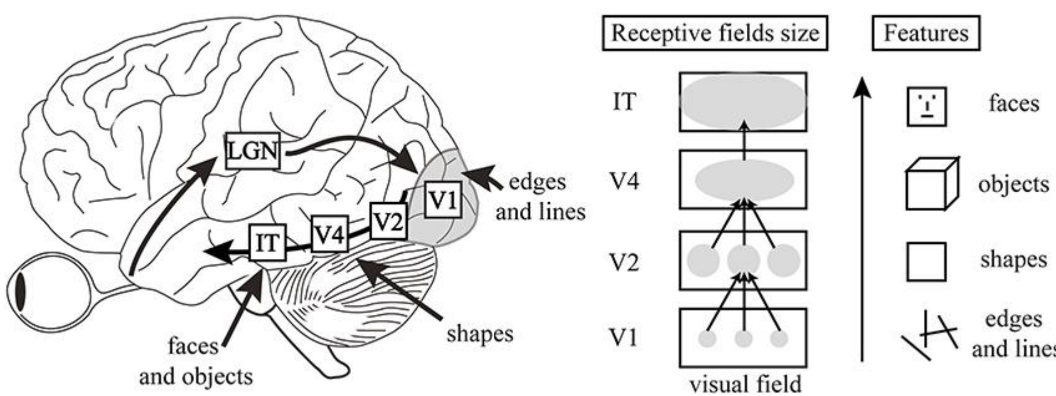
THEORY OF MACHINE LEARNING

LECTURE 18

NEURAL NETWORKS – REPRESENTATION BASICS

RECAP

- Perceptron, or linear threshold
  - Hypotheses of form sign($\langle a, x\rangle$) for an appropriate weight vector *a*
  - Generally, $\sigma(a^T x)$ for some "activation function" σ
- Biologically inspired, arithmetic circuit (with threshold gate)
- Idea behind neural nets:
  - Perceptrons detect "basic" or "primitive" features; 'composing' them allows for complex decision-making
  - Supported by human visual system (V1, V2, …)



RECAP: ARTIFICIAL/DEEP NEURAL NETWORK (DNN)

- **Definition.** A layered "circuit" that takes a vector of input features x, produces output y = $F_r \circ F_{r-1} \circ \cdots \circ F_1(x)$, where each F_i is a function of the form $F_i(z) = \sigma(Az + b)$, for some activation function $\sigma()$ (that acts coordinate-wise)

- Common activation functions:
  - Threshold
  - Sigmoid: (continuous approx.) $\frac{1}{1+e^{-x}}$
  - ReLU, Tanh
  - …

BASICS

- Neural networks are basically a (fairly complex) hypothesis class – takes input x, produces y
- **Question (vanilla supervised learning):** given data $(x_1, y_1), (x_2, y_2), \ldots$ from some distribution D, find *h* in this class that minimizes the risk
- ERM problem usually called neural network "training" – given data, find best hypothesis ($f(x_i) = y_i$) for all *i*

THEORY OF DEEP LEARNING

- Expressibility
  - What kinds of functions can be obtained using a DNN?
- Training complexity & training dynamics for GD and variants
  - Can the ERM problem be solved efficiently? What guarantees are possible?
- Generalization
  - What kind of generalization bounds can we prove? (VC dimension?)

Key: "easy" answers for all questions, but unsatisfactory for realistic settings

EXPRESSIBILITY BASICS

- **Barron's theorem [93].** *Any continuous function f that satisfies an appropriate "niceness" condition (parametrized by C) can be approximated to error ϵ (in L2!) by a 2-layer NN with $\sim \frac{C^2}{\epsilon}$ internal nodes*
- (Nice functions can be approximated by small NNs)
- **Universal approximation** [Cybenko, Hornik '87,'91]. *Any continuous function (over a compact domain) can be approximated by a 2-layer NN with any non-linearity (not a polynomial)*

But wait.. who uses infinitely wide 2 layer nets?

DEPTH VERSUS WIDTH

- Practical intuition:
  - Depth allows "meaningful features" while width is for "brute force memorization"
- Universality results degrade rapidly with dimensions
  - Curse of dimensionality
  - Modern nets work with high dimensional data
- Does higher depth lead to higher expressibility (with much fewer neurons)?
- Bunch of works … [Eldan and Shamir (depth 2 vs depth 3)], [Telgarsky], 2015-16

POWER OF DEPTH

Theorem template. There exists a network of depth D and "size" S that computes some function f that cannot be approximated by the output of *any* network with depth d and size S'   (typically if $d \ll D$, S' will be $\gg S$)

- "Depth versus width" results
- Reminiscent of circuit complexity (original work of Minsky, Pappert)

[Telgarsky 16]. For any $k>0$, theorem holds with:

$D = S \sim k^3$, $d = k$, and $S' = 2^k$ (and ReLU activations)

PROOF OUTLINE

- Consider just one-dimensional inputs and ReLU activations
- **Key insight:**
  - depth D lets us achieve exp(D) many "osciallations" in f
  - getting so many osciallations with depth d requires huge width!