6. Developing common representations in AI and humans

- ***Using human brain activity to guide machine learning*** **Fong et al. (2017)**
  - **Problem** Explore how incorporating human brain activity can improve machine learning performance and make ML representations more "human-like."
  - **Methodologies** The researchers employed a multi-step approach
    - **Brain Data Acquisition:** fMRI data from 7 regions of interest (ROIs) in the brain, all involved in various stages of visual processing.
    - **Model Types:** Both CNNs and Histograms of Oriented Gradients (HOG) were used as feature extraction methods.
    - **Response Strength Derivation:** A binary classifier was trained on fMRI brain activity. The **response strength** for each stimulus is its distance from the decision boundary of this brain-data-trained classifier.
    - **Activity Weighted Loss (AWL):** AWL specifically penalizes misclassified examples that humans find easy to distinguish.
    - **Classifier Training**
      - **Conventional Image Classifier:** A Radial Basis Function SVM classifier was trained.
      - **Brain-Guided Classifier:** An SVM classifier was trained using the AWL function, effectively reweighting the margins based on brain activity data.
    - **Stimuli:** Images from 5 distinct categories were used for classification tasks.
  - **Results**
    - **Improved Feature Representation**
    - **Area-Specific Enhancements:** Different brain regions contribute unique and valuable information for guiding ML.
    - **Human-Like Decisions:** Direct measures of brain activity can effectively guide ML algorithms toward making decisions that more closely resemble human cognitive processes.
- ***Interpretable Semantic Vectors from a Joint Model of Brain- and Text-Based Meaning*** **Fyshe et al. (2014)**
  - **Problem** This study aimed to improve Vector Space Models (VSM) by incorporating human brain activity, believing it would lead to more accurate and human-like word meanings.
  - **Methodologies** The researchers combined text data and brain activity data (fMRI and MEG from subjects reading words). They used a technique called Joint Non-Negative Sparse Embedding (JNNSE). This method allowed them to create a shared "semantic space" that combined information from both text and the brain.
  - **Results**
    - **Better Human-Like Similarity:** JNNSE predicted human judgments of word similarity much more accurately than text-only models.
    - **Enhanced Brain-Based Word Prediction:** Improved prediction of words solely from brain activity.
    - **Generating Text Data for Rare Words:** The models could accurately predict text-based features for words when only brain data was available.
    - **Visualizing Brain Semantics:** They mapped learned semantic concepts directly onto specific brain areas, providing insights into how the brain organizes meaning.
- ***Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features*** **Palazzo et al. (2020)**
  - **Problem** How to identifying the most important parts of an image by combining brain activity data (EEG) with image data. This is done by merging static image features with dynamic, time-varying brain signals.
    - Create a multimodal learning system that brings brain data and image data into a single, unified "**latent space.**"
    - Use brain activity to directly guide machine learning tasks, specifically visual **saliency modeling**.
  - **Methodologies** The researchers used EEG data from 6 people watching 40 categories of ImageNet photos, recording high-resolution brain activity for 0.5 seconds per image. To combine this with image data, they built a Siamese network architecture.
    - **Siamese Network:** The core idea is to train two identical neural networks in parallel. One processes EEG data, and the other processes visual features from images.
    - **Finetuning:** For images, they fine-tuned existing image recognition networks to learn "brain-related" representations.
    - **Image Classification:** tested the effectiveness as feature extractors for a standard 40-class image classification task.
    - **Saliency Detection:** They developed a novel, non-learning method for saliency detection using their trained joint embedding.
      - Masking out a small region of an image.
      - Calculating the similarity between the original image's EEG features and the masked image's visual features in the joint space.
      - The decrease in compatibility indicates how "salient" or important that masked region was.
  - **Results**
    - **Improved Image Classification:** The learned joint representations are highly effective.
    - **Benefits of Joint Features for Classification:** Boosted classification performance, especially for EEG-based classification.
    - **Effective Saliency Detection:** Infusing brain data can directly guide the identification of visually important regions in images.
- ***Human uncertainty makes classification more robust*** **Peterson et al. (2019)**
  - **Problem** This paper investigates if teaching AI with human-like uncertainty, through **soft labels**, can make classification more robust and better at generalizing to new data.
  - **Methodologies**
    - **CIFAR10H:** A dataset of 500,000 human judgments on CIFAR10 images, generating soft-label probability distributions for each image.
    - **Training:** They trained various neural networks using these soft labels, aiming to minimize the difference between the network's output and the human-derived probabilities.
    - **Evaluation:** They evaluated these models' generalization capabilities, calibration (how well confidence matches correctness), and robustness against adversarial attacks.
  - **Results**
    - **Improved Generalization**
    - **Better Calibration:** These models were less overconfident on incorrect predictions while remaining confident on correct ones, mimicking human judgment.
    - **Enhanced Robustness**