CONVOLUTIONAL NETWORKS

ARE ANY NETWORKS EMPLOYING CONVOLUTION INSTEAD THAN MATRIX MULTIPLICATION

CONVOLUTION: $s(t) = \int x(a)\,w(t-a)\,da$ x = INPUT, w = KERNEL, s = FEATURE MAP **DISCRETE:** $s[t] = \sum_{a=-\infty}^{+\infty} x[a]\,w[t-a]$ • IN ML: KERNEL IS LEARNABLE PARAMS AND A PDF, TENSOR

2D: $S[i,j] = \sum_m \sum_n I[m,n]\,K[i-m, j-n] = \sum_m \sum_n I[i-m, j-n]\,K[m,n] = \sum_m \sum_n I[i+m, j+n]\,K[m,n]$ • FLIPPED BY ELEMENT PRODUCT 'BY OVERLAPPING'

COMMUTATIVITY — CROSS-CORRELATION: SAME AS CONVOLUTION BUT KERNEL NOT FLIPPED

REASONS FOR HAVING CONVOLUTION IN ANNS.

- **SPARSE CONNECTIVITY | INTERACTIONS** KERNELS ARE << INPUT MATRIX. EVERY ORIGINAL INPUT INTERACT WITH SMALL SET OF HIDDEN (KERNEL) NODES, ITS WIDE
  • LOT LESS STORAGE SPACE • FASTER PROCESSING • UNITS IN DEEP LAYERS INTERACT INDIRECTLY WITH MOST INPUT STILL.
- **PARAMETER SHARING** AKA TIED WEIGHTS. SAME KERNEL IS USED AT EVERY INPUT POSITION. MORE STORAGE CAPACITY REDUCTION. DRAMATICALLY MORE EFFICIENT
- **EQUIVARIANCE TO TRANSLATION:** IF INPUT CHANGES, OUTPUT CHANGES SAME WAY. $f(x)$ EQUIVARIANT TO g IF $f(g(x)) = g(f(x))$ } BILLIONS FOR STORAGE, 10.000 COMPUTATIONALLY FOR 100 x 100 IMG
  $g(x)[i] = x[i-1]$ SAME REGARDLESS OF ORDER OF APPLICATION. VERY USEFUL DETECTING PROPERTY
- **CONVOLUTIONAL LAYER:** CONVOLUTIONAL STAGE (AFFINE TRANSFORM) → DETECTOR STAGE (NONLINEARITY) → POOLING (if) OR CONSIDER EACH STAGE A LAYER
  - ALL STAGES HAVE SAME KERNEL TENSOR IE RELU

POOLING

REPLACES OUTPUT OF NONLINEARITIES WITH A SUMMARY STATISTIC OF NEARBY OUTPUTS. POOLING TYPES: MAX, AVG, L2 NORM, WEIGHTED AVG FROM CENTER (ON RECTANGULAR NEIGHBORHOOD)

- MAKES REPRESENTATION **INVARIANT TO SMALL TRANSLATIONS** → POOLED OUTPUTS WON'T CHANGE. WE CARE MORE ABOUT PRESENCE OF FEATURE THAN ITS LOCATION.
- IF WE POOL OVER OUTPUTS OF DIFFERENT KERNELS WE CAN MAKE INVARIANCE TO DIFFERENT TRANSFORMATION (ROTATED DIGITS EXAMPLE)
- FILTERS WITH STRIDE > 1 RESULT IN DOWNSAMPLING → EVEN MORE REDUCTION IN COMPUTATIONAL AND STORAGE REQUIREMENTS
- MAY POOL STUFF DYNAMICALLY, RUN CLUSTERING, OR LEARN THE POOLING STRUCTURE ITSELF
- POOLING CAN BE TRICKY FOR NETS WITH TOP/DOWN INFO; BOLTZMANN MACHINES AND AUTOENCODERS
- CAN BE SEEN AS INFINITELY STRONG PRIOR OVER WEIGHTS FOR A FULLY CONNECTED NET SPECIFYING ~~THEN~~ PARAMS MUST BE TIED, WEIGHT ZERO EXCEPT FOR KERNEL REGION, ETC

CONVOLUTION FUNCTION VARIANTS

- IN ANN WE DO MANY CONVOLUTIONS IN PARALLEL ON A LAYER. DIFFERENT FEATURES, MULTIPLE CHANNELS (RGB) • IMPLEMENTATION WITH 4D KERNEL TENSOR: SPATIAL X, Y, CHANNEL INDEX IN BATCH.
- OPS COMMUTATIVE IF INPUT CHANNELS = OUTPUT CHANNELS • **STRIDE:** SKIP SOME POSITIONS TO SAVE RESOURCES, SAMPLE EVERY S PIXELS
- **ZERO PADDING** TO AVOID LOSING MUCH INFO OVER MANY CONVOLUTIONAL LAYERS BECAUSE VALID CONVOLUTION $(m \times m)(n \times n) \to m-n+1 \times m-n+1$ VERY BAD IF IMAGE KERNELS
  → USED TO EITHER MAKE OUTPUT NOT SHRINK (**SAME**) OR TO MAKE EVERY PIXEL VISITED k TIMES IN EACH DIRECTION (**FULL**)
  → PLAY WITH SETTING
  $m+n-1 \times m+n-1$ x OUTPUT
- **LOCALLY CONNECTED LAYERS:** 6D TENSOR. DIFFERENT KERNEL PER PATCH. 'UNSHARED CONVOLUTION' USEFUL WHEN FEATURES ARE OF SMALL PATCH BUT DIFFERENT FEATURES AT DIFFERENT LOCATIONS. IE IN A FACE WE WANT LOOK FOR MOUTH AT THE TOP.
- **CHANNEL-PLAY** CONNECT EACH OUTPUT CHANNEL TO SUBSET OF INPUT CHANNELS. FURTHER EFFICIENTATION.
- **TILED CONVOLUTION:** 'COMPROMISE' BETWEEN CONVOLUTIONAL AN LOCALLY CONNECTED LAYER. LEARNS A SET OF KERNELS WE ROTATE AS WE MOVE THROUGH SPACE
  IF FILTER STACK IS SAME FEATURES TRANSFORMED AND THEN WE MAX POOL → BAM! ARBITRARY TRANSFORMATION INVARIANT. LEARN THROUGH

HOW TO DO BACKPROP THROUGH CONVOLUTION.

↑ RECONSTRUCTION FUNCTION

MULTIPLY BY THE TRANSPOSE OF MATRIX DEFINED BY CONVOLUTION (ALSO IN AUTOENCODERS, RBM). CAREFUL BECAUSE NEEDS TO TAKE PADDING POLICY INTO ACCOUNT; ALSO STRIDE AND SIZE CONCERNS. • CONVOLUTION • BPROP TO WEIGHTS • BPROP TO INPUT IS ALL ITS NEEDED TO TRAIN CONVNETS

- **BIASES:** ONE PER CHANNEL, SHARED | ONE PER STACK | DIFFERENT ONES

STRUCTURED OUTPUTS

WHEN WE WANT TO OUTPUT A HIGH-DIM STRUCTURED OBJECT → IE A PIXEL SEGMENTATION MAP. VECTOR OF SCORES (CATEGORY PROBS) PER PIXEL/PATCH

- **REMOVE FULLY CONNECTED TOP LAYERS:** GETS A SPATIALLY STRUCTURED OUTPUT, TRAIN WITH TARGET LABELS, APPLY SOFTMAX. LAST CONV LAYER MUST HAVE SAME DIMS AS NO. OUTPUT CATEGORIES
  - → WE MIGHT WANT TO REMOVE POOLING, OR SHARE KERNEL WEIGHTS ACROSS LAYERS (RESULTS IN SPECIAL RNN)
  - → THEN USE CONVNET OUTPUT TO TRAIN/FIT A CRF FOR INSTANCE (STRONG PRIOR NEIGHBORS HAVE SAME VALUE) REGIONWISE

DATA TYPES

	SINGLE CHANNEL	MULTI CHANNEL
1D	AUDIO WAVEFORM	SKELETON ANIMATION DATA
2D	AUDIO IN FOURIER DOMAIN → SPECTROGRAM	COLOR IMAGE DATA
3D	VOLUMETRIC DATA, CT SCAN	COLOR VIDEO DATA

- CONVNETS CAN BE USED FOR SOURCE W/ DIFFERENT SIZE BY MAKING # POOLING PROPORTIONAL TO INPUT SIZE.

CONVOLUTION ALGORITHMS

- CONVOLUTION IS MULTIPLICATION IN FOURIER DOMAIN. A LOT BETTER AND FASTER THEN DOING IT EXPLICITLY WITH OUTER PRODUCTS
- IF D-DIMENSIONAL KERNEL IS OUTER PRODUCT OF D VECTORS, KERNEL IS SEPARABLE → NAIVE CONVOLUTION INEFFICIENT
  - → **COMPOSE D 1-DIM CONVOLUTIONS** WITH EACH VECTOR. $O(W \times D)$ VS $O(W^D)$ NOT EXACTLY CONVOLUTION IS LIKE THIS.

RANDOM OR UNSUPERVISED FEATURES

- IF NO SUPERVISED TRAINING FOR FEATURES → RANDOM INITIALIZATION, OR LEARN UNSUPERVISED. THEN USE THEM (ONCE) FOR LAST CLASSIFIER LAYER(S)
- **RANDOM WORKS WELL** → NATURALLY CONVERGE TOWARDS FREQUENCY SELECTIVITY AND INVARIANCE.
- **USE GREEDY UNSUPERVISED PRETRAINING** → LAYER PER LAYER, LIKE MLP → **CONVOLUTIONAL DEEP BELIEF NET**
- TRAIN DENSE UNSUPERVISED MODEL OF IMAGE PATCH, THEN USE THEM AS KERNEL FOR CONVOLUTIONAL LAYER
  - → WE CAN TRAIN CONVNET **WITHOUT EVER USING CONVOLUTION** AT TRAINING → ONLY IN INFERENCE!

NEUROSCIENTIFIC BASIS FOR CONVNETS

FINDING FROM VISUAL SYSTEMS NEUROSCIENCE (HUBEL-WIESEL). EASY!! RETINA → OPTIC NERVE → LATERAL GENICULATE NUCLEUS → **V1 AREA** → V2 → V4 → IT (INFEROTEMPORAL CORTEX)

A CONVNET IS INSPIRED BY STUFF IN V1: 2D SPATIAL MAP OF RETINA, SIMPLE DETECTORS RECEPTIVE FIELDS (KERNELS), COMPLEX SHIFT-INVARIANT DETECTORS (POOLING)

- → **GRANDMOTHER CELLS** IN MEDIAL TEMPORAL LOBE FOR CONCEPTS. FIRE REGARDLESS OF SENSORY MODALITY. **IT:** CLOSEST ANALOG TO CONVNET LAST LAYER OF FEATURES
- → V1 CELLS EXHIBIT RESPONSE FUNCTIONS SIMILAR TO **GABOR FILTERS**: DESCRIBE 2D WEIGHTS AT IMAGE $W = \alpha \exp\left(-\beta_x x'^2 - \beta_y y'^2\right)\cos(fx' + \phi)$ PARAMS CONTROL TRANSLATIONAL, ROTATIONAL, FREQUENCY RESPONSE
  - $x' = (x - x_0)\cos(\tau) + (y - y_0)\sin(\tau)$ ROTATIONS
  - $y' = -(x - x_0)\sin(\tau) + (y - y_0)\cos(\tau)$
- MANY ML ALGOS LEARN GABOR-LIKE EDGE DETECTOR PATTERNS

CONVNET HISTORY

CHECK READING SYSTEM (LECUN). OCR HANDWRITING BY MICROSOFT. FIRST DEEP MODEL TRAINED WITH BACKPROP. TRIGGERED DEEP LEARNING TREND OF TODAY

THEIR EARLY SUCCESS PROBABLY DUE TO THEIR LARGE COMPUTATIONAL EFFICIENCY WRT FULLY CONNECTED NETS.