REPRESENTATION LEARNING

- A GOOD REPRESENTATION IS ONE THAT KEEPS THE INFORMATION AND MAKES FURTHER LEARNING EASIER
- NATURALLY OCCURS WHEN DOING SUPERVISED TRAINING OF DEEP MODELS BUT NOT YET GREAT ON UNSUPERVISED SETTINGS
- PURELY OUT OF UNLABELED EXAMPLES – SEMI-SUPERVISED LEARNING – MANY TASKS SHARING ALL / PARTS OF REPRESENTATION
- TRAINING TASKS W/ ENOUGH LABELED EXAMPLES, TEST TASKS W/ VERY FEW LABELS – TEST TASK SIMILAR BUT DIFFERENT FROM TRAINING TASK

GREEDY LAYERWISE UNSUPERVISED PRETRAINING

EACH LAYER IS PRETRAINED VIA UNSUP. LEARNING USING THE OUTPUT OF THE PREVIOUS LAYER → SIMPLER (HOPEFULLY) DISTRIBUTION. SIMPLIFIES DIFFICULTY OF TRAINING DEEP MODELS.

- GREEDY BECAUSE LAYERS NOT JOINTLY TRAINED, LOWER LAYERS NOT ADAPTED • PRETRAINING BECAUSE LATER, OTHER STAGE, WE DO JOINT TRAINING TO **FINE-TUNE** LAYERS.
- CAN BE SEEN AS PARAM INITIALIZATION SCHEME / REGULARIZER
- AFTER LAST LAYER IS PRETRAINED: STACK A **SUPERVISED LAYER ON TOP** AND FINETUNE • CAN ALSO USED TO INIT ULTIMATELY UNSUPERVISED MODELS
- CAN ALSO HAVE GREEDY SUPERVISED PRETRAINING, FOR OPTIMIZATION OF DEEP SUPERVISED NETS
- **WHY DOES IT WORK?**

  NOT REALLY BETTER IF VERY LARGE LABELED DATASETS AVAILABLE. TRAJECTORY STUDIES DONE
  - TRAJECTORIES DO NOT CONVERGE, END UP IN APPARENT LOCAL MINIMA, BUT DERIVATIVES CLOSE TO SADDLE POINT
  - COVERED REGION OF SPACE SHRINKS WITH MORE TRAINING ITERATIONS, GROWS WITHOUT USPT. LARGE REGION IS BAD FOR GENERALIZATION
  - WITH VS WITHOUT RESULTS IN VERY DIFFERENT FUNCTIONS; NON-OVERLAPPING REGIONS
- BETTER PRETRAINING RESULTS WITH DEEPER MODELS
- SEEMS LIKE IT ACTS AS REGULARIZER → IMPLICIT PRIOR THAT $P(y|x)$ AND $P(x)$, THERE BECAUSE INTERMEDIATE REPRESENTATIONS, SHARE STRUCTURE
- **DISADVANTAGE:** DIFFICULT TO PICK CAPACITY HYPERPARAMS, MAY REQUIRE LARGER REPRESENTATIONS THAN WITHOUT
- NOT SUPER POPULAR TODAY WITH VERY LARGE LABELED DATASETS, REG'D WITH DROPOUT; STILL AN IMPORTANT TOOL ESP W/ SEMI-SUPERVISED, TRANSFER LEARNING, DOMAIN ADAPTATION

TRANSFER LEARNING / DOMAIN ADAPTATION

WHAT HAS BEEN LEARNED IN ONE SETTING (P_1) IS EXPLOITED TO IMPROVE GENERALIZATION IN (P_2)

TRANSFER LEARNING: MANY OF FACTOR OF VARIATION IN P_1 ARE RELEVANT IN P_2. EXAMPLE: VISUAL CATEGORIES DIFFERENT IN P_1, P_2
- SHARE THE **LOWER LAYERS** OF THE ACTUAL NETWORK, LOW-LEVEL VISUAL FEATURES WOULD BE SIMILAR, CORNERS, EDGES, ...
- SHARE THE **UPPER LAYERS** WHEN OUTPUT FORM IS SAME BUT TASK-SPECIFIC PRE-PROCESSING, IE SPEAKER RECOGNITION FROM DIFFERENT PEOPLE

DOMAIN ADAPTATION: SAME TASK, BUT SLIGHTLY DIFFERENT INPUT DISTRIBUTIONS, IE ONLINE COMMENTS | REVIEWS ABOUT DIFFERENT KINDS OF PRODUCTS BECAUSE DIFFERENT TONE / VOCABULARY. DENOISING AUTOENCODERS W/ UNSUP PRETRAINING ARE VERY GOOD FOR THIS.

CONCEPT DRIFT: DATA DISTRIBUTION GRADUALLY SHIFTS OVER TIME.

- DEEPER MODELS → BETTER EFFECTS OF ~~DOM~~ TRANSFER LEARNING; LEARNING CURVE FOR P_2 GOES FASTER → LESSER EXAMPLES NEEDED FOR GOOD GENERALIZATION
- **ONE-SHOT LEARNING:** ONLY ONE EXAMPLE OF NEW TASK IS GIVEN. NEW TASK IS VERY SIMPLE REGION
- **ZERO-SHOT | ZERO-DATA LEARNING:** NO EXAMPLE OF NEW TASK GIVEN, SUCCESSFUL WHEN 'CONTEXT' INFORMATION HAS BEEN GIVEN IN TRAINING, 'TASK' INPUT AS EXTRA. CAN DO INFERENCE ON 'SIMILARITY' BETWEEN TASK SIGNALS. **MACHINE TRANSLATION:** AUTO-RELATE WORD PAIRS BECAUSE WE LEARNED COMMON SEPARATELY, DISTRIBUTIONS; AND FORMED LINK WITH SOME TRANSLATION EXAMPLES → NEW PAIRS ARE AUTO-LINKED UP
- **MULTI-MODAL LEARNING:** SAME THING BUT DIFFERENT DOMAIN MODALITIES FOR REPRESENTATION.

SEMI-SUPERVISED LEARNING

COMBINING UNLABELED EXAMPLES FROM $P(x)$ TO LABELED (x, y) → ESTIMATION OF $P(y|x)$. USING UNSUPERVISED TECHNIQUES WE MAP x_1, x_2 IN NEARBY LOCATIONS OF SPACE OR SAME CLUSTER TO HAVE SIMILAR EMBEDDINGS. THEN WE USE SUPERVISED LEARNING. **EXAMPLE:** PCA PRE-PROCESSING, CLASSIFIER ON PROJECTED DATA

- WE CAN SUP AND UNSUP COMPONENTS TO SHARE PARAMETERS, USE UNSUPERVISED OR GENERATIVE CRITERION. – IMPLIES PRIOR OF $P(x)$, $P(y|x)$ SHARE STRUCTURE
- CONTROL HOW MUCH GEN. CRITERION → BETTER PERFORMANCE
- **IN DL:** INTRODUCE UNSUPERVISED EMBEDDING CRITERION AT EACH LAYER + TOTAL SUPERVISED CRITERION. ALTERNATIVE TO UNSUPERVISED PRETRAINING.
- **WHEN DOES SSL WORK?**

  IDEAL REPRESENTATION → DISENTANGLES UNDERLYING FACTORS OF VARIATION WITHIN THE DATA, **WHEN** $P(y|x)$ SEEN AS FCN OF x HAS SOMETHING TO DO WITH $P(x)$
  - **COUNTEREXAMPLE:** $P(x)$ IS UNIFORMLY DISTRIBUTED
  - **COUNTER-COUNTEREXAMPLE:** x IS FROM A MIXTURE; ONE COMPONENT PER VALUE OF y. IF COMPONENTS WELL-SEPARATED → $P(x)$ ALONE TELLS US EVERYTHING, A SINGLE EXAMPLE ~~OF n~~ WILL BE ENOUGH TO LEARN $P(y|x)$
- WHEN y IS CLOSELY ASSOCIATED WITH ONE OF CAUSAL FACTORS OF x. WE CAN'T KNOW BEFOREHAND WHICH ONE WILL IT BE → SO DISENTANGLE THEM ALL!
  - → IF TRUE PROCESS HAS y AS CAUSE OF x, $P(x|y)$ IS ROBUST TO CHANGES IN y. REVERSE NOT TRUE! **GENERALLY:** CAUSAL MECHANISMS OF STUFF GENERALLY REMAIN INVARIANT, SO A GENERATIVE MODEL FOR h AND $P(x|h)$ IS ALWAYS GOOD

DISTRIBUTED REPRESENTATION

- **DISTRIBUTED REPRESENTATION:** IS ONE EXPRESSING AN EXPONENTIALLY LARGE NUMBER OF CONCEPTS BY ALLOWING TO COMPOSE THE ACTIVATION OF MANY FEATURES
  → VECTOR OF n BINARY FEATURES. TOTAL 2^n CONFIGS. POTENTIALLY A REGION IN SPACE
- **SYMBOLIC REPRESENTATION:** N SYMBOLS, N DETECTORS, N REGIONS OF SPACE. ONE-HOT ENCODING, CLUSTERING, kNN; GMM, KERNEL MACHINES; N-GRAMS, EV. WITH INTERPOLATION BUT STILL SYMBOLIC. **PRO:** ANSWERS CAN BE INDEPENDENTLY CHOSEN FOR EACH REGION
  CON: NO GENERALIZATION TO NEW REGIONS EXCEPT FOR EXTENDING WITH SMOOTHNESS PRIOR.
- IN **DR**, GENERALIZATION ARISES FROM SHARED ATTRIBUTES BETWEEN CONCEPTS. THEY INDUCE A RICH SIMILARITY SPACE. GENERALIZE BY STRUCTURE
- A **SPARSE REPRESENTATION** IS A DR WHERE NO OF ACTIVE ATTRIBUTES IS SMALL COMPARED TO THE TOTAL

- EFFICIENCY

DR EXPLOIT / LEARN, AND GENERALIZE BY STRUCTURE. SR ONLY BY SMOOTHNESS → SUFFERS FROM COD = WE NEED NO. OF EXAMPLE AT LEAST AS LARGE AS THE NO. OF REGIONS

EXAMPLE: REGULAR REPETITIONS / PATTERNS. N FEATURES, D-SPACE. HOW MANY REGIONS IN N HYPERPLANES IN $\mathbb{R}^d$? $\sum_{j=0}^{d}\binom{n}{j} = O(n^d)$ EXPONENTIAL GROWTH IN INPUT SIZE, POLYNOMIAL IN NO ~~FEAT~~ HIDDENS.

- DR IS A PRIOR ON TOP OF SMOOTHNESS PRIOR. WE CAN LEARN HYPERPLANES WITH $O(d)$ EXAMPLES
- PRODUCT OF MIXTURES (RBM), MIXTURE OF PRODUCTS (GMM). GMM TAKES EXP. MANY EXAMPLES TO LEARN RBM DISTRIBUTION

EFFICIENCY FROM DEPTH

UNIVERSAL APPROXIMATION THEOREM. **OK - BUT!** THERE ARE FAMILIES OF FCNS WITH EFFICIENT REPRESENTATIONS W/ K LAYERS REQUIRING EXPONENTIAL NO OF COMPONENTS WRT INPUT SIZE AT INSUFFICIENT DEPTH.

→ **EXAMPLE:** SUM-PRODUCT NETWORKS. DEEP REPRESENTATION ALLOWS PARTIAL RESULTS TO BE REUSED |DEPTH| MANY TIMES.

→ **EXAMPLE:** DEEP RECTIFIER NETWORKS (RELU, MAXOUT) CAN ENCODE NUMBER OF REGIONS $O\left(\binom{n}{D}^{D(L-1)} n^D\right)$. D INPUTS L DEPTH N UNITS PER LAYER
→ EXPONENTIAL IN DEPTH

PRIORS ON UNDERLYING FACTORS

BROAD ASSUMPTIONS THAT CAN HELP THE LEARNER. **NO PRIORS → NO GENERALIZATION**. INDUCTIVE BIAS / NFL THEOREM.

- **SMOOTHNESS** $x \approx y \rightarrow f(x) \approx f(y)$ KILLED BY COD
- **MULTIPLE EXPLANATORY FACTORS:** WHAT IS LEARNED ON ONE FACTOR GENERALIZES FOR OTHERS. DISTRIBUTED REPRESENTATION
- **DEPTH** CONCEPTS CAN BE DEFINED IN TERMS OF OTHER CONCEPTS, HIERARCHY OF ABSTRACTION. DEEP REPRESENTATIONS
- **CAUSAL FACTORS** INPUT x ARE CONSEQUENCES; h ARE CAUSES. ENABLES SEMI-SUPERVISED LEARNING.
- **SHARED FACTORS:** SAME x, DIFFERENT y_i (TASKS). TASKS RELY ON DIFFERENT SUBSETS OF h_i (COMMONS). ALLOWS TRANSFER LEARNING
- **MANIFOLDS** PROBABILITY MASS CONCENTRATES IN LOCALLY CONNECTED REGIONS OF SMALL VOLUME. IMPORTANT IN AUTOENCODERS
- **NATURAL CLUSTERING** DIFFERENT CATEGORICAL VARS → SEPARATE MANIFOLDS. DIFFERENT CLASSES SEPARATED BY REGIONS OF LOW MASS
- **TEMPORAL AND SPATIAL COHERENCE** DIFFERENT FACTORS CHANGE AT DIFFERENT SPATIAL / TEMPORAL SCALES. MANY CATEGORICAL CONCEPTS CHANGE SLOWLY
- **SPARSITY** FOR ANY OBSERVATION x, ONLY A SMALL FRACTION OF FACTORS ARE RELEVANT
- **SIMPLICITY OF FACTOR DEPENDENCIES** GOOD REPRESENTATION ⇄ FACTOR RELATED THROUGH SIMPLE RELATIONS. IE CONDITIONAL / MARGINAL INDEPENDENCE
  BASELINE FOR STACKING LINEAR PRED OR FACTORIZATION ON TOP OF LEARNED REPRESENTATION.