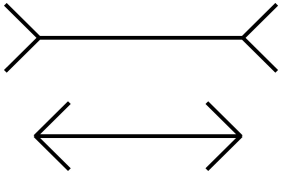
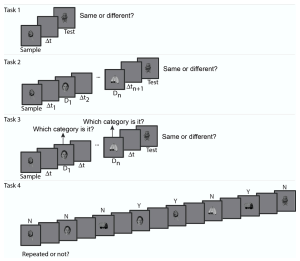
Neuro 140. Biological and Artificial Intelligence
Spring 2024
**List of Potential Projects**

| Project number | Project Title | Brief Description, Hypotheses, Questions | References | Difficulty level [0 = easy, 10 = hard] | Link to more information, data, code |
|---|---|---|---|---|---|
| 1 | **Building models that generalize well. There are three different options:**<br>**1. Weather modifications**<br>**2. Day-night**<br>**3. Real vs Cartoons/Sketches** | Neural networks are notoriously bad at generalizing to test data which is significantly different from train data.<br><br>Recent efforts try to work across such shifts. Implement some recent works, and try to suggest modifications which might do well. | Musat et al 2021<br><br>Robustness to rain<br><br>Madan et al 2024 | 7 | |
| 2 | **The problem of parameters in linear systems** | Current deep convolutional neural networks (CNNs) are typically underdetermined. Why is it that they do not overfit? Compute condition numbers, rademacher averages for underdetermined and overdetermined linear systems to assess robustness | Poggio, Kur, Banburski. Double descent in the condition number<br><br>Radhakrishnan et al 2020 | 6 | |
| 3 | **Sharpened and faded object boundaries** | It is widely known that CNNs are biased to textures rather than shapes.<br><br>Taking a dataset with segmentation maps, sharpen or blur the edges in the training data. How does this impact the texture loving nature of CNNs? | Geirhos et al 2019 | 7 | http://www.image-net.org/ |
| 4 | **Impact of changing Transition Function in Deep RL**<br> | (a) Use reinforcement learning to teach a network to play a video game like PACMAN.<br><br>(b) Transition function defines the probabilities with which PACMAN ghosts move.<br><br>How does the RL agent perform when the probabilities of ghost movements are different in testing than training? | https://github.com/tychovdo/PacmanDQN<br><br>https://www.youtube.com/watch?v=QilHGSYbjDQ<br><br>Bono et al 2024 | 8 | |
| 5 | **Enforcing brain like activations** | Recent works have trained linear models which take as input a CNN layer's activations and map them to neuronal activations collected from brain measurements. | | 5 | |

| | | | | | |
|---|---|---|---|---|---|
| | | Reproduce these results and build on them. | | | |
| **6** | **Graphical humor**<br> | Write an algorithm that will predict human judgments on whether an image is funny or not (or quantitative values on how funny an image is). | Veedant et al 2024 | 10 | Link to data and comments |
| **7** | **Visual illusions**<br>  | Are current computer vision systems susceptible to human visual illusions?<br><br>How do CNNs see these images? Can we create more such images automatically? | Kreiman. The phenomenology of seeing<br><br>Lotter et al 2020<br><br>Williams et al 2018 | 8 | Link to dataset |
| **8** | **Working memory**<br>  | Create a model that can solve a variety of delay match to sample working memory tasks. | Miller. Working memory 2.0<br><br>Xiao et al 2023<br><br>Yang et al 2019 | 9 | Link to discussion and ideas |
| **9** | **Turing project** | Test state-of-the-art algorithms as human imitators | Zhang et al. Human or machine? Turing tests for vision and language | 4 | Link to data |
| **10** | **Interpretability in neural networks** | Evaluate the extent to which unit activations are “interpretable” | Olah et al 2021 | 5 | |

| | | | | | |
|---|---|---|---|---|---|
| | | | Bardon et al 2022 | | |
| **11** | **Which AI said that?** | A lot of work has explored whether it is possible to detect text produced by large language models (e.g., the ones underlying ChatGPT). A related, less well-explored idea is large language model attribution: assuming that a piece of text was produced by AI, which AI was it? GPT-2, GPT-3, GPT-4, Llama, Claude, Bard, Falcon?<br><br>Build on existing work (see references) to implement a classifier that can distinguish between text produced by two or more different large language models. Then, run some experiments with the test data. For example, take some paragraphs written by Llama and ask GPT-3 to rewrite or summarize them (and vice-versa). Does your attribution model say it was written by Llama, or GPT-3?<br>Another idea: can you engineer a prompt for GPT-3 that makes it produce text that consistently "sounds like" Llama such that it fools the attribution model (and vice-versa)?<br><br>This could help test hypotheses about the features attribution models use to make their decisions - e.g. is it about word choice, or sentence structures, or wordiness, or style, content, etc? | Uchendu et al 2020<br><br>Munir et al 2021 | 4 | Github repo from Uchendu et al paper |
| **12** | **Apply perceptual modulation for qualitative interpretability of deep learning models** | Deep neural networks sometimes make predictions in ways that are difficult to explain. For example, CNNs can predict if someone is female or male using a photograph of their retina, and it is not understood why. CNNs can also outperform dermatologists at detecting skin cancer from photographs, for reasons that are not fully understood.<br>In this project, train robustified neural networks on publicly available datasets of sex-labeled retina and/or (separately) skin cancer photographs. Use these networks to perceptually modify the images towards or away from the target class (see Gaziv et al), exaggerating or diminishing features that indicate cancer vs. no cancer, male vs female, etc, and helping to explain why each image was classified a certain way by the network. Compare with other explainability methods such as gradcam. | Gaziv et al 2023<br><br>Krzywicki et al 2023<br><br>Rezvantalab et al 2018 | 6 | See Gaziv et al for code (which was based on this library), Krzywicki et al for fundus (retina) image datasets (make sure to choose datasets that have sex-labeled images, you may have to combine multiple datasets), and Rezvantalab et al for skin cancer dataset |

| | | | | | |
|---|---|---|---|---|---|
| **13** | **Use large language models for structurally focused topic modeling of natural language datasets** | Large language models can be used as powerful tools to extract structured information from natural language. Design LLM prompts to extract carefully selected types of structured information from natural language datasets (e.g., graph representations or summaries with a strictly-defined structure), and train topic models (e.g. latent Dirichlet allocation) on the extracted data to learn about themes in the data and test hypotheses (which are specific to the chosen dataset). For example, this approach was previously applied to a dataset of transcripts of youtube videos that teach linear algebra (e.g., Khan Academy), to investigate the relationship between video popularity and the connections the instructor draws between concepts. However, the approach could be extended to natural language datasets in other domains - e.g. mental health (perhaps postings on mental health themed reddit communities, or patient-therapist interactions that contain cognitive distortions), song lyrics, jokes, blog posts, scientific papers, etc. | Talbot et al 2023 | 8 | https://github.com/Hramir/educational_concept_librarian |
| **14** | **Lifelong learning with latent replay in transformers** | When artificial neural networks try to learn more than one task in a sequence, they "catastrophically forget" earlier tasks. E.g., if a network trained to recognize dogs vs cats is subsequently trained to distinguish cars vs trucks, its performance on dogs vs cats drops to chance levels very quickly. This is a fundamental problem in deep learning that has been addressed in a variety of ways - one promising group of approaches is referred to as feature-level/latent replay (see Pellegrini et al 2020). This has been successful in CNNs - in this project, try to implement a similar approach in (vision or language) transformers. | Hadsell et al 2020<br><br>Pellegrini et al 2020<br><br>Talbot et al 2023 | 8 | |
| **15** | **Neural networks as models of learners** | One goal in ML research focused on education is to develop models that can predict how students learn - for example, at what stage is each student "ready" to work on a problem of a given difficulty, and when are they vulnerable to forgetting each piece of knowledge?<br><br>In this project, train a tiny CNN on a series of tasks in a continual learning setting (e.g., learn two digits from MNIST | | 9 | |

| | | | | | |
|---|---|---|---|---|---|
| | | before moving on to the next two, for 5 binary classification tasks in total). Then, try to predict the tiny network's learning behavior using a more powerful model. This could be a bigger CNN that learns the same sequence of tasks and "aligns" its behavior to the smaller network, or perhaps an LSTM that makes predictions over time, etc.<br><br>The ultimate goal is a model that can predict the learning behavior of a human, as opposed to a CNN model - this would be an ambitious stretch goal for a one-semester project. | | | |
| **16** | **Does the effortful retrieval hypothesis translate to artificial neural networks?** | When reviewing previously-learned knowledge, humans improve the strength of their memory the most when they review something that is almost, but not quite, forgotten. The harder it is to retrieve something from memory, the more that memory is strengthened during the retrieval process - this is referred to as the "effortful retrieval hypothesis," and it is supported by a variety of behavioral studies in humans.<br><br>This could be an evolutionary adaptation - maybe if an ancient human ancestor tried really hard to remember where the raspberry bushes are, that is a signal indicating that particular memory is very important, and the brain evolved some special mechanism to preserve such memories. And/or, alternatively, perhaps there is some fundamental property of neural networks such that almost-forgotten information is consolidated by "retrieval" (e.g., backpropagation in artificial neural networks) better than easily accessible information that was learned or retrieved relatively recently.<br><br>Test these hypotheses using a continually-learning CNN, by systematically manipulating the timing with which each image, task, or category is presented to the network during training. Stretch goal: if there is no retrieval effort effect in CNNs, can we somehow artificially emulate one to produce a more human-like learner? (see also "neural networks as models of learners" project above) | Cognitive science references:<br>Pyc et al 2009<br><br>Benjamin et al 2010 | 7 | |

| | | | | | |
|---|---|---|---|---|---|
| **17** | **Detect disease in chest X-rays, and investigate how to generalize to a new clinical population** | Deep neural networks are easily confused by changes in context, which is a problem in many applications such as clinical medicine. For example, a CNN trained to interpret chest x-rays using data from the US might perform well within the US but poorly in another country. Reasons include different patient populations, different X-ray equipment, different image pre-processing, different prevalence of diseases, etc. all of which can introduce subtle shifts in the training data. This touches on issues of equity and fairness in ML, which is especially critical in clinical ML.<br><br>One promising approach is to collect as diverse of a dataset as possible, but in practice this can lead to a "jack of all trades, master of none" effect where the model performs decently in general but doesn't perform optimally in any given context (see Futoma et al 2020).<br><br>In this project, train a CNN to detect diseases in chest X-rays, combining datasets from diverse sources in various ways. What is the best method to produce an optimal CNN that reads chest X-rays for a given clinical site (hospital, region, country, etc), especially when data from the target site is limited?<br>1. Train only on the limited data from the target site?<br>2. Train on a diverse dataset from many different (other) sources and hope the model generalizes to the target site?<br>3. Train on a diverse dataset that includes data from the target site?<br>4. Train on a diverse dataset, then fine-tune to data from the target site?<br>5. Number 4, but mix in diverse data during the fine-tuning process?<br>6. In fine-tuning approaches, how much data is needed to get good results? Does the size and/or diversity of the pretraining x-ray dataset matter?<br><br>In this project, pay special attention to issues of calibration and dataset balance. Consider the following: a model is trained to predict "pneumonia" vs "no pneumonia" from | Futoma et al 2020<br><br>Deng et al 2022 | 3 | https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection<br><br>https://www.kaggle.com/datasets/nih-chest-xrays/data<br><br>https://stanfordmlgroup.github.io/competitions/chexpert/<br><br>https://bimcv.cipf.es/bimcv-projects/padchest/<br><br>Search the web for more! |

| | | | | | |
|---|---|---|---|---|---|
| | | chest x-rays. Imagine that 2% of people getting X-rays in the US have pneumonia. The model can achieve 98% accuracy by ALWAYS guessing “no pneumonia”. But maybe in Australia, for some reason 5% of people getting X-rayed have pneumonia. Now, the same model only has 95% accuracy.. Even a model that is doing more than just always guessing no-pneumonia will tend to be “calibrated” to the base disease rate of the population it was trained on, which might be different from the target population. | | | |