When and how do DNNs generalize to understand unseen category-pose combinations of objects?

Spandan Madan[1], Timothy Henry[2,3], Jamel Dozier[2,3], Helen Ho[3], Nishchal Bhandari[3], Tomotake Sasaki[4], Fre´do Durand[3], Hanspeter Pfister[1], Xavier Boix[2]

[1] SEAS, Harvard University [2] Center For Brains, Minds and Machines, MIT [3] MIT CSAIL [4] Fujitsu Laboratories Ltd.

Generalizing to new What and Where : essential for vision

*What does it mean to see? The plain man's answer (and Aristotle's, too) would be, to know **what** is **where** by looking.*
- David Marr

- Networks are brittle: don't work well beyond training distribution.
- In a multi-task setting: Can't see all combinations of all tasks. We need to generalize!
- Object Detection, VQA, Image Captioning all need to jointly reason about what and where simultaneously.
- Can networks generalize to unseen category-pose combinations?
- Ex: Can a model trained on Ford shown from front + Mitsubishi from side, predict category and pose for the third image below?


Ford Thunderbird, Front


Mitsubishi Lancer, Side


Ford Thunderbird, Side

Category-Pose Datasets

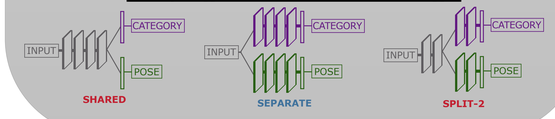
- MNIST variations: (i) MNIST-Position (ii) MNIST-Scale
- iLab dataset: Natural image dataset of physical toy objects on a turntable viewed from different viewpoints + intraclass variation.
- (Introducing our new) *Biased-Cars* dataset:
  - **Controlled:**
  - **Diverse**
  - **Photorealistic**
  - **Randomized**



Adding quantitatively controlled bias



Varying architectural design



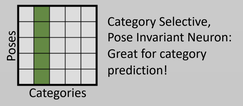
Neural Mechanisms: Selectivity and Invariance

- Selective Neuron: Fires if a feature is present.
- Invariance: Unaffected by every other feature.
- Intuition: A "ford selective + pose invariant" will do perfect classification.

$$S_c^k = \frac{\hat{a}^k - \bar{a}^k}{\hat{a}^k + \bar{a}^k}, \quad \text{where}$$

$$\hat{a}^k = \frac{1}{N}\sum_j a^k_{i^*kj}, \quad \bar{a}^k = \frac{\sum_{i \neq i^*}\sum_j a^k_{ij}}{N(N-1)}$$

$$I_p^k = 1 - \left(\max_j \; a^k_{i^*kj} - \min_j \; a^k_{i^*kj}\right)$$



Category Selective, Pose Invariant Neuron: Great for category prediction!



Pose Selective, Category Invariant Neuron: Great for pose prediction!

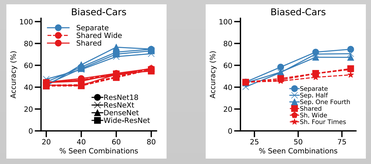
Specialization: A metric for division of labour in Multi-Task settings

- **Extending this framework to multi-task learning:** Network invariant to category cannot predict category! Similarly for pose.
- Possible solution: A division of labour: Splitting into two specialized networks: *Specialization Score.*

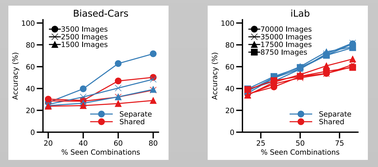
$$\Gamma^k = \begin{cases} \sqrt{S_c^k I_p^k} & \text{if } S_c^k > S_p^k \quad \text{(category neuron)} \\ \sqrt{S_p^k I_c^k} & \text{if } S_c^k \leq S_p^k \quad \text{(pose neuron)} \end{cases}.$$

Additional Controls

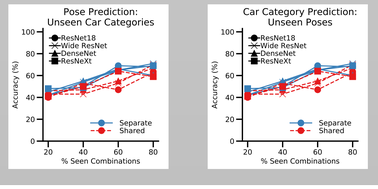
Control #1: Number of units in the network



Control #2: Dataset size, as opposed to dataset diversity

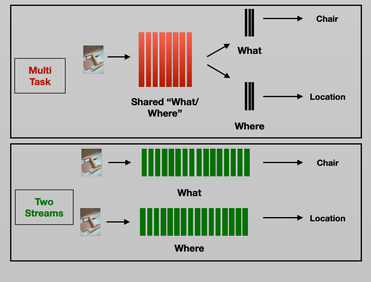


Control #3: Controlling separately for pose and category



Research Questions

Q1. When do neural networks generalize to unseen what-where compositions?
- What is the impact Data Diversity?
- What is the impact of Architectural Choices?

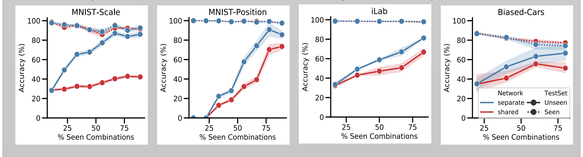


Q2. What are the neural mechanisms driving this behavior?
- Do selective and invariant representations have a role to play?
- What behaviors emerges at the individual neuron level?
- How do findings change in a multi task setting?

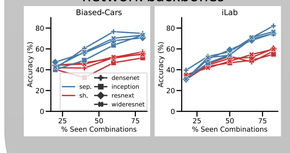
Results I: *When* do networks generalize?

Finding I. Data diversity helps generalizate even to unseen combinations.

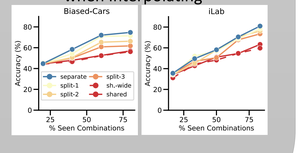
Finding II. Separate "what" and "where" gives better generalization.



Holds true for multiple network backbones

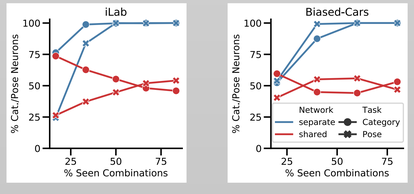


Behavior transitions smoothly when Interpolating



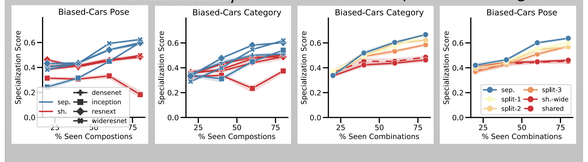
Results I: *how* do networks generalize?

Finding III. Do neurons get specialized? Yes!



Specialization increases with more data diversity

Specialization increases with separate branching



Finding IV. Specialization correlates with generalization, and may explain why networks generalize with (i) more diversity, (ii)separate branches

Conclusions

- Data diversity can help networks ***extrapolate.***
- Separate architectures extrapolate better than shared ones.
- S*pecialization* serves as a quantitative framework to understand generalization in multi-task settings.
- Separate architectures facilitate specialization, which leads to better performance.

Future Work

- Extend to other tasks like depth, and segmentation among others.
- Mixed selectivity of neurons.
- Can electivity and invariance be defined for a group of neurons?
- Extending beyond 2 tasks.