed_mat_spa	0.739 ± 0.14	$\mathbf{0.941 \pm 0.02}$	0.894 ± 0.03	$\mathbf{0.936 \pm 0.03}$

- IL improves performance on all but one CLOSURE category
- Tensor-NMN with IL leads to CLEVR accuracy similar to previous works with far fewer programs
  - 18,000 for Johnson et al., 2017[1], 1,000 for Vedantam et al., 2019[2]

[1] Johnson et al. *Inferring and executing programs for visual reasoning.* CVPR 2017.
[2] Vedantam et al. *Probabilistic neural symbolic models for interpretable visual question answering.* ICML 2019.

Article 1

Conclusion

Key takeaways

Iterated learning is more broadly applicable where hard-to-learn compositional latents are desired, beyond simple referential games

Iterated learning amplifies compositionality when a preference for it exists

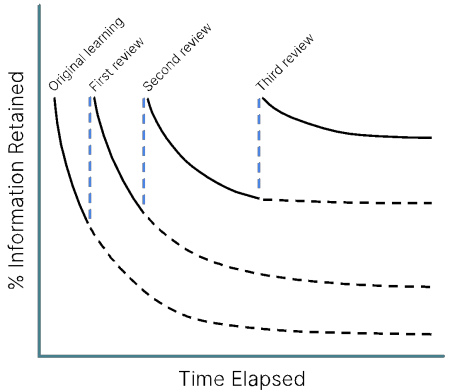
- In addition to model architecture, this bias can be imparted through few-shot supervision
- Emergent compositional structure can improve downstream systematic generalization
    - We find similar gains in simpler (SHAPES) and larger real-world image (GQA) settings

2

Forgetting and Relearning

Forgetting and relearning in humans

- Forgetting can be frustrating, but it is beneficial
- Spacing effect[1]
  - More effective learning with spaced-out study sessions
- Forgetting allows integration of new perspectives and knowledge unburdened by biases of the past

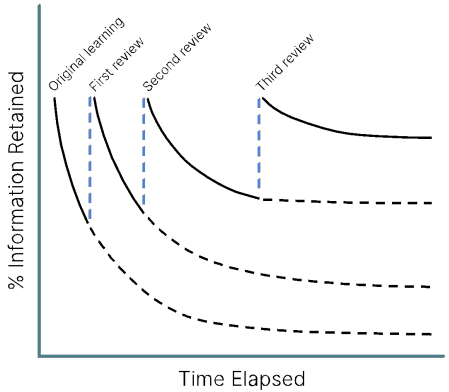


(Figure based on [1])

[1] Ebbinghaus. *Memory: A Contribution to Experimental Psychology.* Duncker & Humblot, 1885.

Forgetting and relearning in humans

- Forgetting can be frustrating, but it is beneficial
- Spacing effect[1]
  - More effective learning with spaced-out study sessions
- Forgetting allows integration of new perspectives and knowledge unburdened by biases of the past



(Figure based on [1])

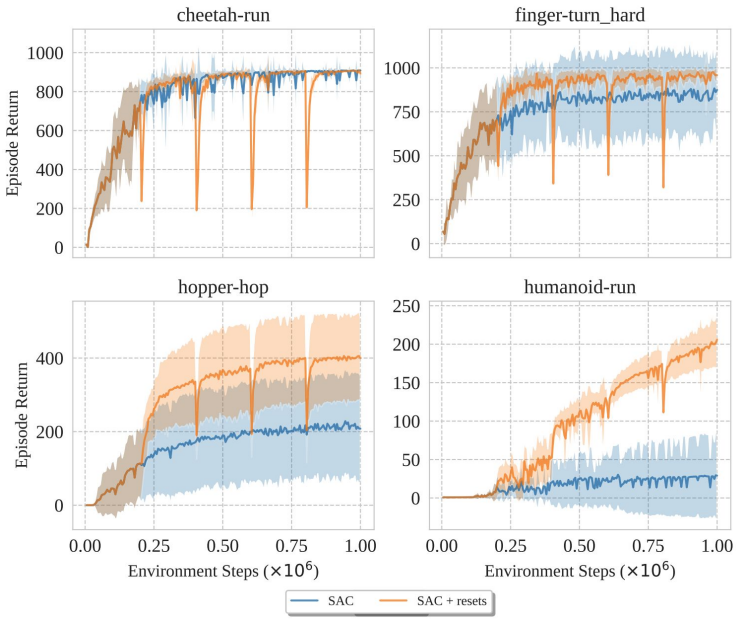
Under continuous forgetting and relearning:

Less relevant and specialized information fades away, while more pertinent and structurally coherent knowledge is retained

[1] Ebbinghaus. *Memory: A Contribution to Experimental Psychology.* Duncker & Humblot, 1885.

Forgetting and relearning in ML

- The forget-and-relearn[1] paradigm typically partially resets weights to “forget” before continuing training for “relearning”
- Iterated learning can be seen as an instance of forget-and-relearn
- Iterative resetting counters the primacy bias in reinforcement learning [2]



(Figure taken from [2])

[1] Zhou et al. *Fortuitous forgetting in connectionist networks.* ICLR 2022.
[2] Nikishin et al. *The primacy bias in deep reinforcement learning.* ICML 2022.

Article 2

Forget Sharpness:
Perturbed Forgetting of Model Biases Within SAM Dynamics


Ankit Vani


Frederick Tung


Gabriel Oliveira


Hossein Sharifi-Noghabi

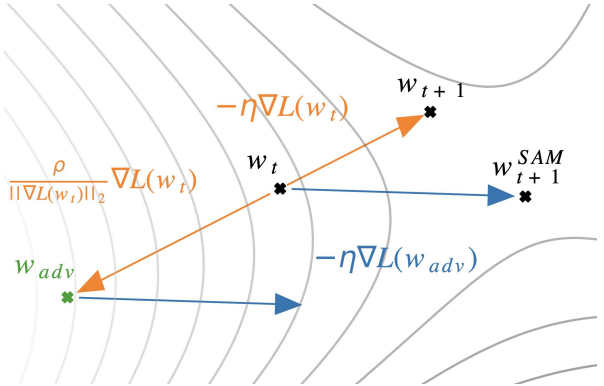
International Conference on Machine Learning (ICML), 2024

Sharpness-Aware Minimization (SAM)

Assumption: Flatter loss surface regions generalize better

SAM[1] is motivated to guide optimization to flatter regions of the loss surface

- Reduce sharpness by approximately minimizing a PAC-Bayes upper bound
- Algorithm:
  - Perturb by maximizing loss
  - Compute minimization gradient
  - Step from unperturbed weights



(Figure taken from [1])

[1] Foret et al. *Sharpness-aware minimization for efficiently improving generalization*. ICLR 2021.

Unresolved concerns with SAM

Improved generalization with SAM deviates from its motivating theory

- The PAC-Bayes bound does not sufficiently explain SAM's benefits [1]
  - Bound derived for random perturbations, which perform worse in practice
  - Steepest ascent perturbations loosen the bound but perform better
  - ***m*-Sharpness:** Why do smaller perturbing batches perform better?
- Empirically, flatness and generalization do not correlate in general [2]

[1] Andriushchenko et al. *Towards understanding sharpness-aware minimization.* ICML 2022.
[2] Andriushchenko et al. *A modern look at the relationship between sharpness and generalization.* ICML 2023.

Article 2

Method

“Perturbed forgetting” perspective

- SAM dynamics perform forgetting and relearning without erasing learned state
- Perturbations can discard undesirable model biases

Steepest ascent gradients for a small set of *m* samples can reveal undesirable shortcuts the model learned for them

- Perturbing discards them
- Predictions then utilize the global learned structure
- Gradients strengthen this global structure

Information-theoretic argument

Generalization gap bound from [1]: $\Delta(\boldsymbol{\theta}) \in \tilde{O}\left(\sqrt{\frac{I(\boldsymbol{X};\hat{\boldsymbol{Y}} \mid Y) + I(\boldsymbol{\theta};\mathrm{D})}{n}}\right)$

Avoid increasing $I(\boldsymbol{\theta};\mathrm{D})$

- Minimizing the loss makes it easier to identify D as the training dataset
- Perturbing does not minimize loss

Decrease $I(\boldsymbol{X};\hat{\boldsymbol{Y}} \mid Y)$

- Model biases can be exposed via outputs $\hat{\boldsymbol{Y}}$
- Perturbing can target and discard them

[1] Kawaguchi et al. *How does information bottleneck help deep learning?* ICML 2023.

Output bias forgetting (OBF)

Gradient of cross-entropy (CE) with logits z:

$$\nabla_{\boldsymbol{\theta}} L^{\mathrm{CE}}(y, \hat{\boldsymbol{y}}) = \mathbb{E}_{i \sim \hat{\boldsymbol{y}}} \left[\nabla_{\boldsymbol{\theta}} z_i \right] - \nabla_{\boldsymbol{\theta}} z_y$$

Minimizing (for learning) pushes down more on worse predictions 

Maximizing (for perturbing) pulls up more on worse predictions

Forgetting some biases but amplifying others 

Output bias forgetting (OBF)

Gradient of cross-entropy (CE) with logits z:

$$\nabla_{\boldsymbol{\theta}} L^{\mathrm{CE}}(y, \hat{\boldsymbol{y}}) = \mathbb{E}_{i \sim \hat{\boldsymbol{y}}} \left[\nabla_{\boldsymbol{\theta}} z_i \right] - \nabla_{\boldsymbol{\theta}} z_y$$

Minimizing (for learning) pushes down more on worse predictions ✓

Maximizing (for perturbing) pulls up more on worse predictions

Forgetting some biases but amplifying others ✗

The OBF perturbation minimizes target likelihood without bias amplification

$$\nabla_{\boldsymbol{\theta}} L^{\mathrm{BF}}(y, \hat{\boldsymbol{y}}) = \mathbb{E}_{i \sim \mathrm{Uniform}} \left[\nabla_{\boldsymbol{\theta}} z_i \right] - \left(\alpha \mathbb{E}_{i \sim \hat{\boldsymbol{y}}} \left[\nabla_{\boldsymbol{\theta}} z_i \right] + (1 - \alpha) \nabla_{\boldsymbol{\theta}} z_y \right)$$

Like cross-entropy, without bias amplification

$\alpha = 0$ ⟷ $\alpha = 1$

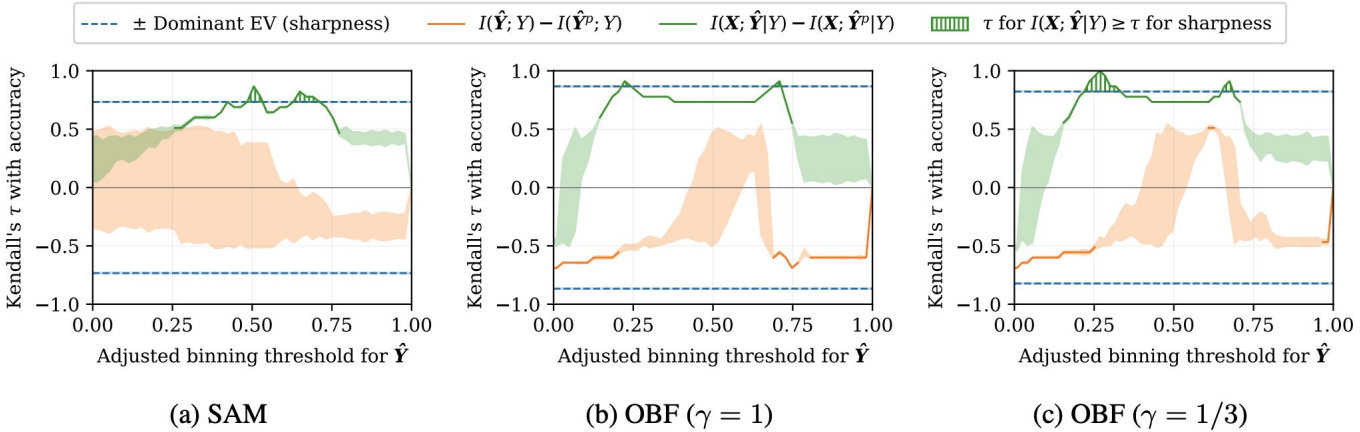
Negative cross-entropy with uniform targets

Article 2

Results

Forgetting *vs.* generalization

- Train ViT-S/32 models on CIFAR-10 with perturbing batch sizes $m \in \{2^k \mid k \in \{0,\ldots,9\}\}$
- Discretize model outputs with varying thresholds



There exist thresholds where generalization correlates more strongly with forgetting than flatness of the loss surface.

OBF *vs.* standard SAM variants

Model	Method	Perturb	ImageNet- V1	ImageNet- Real	ImageNet- V2	ImageNet- R	ImageNet- Sketch	Sharpness
ViT-S/32	AdamW	None	69.29±0.26	75.31±0.28	55.48±0.58	19.02±0.47	16.38±0.34	165.6±15.2
	SAM[1]	Standard	72.77±0.06	78.89±0.05	58.81±0.33	21.63±0.23	19.68±0.50	14.9±1.1
		OBF	74.49±0.04	81.31±0.05	**61.13±0.18**	**25.31±0.41**	**22.58±0.13**	3.9±1.4
	GSAM[2]	Standard	73.41±0.05	79.48±0.08	59.94±0.15	22.18±0.15	20.28±0.15	11.6±1.2
		OBF	74.41±0.12	**81.41±0.11**	**61.08±0.18**	25.15±0.23	22.24±0.07	**3.1±0.7**
	ASAM[3]	Standard	74.45±0.11	81.23±0.11	60.78±0.25	24.07±0.12	21.68±0.23	6.5±0.4
		OBF	**74.73±0.19**	81.24±0.25	60.95±0.28	24.65±0.26	22.40±0.10	30.3±11.6
ResNet-50	SGD	None	76.86±0.07	83.28±0.11	65.00±0.14	20.29±0.36	20.53±0.46	230.4±42.7
	SAM[1]	Standard	77.49±0.06	83.78±0.05	65.26±0.21	21.08±0.16	21.18±0.32	170.1±18.9
		OBF	77.67±0.07	84.01±0.03	65.70±0.45	21.63±0.18	22.17±0.26	164.4±25.0
	GSAM[2]	Standard	77.43±0.12	83.79±0.19	65.37±0.26	21.37±0.21	21.52±0.56	171.0±16.8
		OBF	77.66±0.08	84.09±0.07	66.01±0.09	21.76±0.23	22.26±0.47	161.4±10.9
	ASAM[3]	Standard	77.30±0.02	84.07±0.03	65.55±0.16	21.71±0.02	21.75±0.15	**33.6±2.99**
		OBF	**78.17±0.07**	**84.66±0.05**	**66.55±0.15**	**23.84±0.12**	**24.21±0.42**	39.1±1.28

[1] Foret et al. *Sharpness-aware minimization for efficiently improving generalization.* ICLR 2021.
[2] Zhuang et al. *Surrogate gap minimization improves sharpness-aware training.* ICLR 2022.
[3] Kwon et al. *ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks.* ICML 2021.

Article 2

Conclusion

Key takeaways

Targeted forgetting and relearning can be a powerful driver of generalization

- Gradient-based approaches are effective at targeting undesirable biases
- Perturbed forgetting protects the global model state against suboptimal forgetting steps

SAM's training dynamics are more important than loss surface flatness

- Is the pursuit of flatter minima misleading?

3

Selective Attention

Selective attention in humans

- Human working memory is extremely limited
  - It can can hold at most ~4 “chunks” of information at a time [1]
- *Selective attention* is critical in managing these limited mental resources
  - We prioritize stimuli that is most relevant, and ignore the rest [2]
  - Humans struggle to simultaneously attend to separable features (color, shape, orientation) [3], or novel objects at the same location [4]

[1] Cowan. *The magical number 4 in short-term memory: A reconsideration of mental storage capacity.* Behavioral and brain sciences, 2001.
[2] Treisman. *Strategies and models of selective attention.* Psychological review, 1969.
[3] Treisman and Gelade. *A feature-integration theory of attention.* Cognitive psychology, 1980.
[4] Rock and Gutman. *The effect of inattention on form perception.* Journal of Experimental Psychology: Human Perception and Performance, 1981.

Selective attention in humans

- Human working memory is extremely limited
  - It can can hold at most ~4 “chunks” of information at a time [1]
- *Selective attention* is critical in managing these limited mental resources
  - We prioritize stimuli that is most relevant, and ignore the rest [2]
  - Humans struggle to simultaneously attend to separable features (color, shape, orientation) [3], or novel objects at the same location [4]

A preference for learning separable concepts eases learning compositional representations of our knowledge, which is essential to act effectively with limited cognitive resources in the world

[1] Cowan. *The magical number 4 in short-term memory: A reconsideration of mental storage capacity.* Behavioral and brain sciences, 2001.
[2] Treisman. *Strategies and models of selective attention.* Psychological review, 1969.
[3] Treisman and Gelade. *A feature-integration theory of attention.* Cognitive psychology, 1980.
[4] Rock and Gutman. *The effect of inattention on form perception.* Journal of Experimental Psychology: Human Perception and Performance, 1981.

Selective attention in machine learning

- The attention mechanism[1] is closely related to selective attention
  - Refined and popularized by the Transformer[2] architecture
- In transformers, selective attention is implicitly performed for hierarchical construction of internal representations
- But when explicitly learning representations for future downstream tasks, a prior for separately-attendable concepts is not automatically granted
  - Deprives downstream tasks easy access to task-relevant aspects of the data's structure

[1] Bahdanau et al. *Neural machine translation by jointly learning to align and translate.* ICLR 2015.
[2] Vaswani et al. *Attention is all you need.* NeurIPS 2017.

Article 3

SPARO:
Selective Attention for Robust and Compositional Transformer Encodings for Vision


Ankit Vani


Bac Nguyen

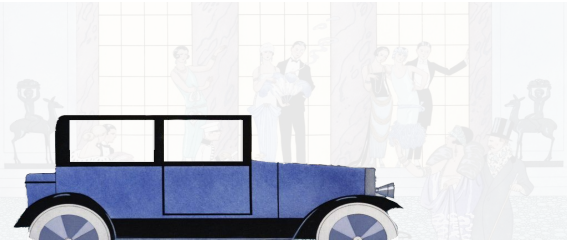

Samuel Lavoie


Aaron Courville

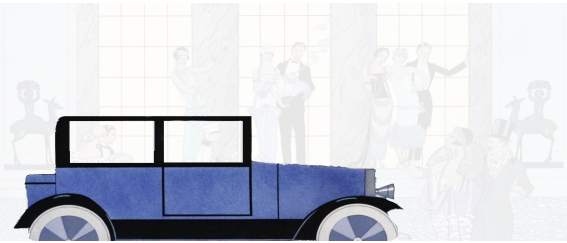

Ranjay Krishna

European Conference on Computer Vision (ECCV), 2024

What is the make and model of the car?

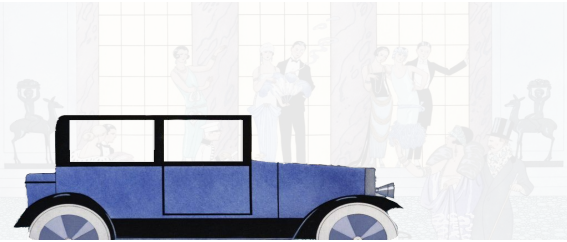


What is the make and model of the car?



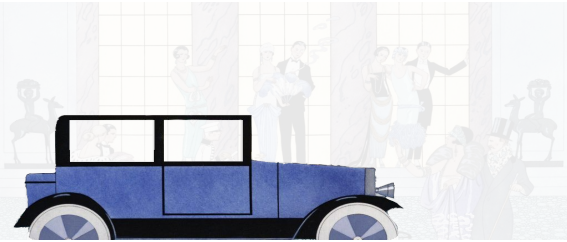
What is the make and model of the car?

Describe the building architecture.



What is the make and model of the car?

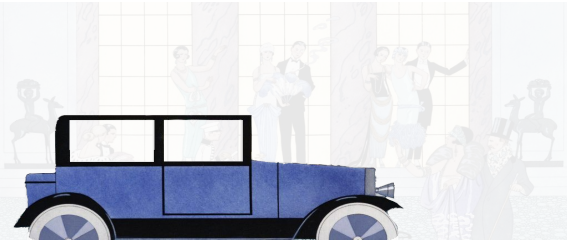
Describe the building architecture.



What is the make and model of the car?

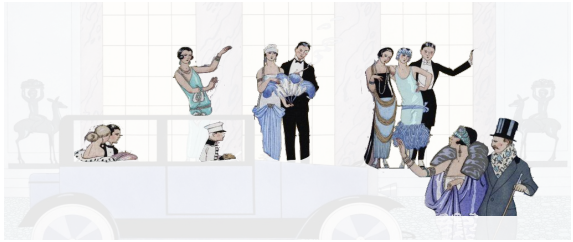
Describe the building architecture.

Are people arriving or leaving?

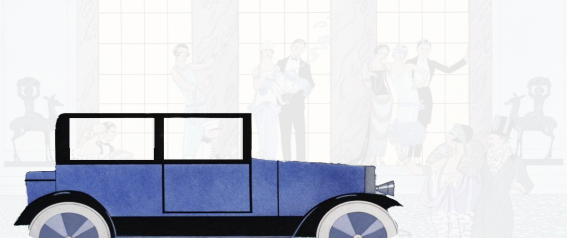


What is the make and model of the car?

Describe the building architecture.



Are people arriving or leaving?

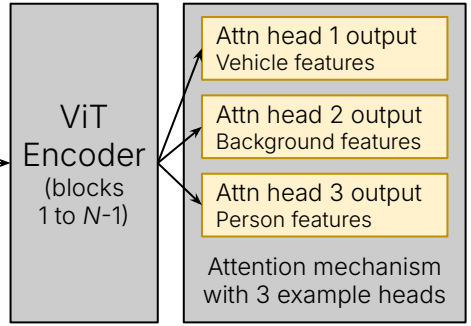


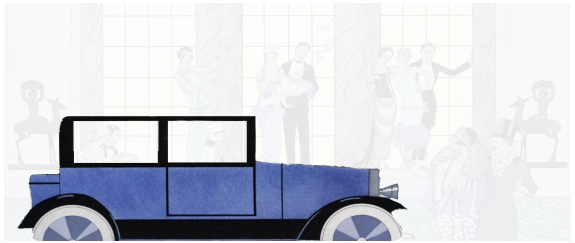
What is the make and model of the car?

Describe the building architecture.

Are people arriving or leaving?

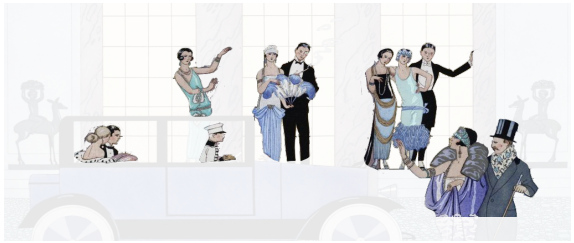
Input image





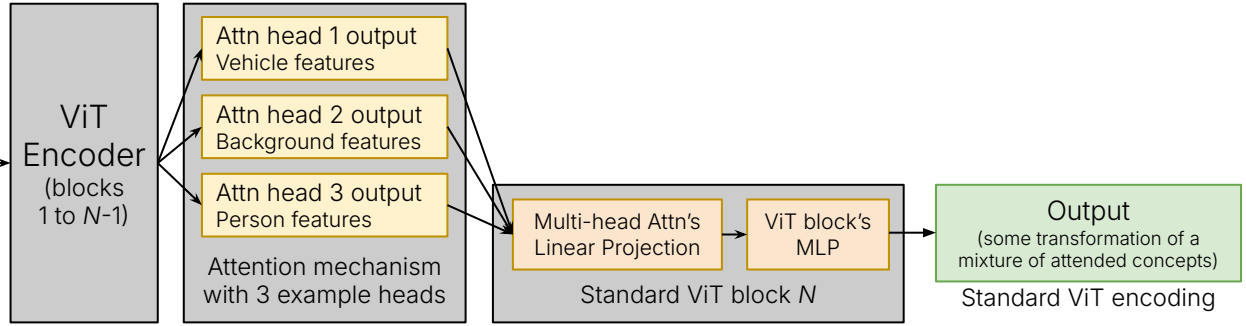
What is the make and model of the car?

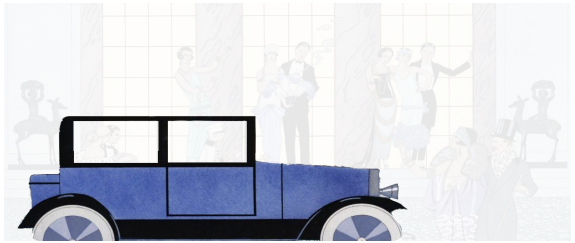
Describe the building architecture.



Are people arriving or leaving?

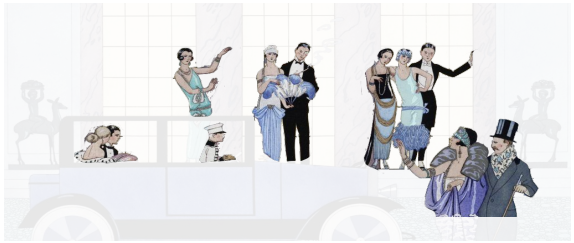
Input image





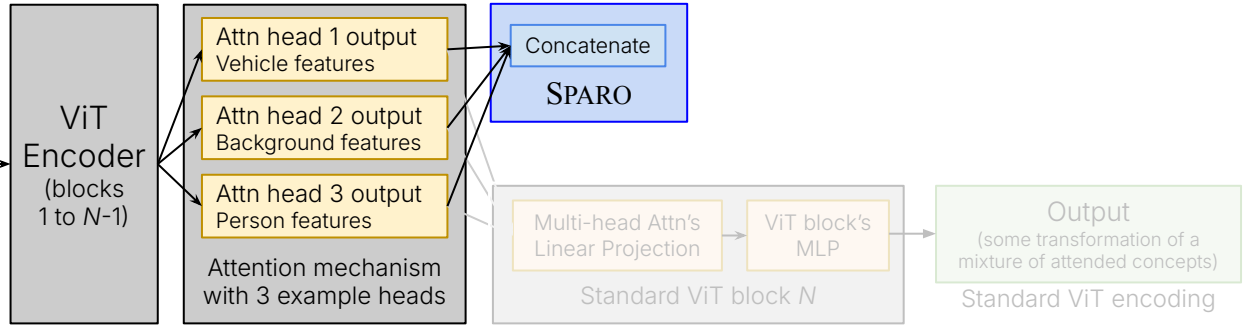
What is the make and model of the car?

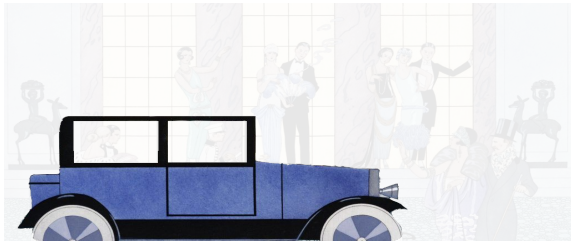
Describe the building architecture.



Are people arriving or leaving?

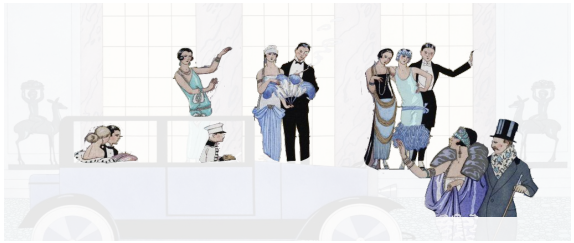
Input image





What is the make and model of the car?

Describe the building architecture.



Are people arriving or leaving?

Input image

ViT Encoder (blocks 1 to *N*-1)

Attn head 1 output
Vehicle features

Attn head 2 output
Background features

Attn head 3 output
Person features

Attention mechanism with 3 example heads

Concatenate

SPARO

Slot 1 (selective attn on vehicles) | Slot 2 (selective attn on background) | Slot 3 (selective attn on persons)

SPARO output encoding

Explicitly structures encodings as attended concepts

Multi-head Attn's Linear Projection → ViT block's MLP

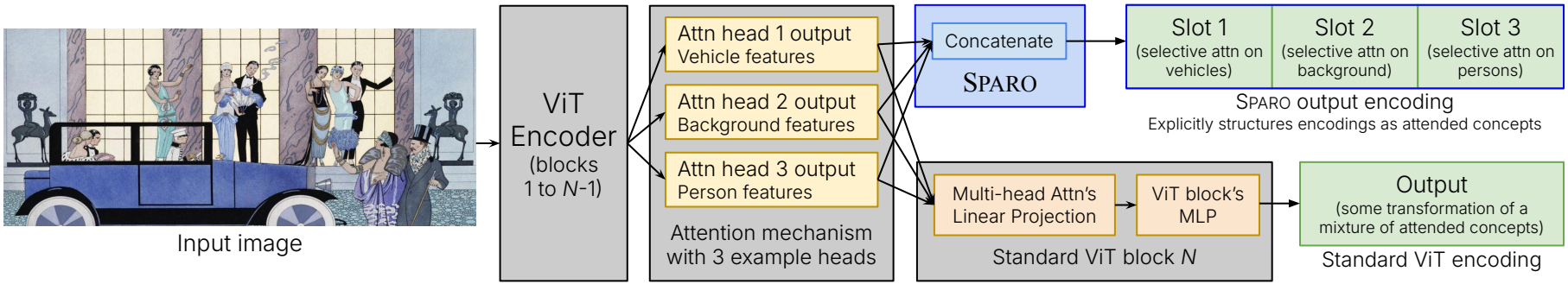
Standard ViT block *N*

Output (some transformation of a mixture of attended concepts)

Standard ViT encoding

SPARO
(Separate-head Attention Read-Out)

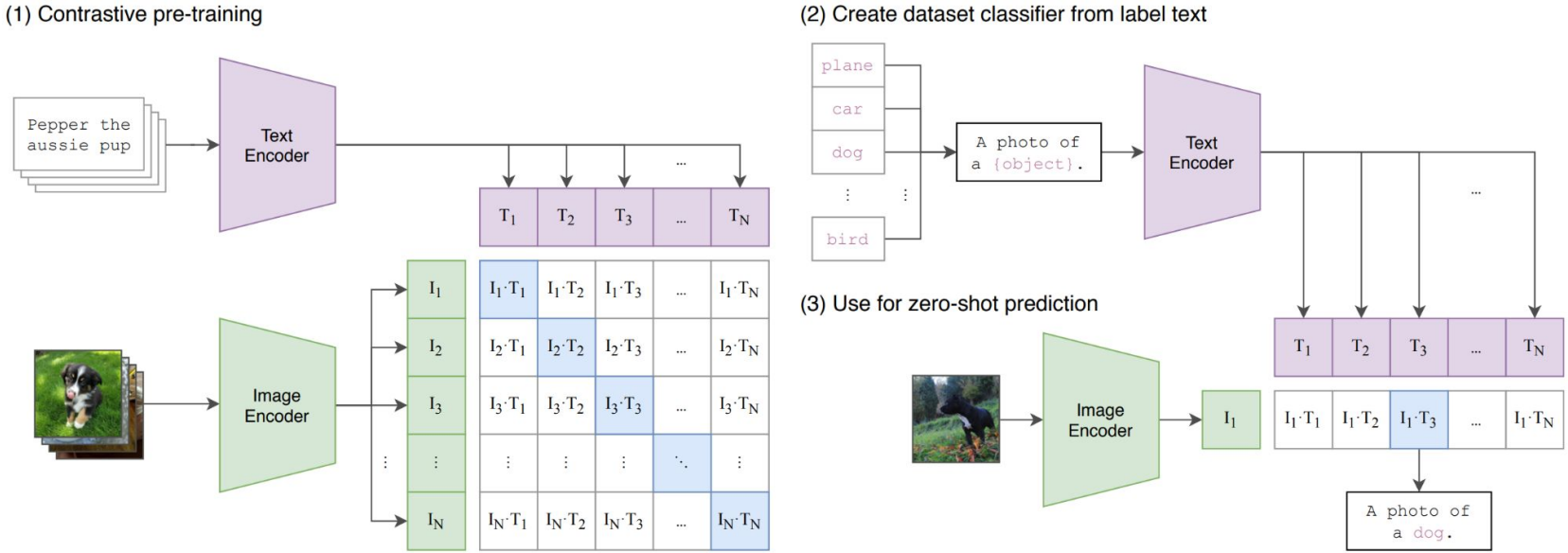
SPARO is a read-out mechanism for transformers that explicitly structures encodings as a collection of separately-attended concepts.



Article 3

SPARO with CLIP

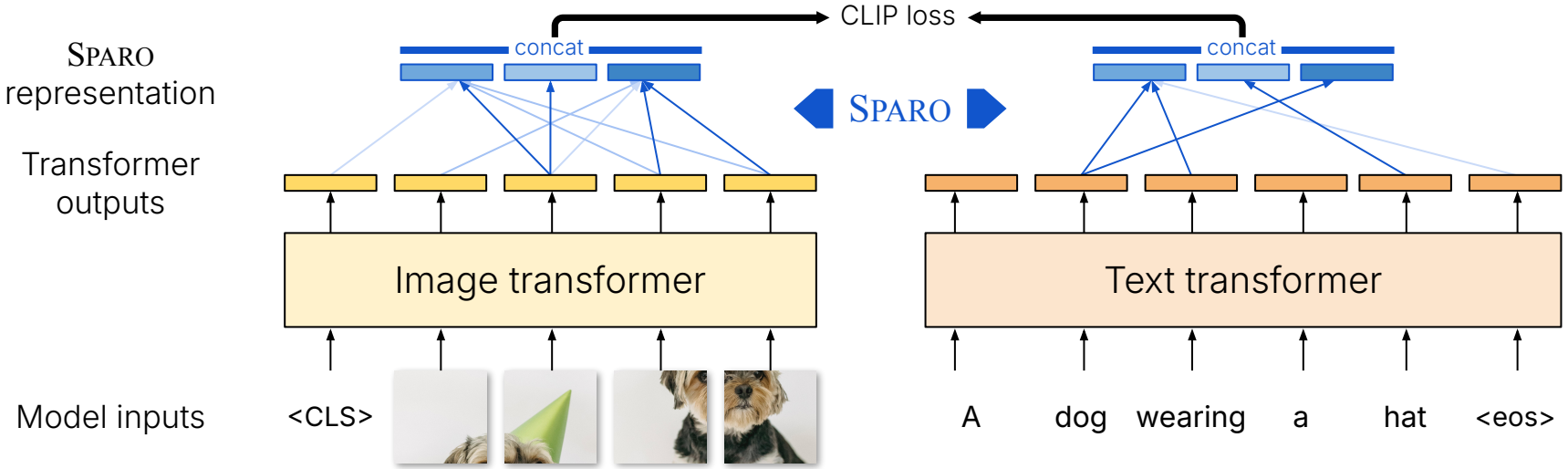
CLIP[1] (Contrastive Language-Image Pre-training)



(Figure taken from [1])

[1] Radford et al. *Learning transferable visual models from natural language supervision.* ICML 2021.

SPARO with CLIP



SPARO imposes a prior that both modalities share the same compositional world with the same attendable concepts

Article 3

CLIP Experiments

Zero-shot and linear probe accuracies

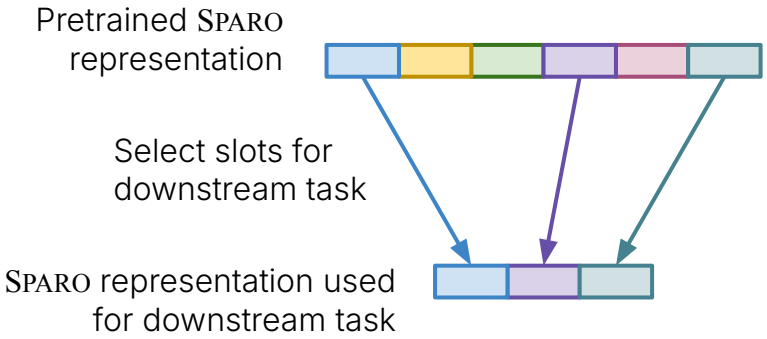
Train	Model	ImageNet-					Object Net	Sugar Crepe
		V1	V2	Sketch	A	R		
CC3M	CLIP^{16} $(\mathcal{C})$	0.141	0.122	0.068	0.033	0.177	0.080	0.611
	$\mathcal{C}$+GAP	0.156	0.134	0.069	0.033	0.187	0.081	0.616
	$\mathcal{C}$+SPARO	**0.170**	**0.140**	**0.088**	**0.035**	**0.221**	**0.098**	**0.625**
CC12M	CLIP^{16} $(\mathcal{C})$	0.361	0.311	0.249	0.091	0.467	0.218	0.697
	$\mathcal{C}$+GAP	0.382	0.330	0.262	0.101	0.501	0.241	0.695
	$\mathcal{C}$+SPARO	**0.406**	**0.350**	**0.298**	**0.113**	**0.559**	**0.268**	**0.723**
CC15M	CLIP^{16} $(\mathcal{C})$	0.384	0.337	0.268	0.105	0.503	0.238	0.699
	$\mathcal{C}$+GAP	0.399	0.343	0.287	0.114	0.531	0.252	0.701
	$\mathcal{C}$+SPARO	**0.437**	**0.378**	**0.317**	**0.145**	**0.579**	**0.279**	**0.730**
L400M	CLIP^{32} $(\mathcal{C})$	0.617	0.531	0.482	0.202	0.719	0.423	0.748
	$\mathcal{C}$+GAP	0.623	0.537	0.492	0.212	0.725	0.440	0.732
	$\mathcal{C}$+SPARO	**0.635**	**0.552**	**0.507**	**0.231**	**0.747**	**0.459**	**0.770**

Model	ImageNet linear probe			
	Train: CC3M	CC12M	CC15M	L400M
CLIP $(\mathcal{C})$	0.469	0.630	0.646	0.743
$\mathcal{C}$+GAP	0.504	0.649	0.664	0.747
$\mathcal{C}$+SPARO	**0.561**	**0.700**	**0.711**	**0.755**

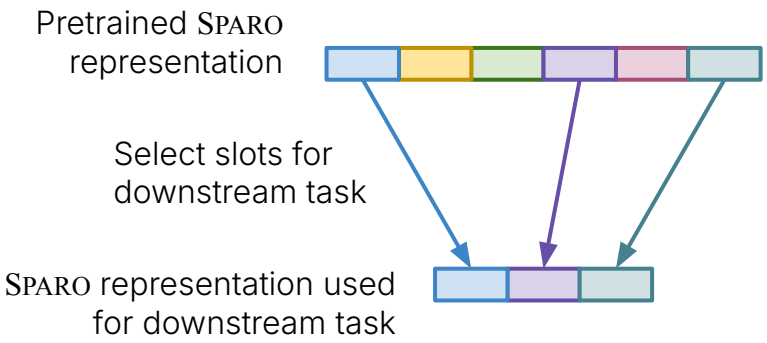
Post-hoc concept selection

Pretrained SPARO representation

Post-hoc concept selection



Post-hoc concept selection



SugarCrepe		CC15M		L400M	
	Slots:	100%	50%	100%	25%
Replace	Object	0.896	**0.912**	0.927	**0.929**
	Attribute	**0.798**	0.797	0.822	**0.854**
	Relation	0.666	**0.713**	0.678	**0.705**
Swap	Object	0.593	**0.659**	0.585	**0.646**
	Attribute	0.700	**0.703**	0.719	**0.740**
Add	Object	0.779	**0.809**	**0.866**	0.857
	Attribute	0.679	**0.737**	0.794	**0.822**
Average		0.730	**0.761**	0.770	**0.793**

Intervening to select a subset of slots with the highest ImageNet zero-shot performance

Visualizations

A herd of **cattle** walking down a road being followed by a cowboy

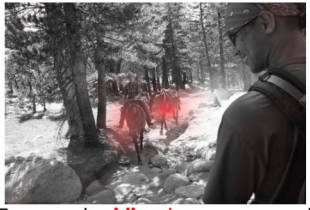
Several **surfboards** standing in a row on the beach

Two people riding **horses** on a rock path

A **man** sitting alone on a park bench in a park

A herd of cattle **walking** down a road being followed by a cowboy

Several surfboards **standing** in a row on the beach



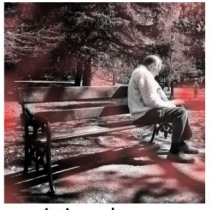
Two people **riding** horses on a rock path

A man sitting alone on a park **bench** in a park

A herd of cattle walking down a **road** being followed by a cowboy

Several surfboards standing in a row on the **beach**

Two people riding horses on a rock **path**



A man sitting alone on a park bench in a **park**

Article 3

Conclusion

Key takeaways

Attention offers a scalable solution to impart a stronger bias for compositionality

- Constructing encodings as slots produced by bottlenecks of separate attention heads encourages learning of data variations in terms of concepts represented by the heads
  - In CLIP, adds an additional prior for the modalities to share the same set of concepts

- Models trained with selective attention constraints achieve better downstream generalization

Conclusion of the presentation

Discussion

- Caveats in taking insights from cognitive science
  - Human learning utilizes a vast repository of prior knowledge, experience, and is intertwined with motivation, curiosity, and emotions
  - A large number of complex cognitive phenomena occur simultaneously and cognitive science models often simplify processes to make them tractable for study
  - Directly translating these simplified models into machine learning risks overlooking the nuanced nature of human cognition and the fundamental differences in optimization
- Humans do not always systematically generalize
  - Ability to perform complex, systematic logical inferences often requires thorough practice
  - Machine learning models, however, often fail at tasks human common sense succeeds at
- Systematic generalization is not sufficient in the real world
  - Models can systematically generalize on systemic biases

Open questions

- Iterated learning (IL)
  - Can IL help when there is no clear preference for a beneficial structure?
  - Can we dynamically adapt the learning bottleneck during training?
  - Can we distill IL's dynamics into a method that does not require a speaker-listener setup?

- Perturbed forgetting and SAM
  - Can we separate perturbed forgetting into a more general framework beyond SAM?
  - In the perturbed forgetting perspective of SAM, what really are the undesirable biases?
  - Is there a relationship between perturbed forgetting dynamics and loss surface flatness?

- SPARO
  - Can IL amplify SPARO's preference for compositionality for cleaner disentanglement?
  - Can we guide the type of concepts SPARO should prioritize learning?
  - Can SPARO offer insights for improving the internal attention mechanisms of transformers?

Thank you!


Aaron Courville


Aishwarya Agrawal


Alessandro Sordoni


Arian Hosseini


Bac Nguyen


Christos Tsirigotis


Dzmitry Bahdanau


Eeshan Dhekane


Frederick Tung


Gabriel L. Oliveira


Hattie Zhou


Hossein Sharifi-Noghabi


Hugo Larochelle


Jose Gallego


Max Schwarzer


Michael Noukhovitch


Ranjay Krishna


Samuel Lavoie


Sarvjeet Singh Ghotra


Simon Lacoste-Julien


Yuchen Lu