Relatively new focus on using videos as a more natural representation of how the visual system receives info in the real world and processes this. Allows for more accurate representations of visual pathways in artificial systems.

There are a few github sources that provide algonauts 2021 analysis

Regarding feature identification - generally a classification model is not necessary for feature extraction (the features are represented by abstract values which can correspond to specific voxels in a given region)

Fisher Vector encoding comes up a few times

Agrawal 2014

* Encoding models create nonlinear map between stim and brain activity
* *“A fully satisfying model of human vision should predict activity across the entire visual hierarchy directly from pixels, without the need for any human intervention.”*
  + Autoencode semantic info directly from pixels
* ConvNet model can be used to recover the visual receptive fields from single voxels.
  + *Regularized linear regression was used to regress the features (bottom left) extracted from the 1260 images in the estimation set (top left) onto brain activity recorded from each voxel (bottom right). The resulting encoding model weights (bottom center) can be interpreted as a tuning curve for the image features.*
* encoding model consists of a feature space that provides a linearizing transformation between the stimulus images and measured brain activity.
  + for every voxel, we used regularized linear regression to find a set of weights that predicted voxel activity from the feature-space representations of each image.

Wu 2018

* Video classification and captioning are extensions of image classification - highlights ways this can be done - summary of image class provided.
* CNNs: takeaway is that AlexNet was the first to solve ImageNet Challenge successfully, however since then, CNNs with more discriminative power have been created primarily through increasing depth of network (16, 22, 152, etc) increased
  + “depth allows the network to better approximate the target function, generating better feature representations with higher discriminative power.”
* RNNs: better explore temporal info if sequential data - they allow cyclical connections/”memory of previous states
  + Be aware of vanishing and exploding gradients (LTSM solves this?)
  + LTSM popular RNN for solving video classification
* off-the-shelf CNN features coupled with kernel SVMs can obtain decent recognition performance.

THINGSvision

* Library that allows quicker DNN calls for extraction etc. - contains all pytorch nn.Modules in addition to a few newer ones that are all focused on vision related analysis
* a readily-available package for a streamlined extraction of neural network activation.
* feedforward and recurrent variants of CORnet, a recent DNN model that was inspired by the architecture of the non-human primate visual system and that leverages recurrence to more closely resemble biological processing mechanisms
* Section 3 covers the application of the neural nets in TORCHvision
  + The correspondence of a DNN’s penultimate layer to human behavioral representations has been studied extensively and is therefore often used when investigating the representations of abstract visual concepts in neural network models

Khosla 2021

* build group-level models of neural activity that incorporate several inductive biases about neural information processing, including hierarchical processing, temporal assimilation, and auditory-visual interactions.
* hierarchical feature extractor based on feature
* pyramid networks (19) that combines features from early, intermediate, and later layers simultaneously. The detailed architectures of both components, including the feature extractor and convolutional response model, are described in fig. S3.
* we employed WordNet labels that were provided as part of the HCP movie-watching data pipeline
* <https://github.com/mk2299/MultimodalEncoding>

