OVERVIEW OF ML APPROACHES

- BARRET et al. → synergy between NN and artificial NN (DNN)

DNN

Millions parameters — No engineered features — High performances

Questions for both:
→ how are data processed to transform stimuli into human level knowledge
→ how do they deal with high dim. data

tonotopy: like retinotopy but for audio stimuli

retinotopy

RECEPTIVE FIELDS → • relations between scene areas and features processed there
→ • small receptive fields in first layers of visual cortex
→ • receptive fields grows as information is transmitted to higher level areas of visual cortex
× this enables sensitivity to larger areas of space
× more complex features and abstraction → "concept cells" → recognise object but no appearance

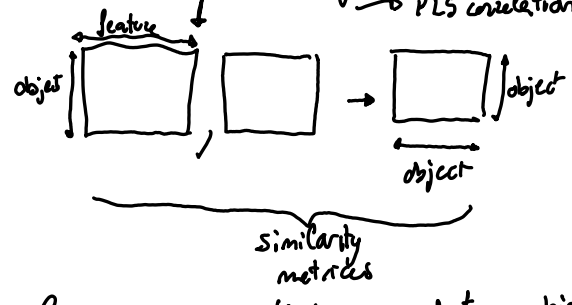
Does this hold also for DNN?
• patterns
• pooling

ANALOGIES

ABLATION: → studying brain lesions offer much information about potential function of brain areas (& REORGANIZATION) → can be achieved with various tools [?]
↳ we can always consider effect of silencing neurons on the output (structural pruning) ← remove neuron with its outgoing weights
• NN which have good generalization are more robust to ablation

DIMENSIONALITY REDUCTION:
→ brain store infos. in a distributed way → needs multivariate analysis → redundancy
→ distributed manner dependencies
→ correlations of units means redundancy
↳ In DNN we have a lot of redundancy and correlation

REPRESENTATIONAL GEOMETRIES
→ how does a DNN represent informations across different layers in space and time
↳ comparison of different states using → canonical correlation analysis → identify low level factors that maximize correlation → SUPERVISED!
→ PLS correlation



→ Linear regressions methods can predict neurobiological activation

SPATIAL METHODS: ← Spicer & Sanborn

place items on a multidimensional space and use their location to draw conclusions about categorization

→ CLASSIFICATION: locate items with hyperplane
  ↳ ASSUMPTION: there exist a statistical model with a center and some parameters that are learnable in training

EG: GENERALIZED CONTEXT MODEL

$$P(C_j \mid i) = \left(\sum_{j \in i} s_{ij}\right)^{\gamma} / \sum_k \left(\sum_{k \in K} s_{ik}\right)^{\gamma}$$

(for simplicity = 1; γ: deterministic regulator) → No Softplus

prob. that i belongs to C_j

s_{ij}: similarity of i to j

LOGICAL METHODS

concepts are based on definition that is applied

the features of the object. → SEARCH FOR RULES THAT MAXIMISE DISCRIMINATION BETWEEN STIMULI } also probabilistic

eg 1st order Logic

ANN: (artificial NN)

do not make assumptions about the representation involved, but offer an implementation method

Which model is the most accurate? Depends on the area of science you are working on → e.g. do you need underlying representations?

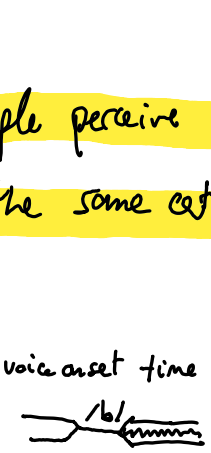
CATEGORICAL PERCEPTION

" The phenomenon in which people perceive stimuli from different categories as more different from each other that stimuli from within the same category " → example: sound

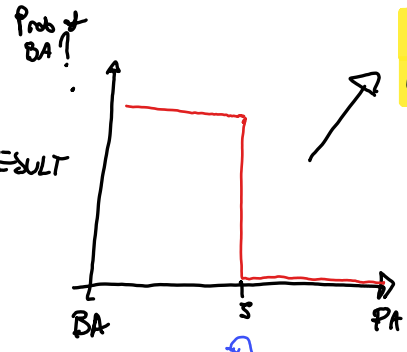
EXAMPLE

stimuli 100-800Hz → → stimulus judgements / generalization / confusion

VOT: voice onset time of consonant



→ RESULT



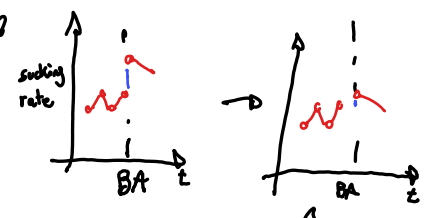
we are very good to distinguish 4 to 6 but we are unable to distinguish 2 to 4

CATEGORICAL PERCEPTION:
If there is no category boundary, then the two stimuli are the same for us

EVIDENCES →

- CHANGE DEAFNESS: participants were repeated words in a stream. Halfway the voice changed but just 40% of participants noticed it

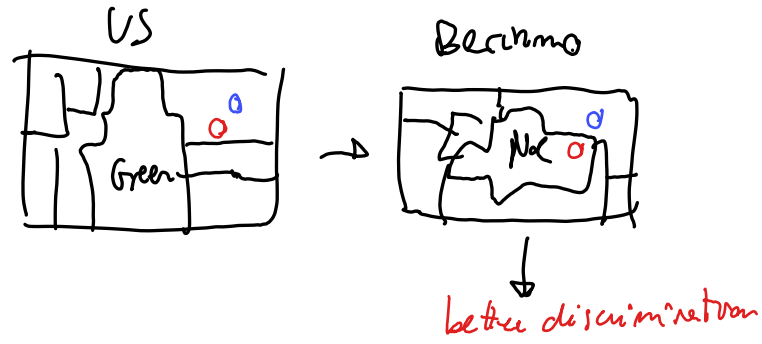
} we are fast to learn that category



4 months babies

- LINGUISTIC CATEGORICAL PERCEPTION (colors)
  - more easy to distinguish blue than green than shades even if they have same distance
  - ↓
  - discriminate better when we introduce categories

} Berinmo tribe distinguish better some colors because they give a particular name to some colors which we usually consider just shades



↓
better discrimination

CONCEPTUAL STRUCTURE

- Typical approach → assume that words are associated with concepts representations

→ Problem of polysemy → we need context (Murphy) → Klein and Murphy: we can store multiple meaning for the same word (1 ↔ 1)

→ Problem of homophones (bait, bate) → need context
- homonyms (bank, bank) → need context

*anomia → a form of aphasia, difficulty to retrieve lexical form of a concept

Classic View of word meaning

- Definitions are sets of NECESSARY conditions that are jointly sufficient
- we only learn these defining features

HIERARCHY AND INHERITANCE
graph of relations

classic example of representation

Propositional Networks

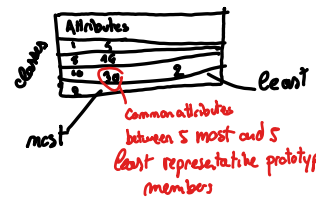
ISSUES!
- → some features are stored, others inferred
- → does not take into account the frequency of encountered statements
- → verification time does not scale accordingly to the # of ISA link
- → logic does not hold sometimes (chair ex.)

(Collins & Quillian 1969)

Refutate by Rosch

- Problem: 'Game' concept is not grabbable by some necessary/sufficient conditions

→ game is defined by similarity between games



→ fuzzy categories boundaries problem: is stroke a disease?

→ insecurity, fast changes in opinions

Typicality (?)

Categories have central and peripheral members

Why do we choose a particular word? (cat)
- level of information (why add useless information?)
- distinctiveness (what makes it different)

- ABSTRACTION

SUPERORDINATE
INF ✗
DIST ✓

BASIC
INF ✓
DIST ✓

SUBORDINATE
INF ✓
DIST ✗
→ expensive

MODELING CONCEPTUAL ORGANIZATION (Lake et al.)

→ Notion of typicality → Can DL systems imitate cognitive models?
↳ changes categorization speed

METHOD

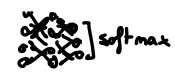
behavioral

Dataset of ratings of typicality for images drawn for 8 categories
↓
Average of typicality ratings for across all s.s.

VS

computational

- 3 convnet architectures



Result: correct label appears in TOP5 for 85% of images

Evaluating Classification strenght (in general)

Raw typicality

$$y_j = \sum_{i=1}^{4096} w_{ij} x_i$$

y_j → particular abstract representation

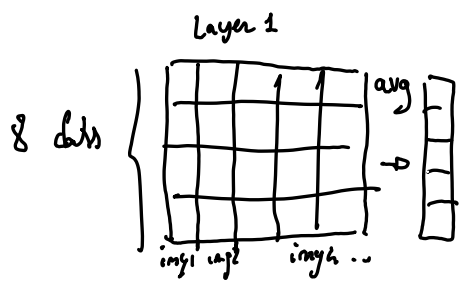
Similar result →

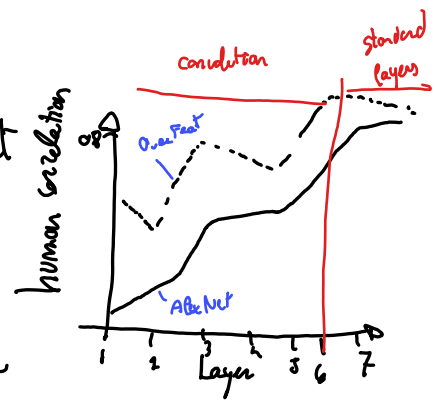
Contrast typicality

$$z_j = \frac{e^{y_j}}{\sum_{j=1}^{1000} e^{y_j}}$$

Conclusion: for everyday categories it is possible to predict human typicality ratings

→ More in depth: examination within layers: → pass dataset to NN
→ get activation vectors
→ average by rows over all activation matrices to get typicality response of a class
→ compute it with single image response





Words features as models of cognition → Mitchell et. al 2008

→ How do humans represent concepts in the brain?

→ Which features (basis functions)

Previously: regression on FMRI on words sampled from different categories → pure decoder

FMRI —predict→ word

→ predict activation for specific word

→ Problem of high # of features using encoder (they use just the 500 already known to be significant) [for voxels]

⇓

GENERATIVE ENCODER MODEL

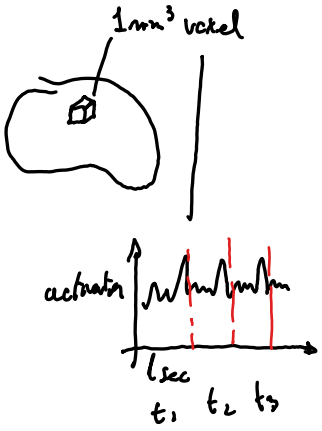
① encode words into semantics vectors that should capure meaning.

- useful
- grey
- pet
- ....

$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$ → 25 features

extracted from corpus counting co-occurences

celery:
- eat → 0.83
- taste → 0.3461
- 25 verbs
- ⋮



② single voxel analysis: collect activation for each voxel for each semantic feature

collect fMRI data

	t_1	t_2	t_3	avg
1	x_1	x_2	x_3	$\bar{x}$
2	y_1	y_2	y_3	$\bar{y}$
.				
.				
25	z_1	z_2	z_3	$\bar{z}$

	t_1 ---- t_3
1	
⋮	
58	

→ ∀ voxel

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{58} \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{58} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{58} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1\,25} \\ \vdots & & & \\ 1 & X_{58\,1} & & X_{58\,25} \end{bmatrix}$$

- Y: labels — activation of a SINGLE voxel for each of the 58 words
- ε: bias
- β: 25 weights
- X: semantic features — from corpus statistics

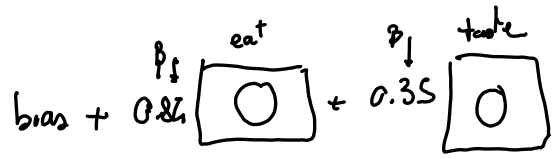
		eat	taste	...
celery	1	0.8368	0.3461	...
...				

③ predict neural fMRI activation of every voxel location as a weighted sum of neural activations contributed by each of the semantic features (58/60)

We want to predict 60th word using the other 59

↓

Just multiply the semantic features vector of that word by betas and add bias



$$\text{prediction} = \sum_{i=1}^{35} f_i(w) c_{vi}$$

↳ redo it for each word using leave-two-out

1770 test

↓

Mean accuracy 0.79

④ Examine importance: → activation of area of the brain that are known to be related to some semantics (eg. food) → most activating words are in those semantics

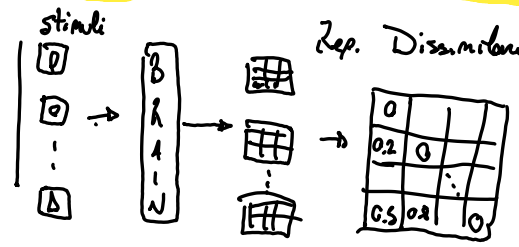
→ random basis set is crucial → random sets always do worst (0.6 vs 0.79)

STUDYING BRAIN REPRESENTATION VIA SIMILARITY SPACES

Kriegeskorte 2008
RSA

① CONSTRUCT A SIMILARITY MATRIX

RSA is 2nd order RDM

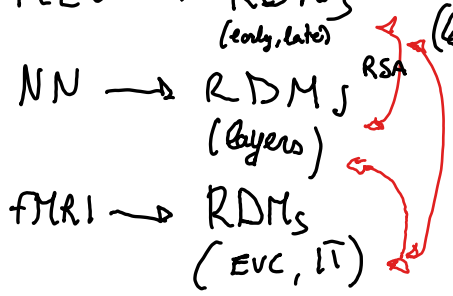


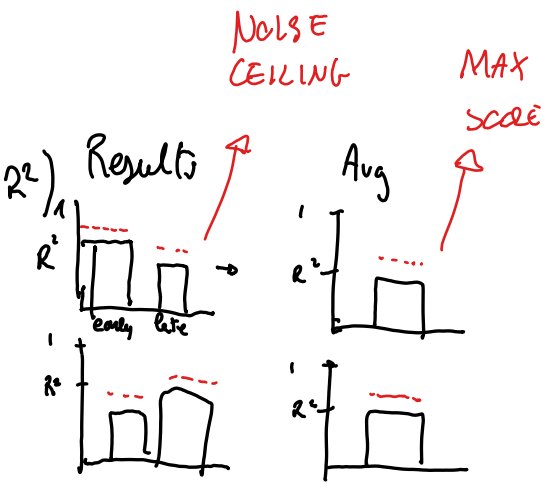
→ in time I have RDMs, one for each time interval

② COMPARE DISSIMILARITY MATRICES

obj x obj

RSA allows us to compare different similarity measures





NOISE CEILING → maximum value given the noise → comes from stochasticity of human behaviour
↳ can be measured repeting some experiment and computing correlation

obviously this lead to a loss of dimension if #features > #obj

eg: obj × voxels → obj × obj

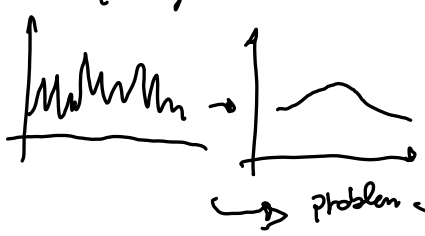
WHY IS THIS RELEVANT?

↳ We can check if different areas react to the same stimuli.
- test hypothesis of area where informations are processed
- compare goodness of some brain areas for some tasks.

COMPUTE RDMs

NEURO BIO

Stimuli x voxel
↓
a lot of variation (common problem of the UNIVARIATE ANALYSIS)
↓
introduce gaussian filter
↓
spread intensity at each voxel to neighboring voxels

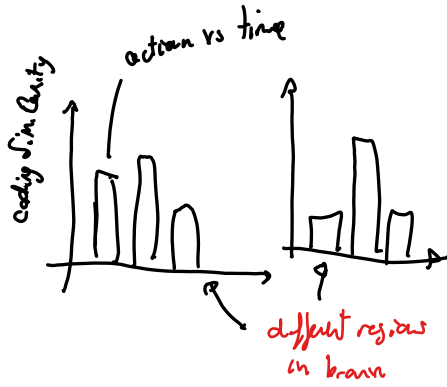


↳ problem

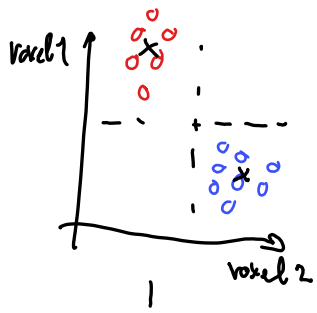
COMPUTATION MODELS
Stimuli x feature

MULTIVARIATE ANALYSES

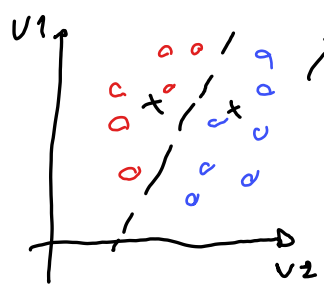
Consider the entire set of values in each region and quantify how similar those patterns are for the 3 conditions → spatial, temporal, action-related



MULTIVOXEL PATTERN ANALYSIS



UNIVARIATE ANALYSIS IS ENOUGH



similar joint activation

UA NOT SUFFICIENT

EVALUATING AND IMPROVING THE CORRESPONDENCE BETWEEN DNN → RSA AND HUMAN REPRESENTATIONS → similarity judgements

Joshua C. Peterson

DATASET: 6 categories, 120 images each

for HUMAN → Amazon Mechanical Turk → pairwise similarity ratings (0 : not similar … 10 : very similar) → 6 120×120 matrices

$$S = F \cdot F^T$$

- S: human similarity RDM (0,0)
- F: feature matrix (dot product)

features × objs (e.g. 120) × 4096 × 4096 → similarity matrix (120×120)

for DNN → consider activations values

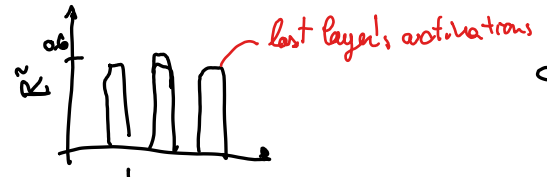
↳ used

neurons in layer

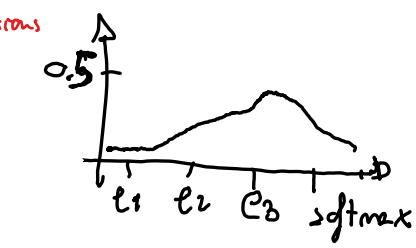
$$\begin{array}{c} \\ 1 \\ \\ \\ m \end{array} \begin{array}{c} \overset{1}{} \quad \overset{n}{} \\ \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \end{array}$$

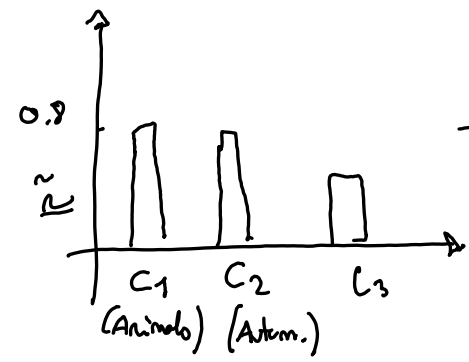
Multidimensional feature representation → matrix of activations for each image

COMPARISON → compute Pearson correlation between human and DNNs similarity matrix



F_{DNN} (n×120×120) and F_{human} (120×120)



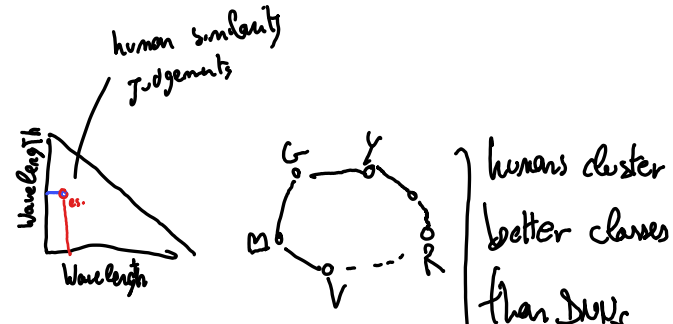


→ very good approximation (VGG is the best)

OTHER METHODS:

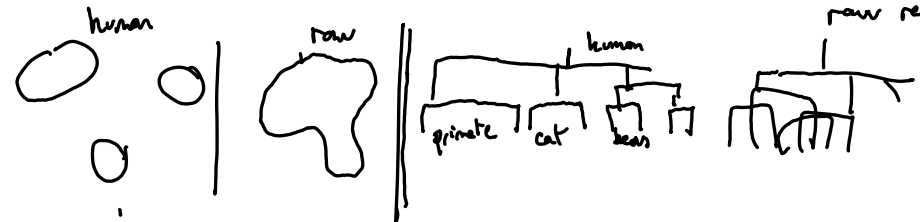
used for better understanding how DNNs succeed and fail in representing psychological concepts

1. **NON-METRIC MULTIDIMENSIONAL SCALING:** (convert similarity into spatial representation)



humans cluster better classes than DNNs

2. **HIERARCHICAL CLUSTERING**



→ Try to improve alignment of DNNs representations

$$S = FWF^T \qquad Y_i = B_1 X$$

$$\begin{vmatrix} & \begin{bmatrix} a & & c \\ 0 & \beta_\gamma & \\ & & \varepsilon \end{bmatrix} \end{vmatrix}$$

$$S_{ij} = \sum_k \omega_k f_{ik} f_{ik}$$

two images — feature

DNN's learn the correct basis set of features but with incorrect scale of saliency → reweighting improves a bit accuracy

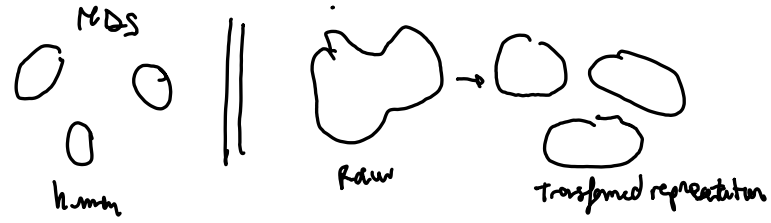
OVERFITTING PROBLEM

$\frac{120 \times 120}{2} \simeq 7000$ samples against 1024 parameters

↓

→ Feature selection
→ regularized L2 ($w^T w$)

→ GOOD RESULTS → but we are weighting dot product ⇒ NOT LINEAR MAPPING!



Conclusion: reweighting standard AI approaches is sufficient to obtain a good approximation

↓

They then asked themself if they can use the same W for different categories

IMPROVING ISOMORPHISM

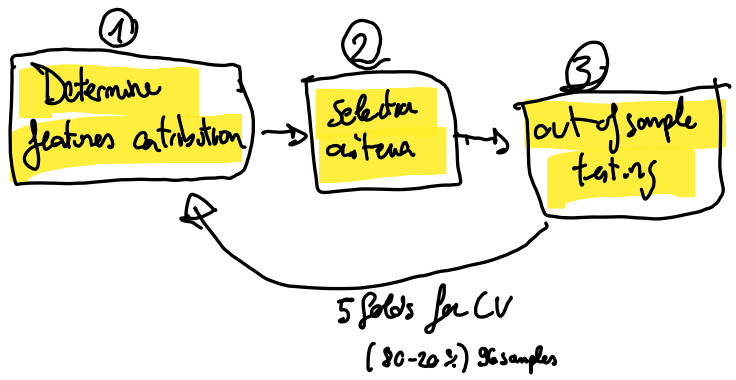
(Pruning vs Reweighing)

IDEA: DNNs do acquire the correct saliency of the best features but they're diluted by irrelevant ones

[es: considering every voxel for the task]

HYPOTHESIS: DNN's of features embeddings approximate HSS better than 2nd order isomorphism (Pearson Correlation)

& feature pruning by ranking can improve the fit between DNNs and human



(1) Done by removing each feature once at the time and retrain → 4096 features → RANKING
Comparing score with a baseline

(2) Add one feature at the time from scratch based on ranking and compute RDM

(3) Prune the DNN based on features → check activation in pruned layers

Check with different architecture → PRUNED outperforms the other architectures (R^2) (LASSO, Sim-DR, PAC-18...)

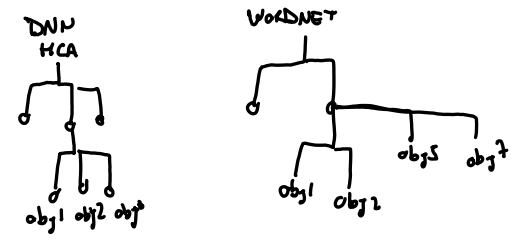
Problems

→ Pruning with 2D1 reflects ranking of pairwise similarities but does not directly address the hierarchical structure

↓

sol: Compare the H. structure latent in DNNs to WordNet

(1) Produce clusters from DNNs
(2) Define a Neighbours for each leaf
(3) Jaccard distance (fitness of neighbours)



$IoU = \frac{A \cap B}{A \cup B} = \frac{(1,2)}{(1,2,3,5,7)}$

how the net is capable of learning better hierarchy

Another test: PRUNE from brain similarity rather than human similarity

DATASET: 144 images with brain activity for each one (FMRI) → combined in 8 brain areas

Result → Pruned DNN performs a lot better than raw

↓

Why?

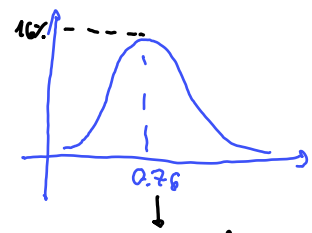
Hu et. al

- Overfitting → STUDIED WITH THE NOTION OF AP_{02}

APOZ: Average percentage of zeros of a single neuron → % of Ø activations of that neuron after the ReLU mapping

$$\frac{\sum_{k}^{N}\sum_{j}^{M} f\left(O_{c,j}^{(i)}(k)=0\right)}{N+M}$$

N: samples, M: features



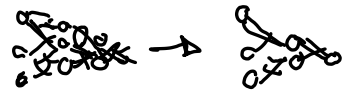
ex: 10% of neurons have 0.76 APOZ

↓

Consider the entire curve!!

ABLATION → test → fine tuning → test

Find the useless nodes and prune them!



68% → 69%

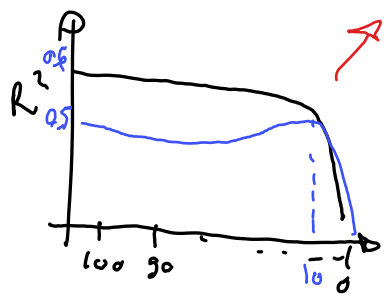
→ then retrain (fine tuning)

→ They then add one feature at the times according to their APOZ

H to L → very good R^2

L to H → good R^2 → This means that informations are redundant, so some neurons fire with low freq. bc others do the same task

→ Then they try to remove H → L



Prune is good even to approximate human 55

↓

% nodes remain