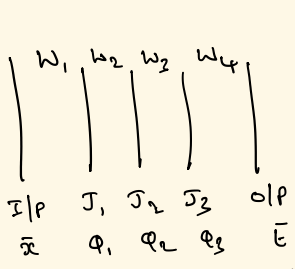
CNNs

Friday, 23 February 2024 11:25

Convolutional Neural Network (CNN)

MLFFNN :



Fully connected layers

I/p	J_1	J_2	J_3	o/p	
$\bar{x}$	ϕ_1	ϕ_2	ϕ_3	E	$\bar{y}$
d	$\bar{z}_1$	$\bar{z}_2$	$\bar{z}_3$	K ϕ^o	K

Hand-crafted features

Hand-picked features

$\bar{z}_1 = \bar{\Phi}_1(W_1\bar{x} + \bar{b}_1)$

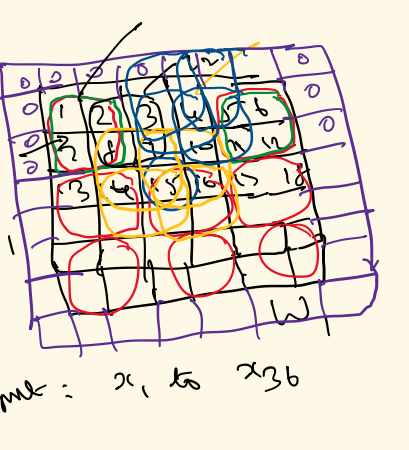
$\bar{z}_2 = \bar{\Phi}_2(W_2\bar{z}_1 + \bar{b}_2)$

$\bar{z}_3 = \bar{\Phi}_3(W_3\bar{z}_2 + \bar{b}_3)$

$\bar{y} = \bar{\Phi}^o(W_4\bar{z}_3 + \bar{b}_o)$

Global feature extraction   Raw data as input

Local feature extraction   Representation learning



Input : x_1 to x_{36}

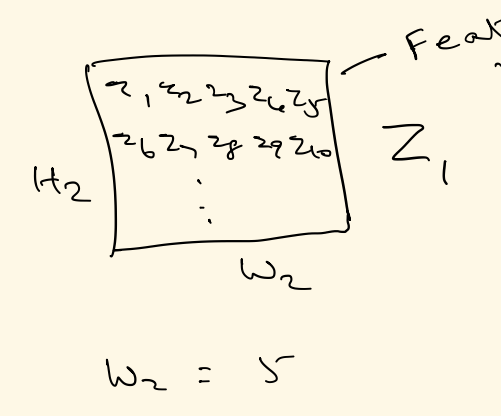
Filter size : $F \times F$ $F = 2$

Stride : $S = 1$ $S < F$

Weight sharing

Padding size : $P = 1$

w_1	w_2
w_3	w_4



$z_1 z_2 z_3 z_4 z_5$
$z_6 z_7 z_8 z_9 z_{10}$
H_2
W_2
Feature map
Z_1

$W_1 = 6$ $W_2 = 5$

$H_1 = 6$ $H_2 = 5$

$a_1 = w_1x_1 + w_2x_2 + w_3x_7 + w_4x_8 + b$

$z_1 = s_1 = \Phi(a_1)$

$W_2 = \left\lfloor \frac{\overset{6}{W_1} - \overset{2}{F} + 2P}{S} \right\rfloor + 1$

$H_2 = \left\lfloor \frac{H_1 - F + 2P}{S} \right\rfloor + 1$

$a_2 = w_1x_2 + w_2x_3 + w_3x_8 + w_4x_9 + b$

$z_2 = \Phi(a_2)$

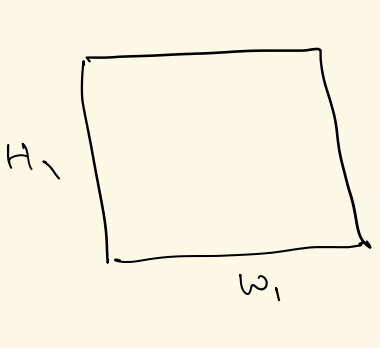
Multiple filters (kernels)

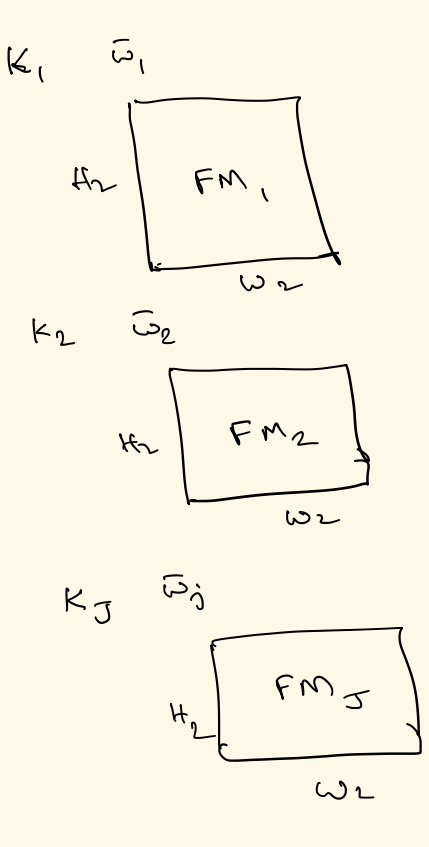
Multiply Shift $\rightarrow$ Convolution

Receptive field of a neuron : Part of the image given as input to a neuron

Size of receptive field : Size of filter or kernel.

2-d kernel $F \times F$





K_1 $\bar{w}_1$
H_2 FM_1 W_2
K_2 $\bar{w}_2$
H_2 FM_2 W_2
K_J $\bar{w}_J$
H_2 FM_J W_2

Convolution Layer (CL)

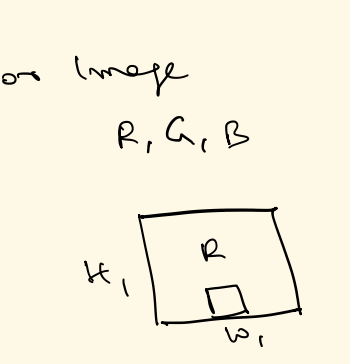
No. of parameters : $(F \times F + 1) \times J$

No. of connections : $W_2 \times H_2 \times (F \times F + 1) \times J$

Weight sharing

Color Image

R, G, B



H_1 R W_1
H_1 G W_1
H_1 B W_1

Input volume

$W_1 \times H_1 \times 3$

3 channels

$F = 2$

3-d kernel : $F \times F \times 3 \leftarrow d_1 = 3$

$x_1 \quad x_2 \quad x_3 \quad x_4$

x_{1R}	x_{2R}	x_{3R}	x_{4R}
x_{1B}	x_{2B}	x_{3B}	x_{4B}
x_{1G}	x_{2G}	x_{3G}	x_{4G}

$z = \underset{F \times F \times 3}{\bar{w}^t}\bar{x} + b$

$F \times F \times d_2$ ← depth of kernel

CL_1 J_1 FMs

$W_2 \times H_2 \times J_1$

Kernel : $F \times F \times d$ ← Depth of 2-d kernel

$d = 3$

Stride : S

Padding : P

$W_2 = \left\lfloor \frac{W_1 - F + 2P}{S} \right\rfloor + 1$

$H_2 = \left\lfloor \frac{W_2 - F + 2P}{S} \right\rfloor + 1$

No. of kernels : J_1

CL_2 J_2 FMs

$d_2 = 2$ Kernel : $F \times F \times d$

J_1c_2 $d = J_1$

J_1c_3 $d < J_1$

J_1c_4 $d = 2, 3, \ldots, J_1 - 1$

J_1c_d K_2 features

$J_1 = 4$

FM_1, FM_2, FM_3, FM_4

CL_2 : $d = 2$

Combinations of FMs : 1,2 1,3 1,4 2,3 2,4 3,4

K_2

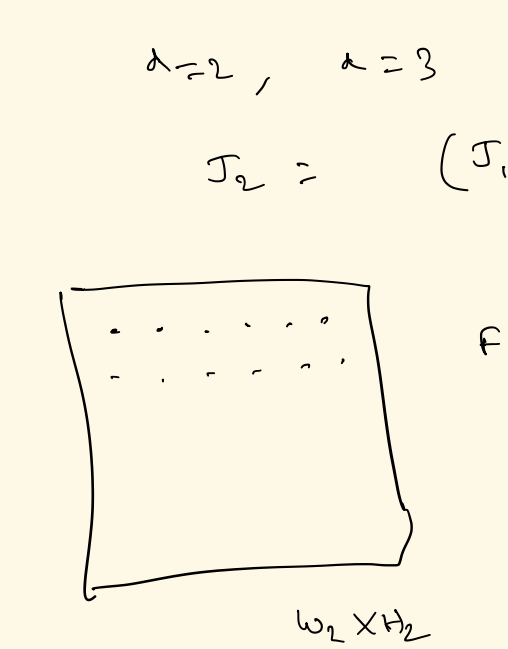
J_2 : $\underset{4c_2}{6} \times k_2$ $J_2 = J_1c_2 \times k_2$

$d = 3$

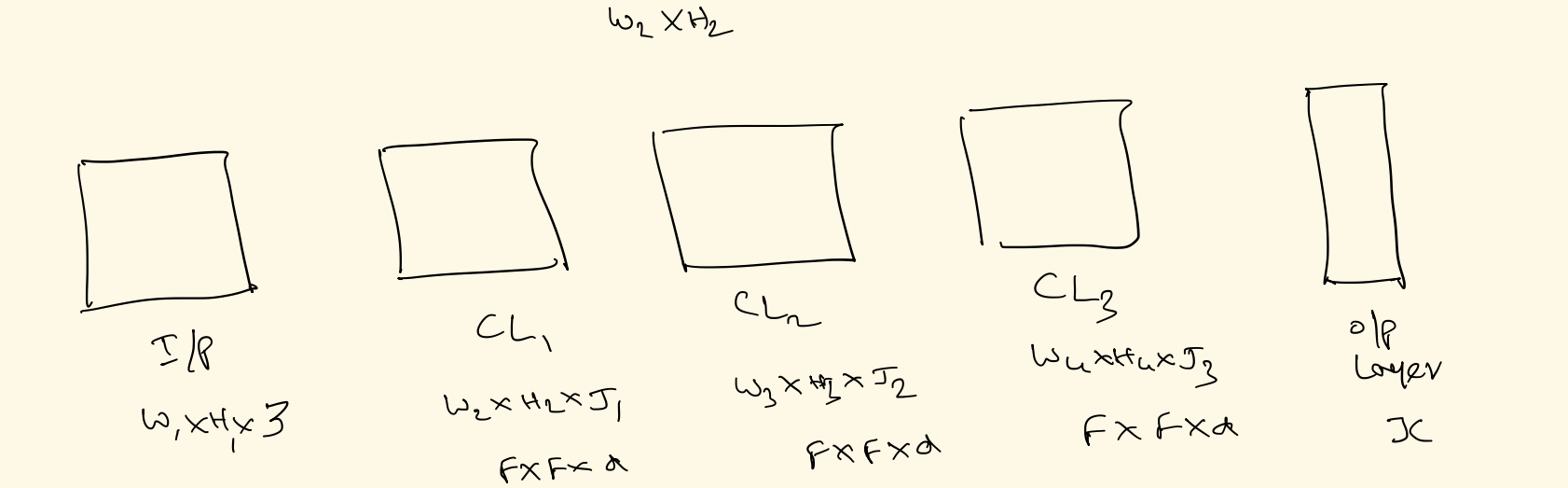
$J_1c_3 \times k_2$ $J_2 = J_1c_3 \times k_2$

$d = 2$, $d = 3$

$J_2 = (J_1c_2 + J_1c_3) \times k_2$



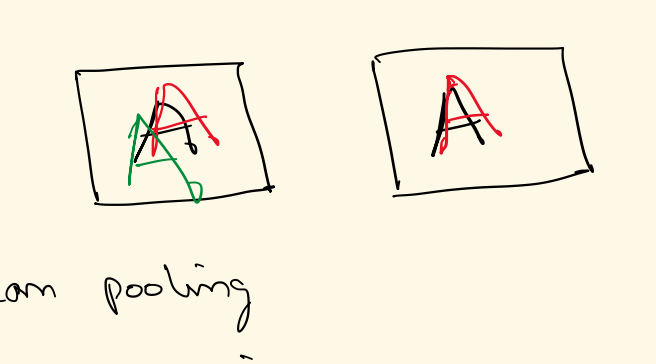
$F \times F \times d$



I/P
$W_1 \times H_1 \times 3$
CL_1
$W_2 \times H_2 \times J_1$
$F \times F \times d$
CL_2
$W_3 \times H_3 \times J_2$
$F \times F \times d$
CL_3
$W_4 \times H_4 \times J_3$
$F \times F \times d$
o/p Layer
K
Flattening
$W_4 \times H_4 \times J_3$

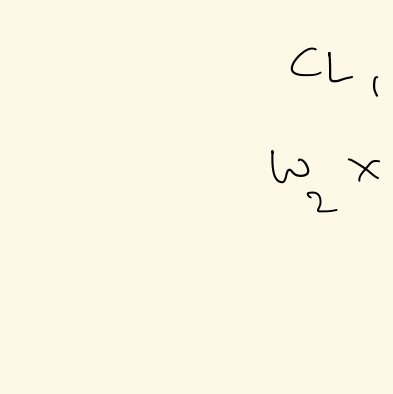
Pooling Layer

- After one or more CLs
- Reduce the size of CLs
- Translational invariance or shift invariance



Mean pooling

Max Pooling



$y_1 = \underset{\text{or max}}{\text{Mean}}(z_1, z_2, z_5, z_6)$

2×2

$S = F$

CL_1	PL_1	CL_2
$W_2 \times H_2 \times J_1$	$F = 2$ $W_3 \times H_3 \times J_1$	$W_4 \times H_4 \times J_2$
	$W_3 = W_2/F$	$W_4 = \left\lfloor \frac{W_3 - F + 2P}{S} \right\rfloor + 1$
	$H_3 = H_2/F$	$H_4 = \left\lfloor \frac{H_3 - F + 2P}{S} \right\rfloor + 1$

I/P Layer	CL	o/p Layer
$W_1 \times H_1 \times 1$	$W_2 \times H_2 \times J_1$	K

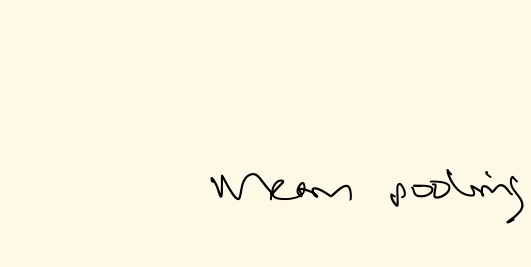
$\Delta w = -\eta \frac{\partial E}{\partial w}$

FM_1 : $W_2 \times H_2$ $\underset{F \times F}{\bar{w}_1}, b_1$ Kernel 1

Updated weights for nodes in FM_1 : $\bar{w}_{1l}, b_{1l}$, $l = 1, 2, \ldots, W_2 \times H_2$

Updated weights for shared weights $\bar{w}_1$ and $\bar{b}_1$ are obtained as average of $\bar{w}_{1l}$ and b_{1l}, $l = 1, 2, \ldots, W_2 \times H_2$, respectively

I/P Layer	CL	PL	O/P Layer
$W_1 \times H_1 \times 1$	$W_2 \times H_2 \times J_1$	$W_3 \times H_3 \times J_1$	K



FM_1
CL
FM_1
PL

Mean pooling

$F = 2$

$S = 2$

Δw_{jk} ← CL FM node, o/p Layer node

$\Delta w_{jk} = \eta\, \delta_k^o\, s_j^h \cdot r$ ← scale factor

Scale factor depends on the type of pooling used.

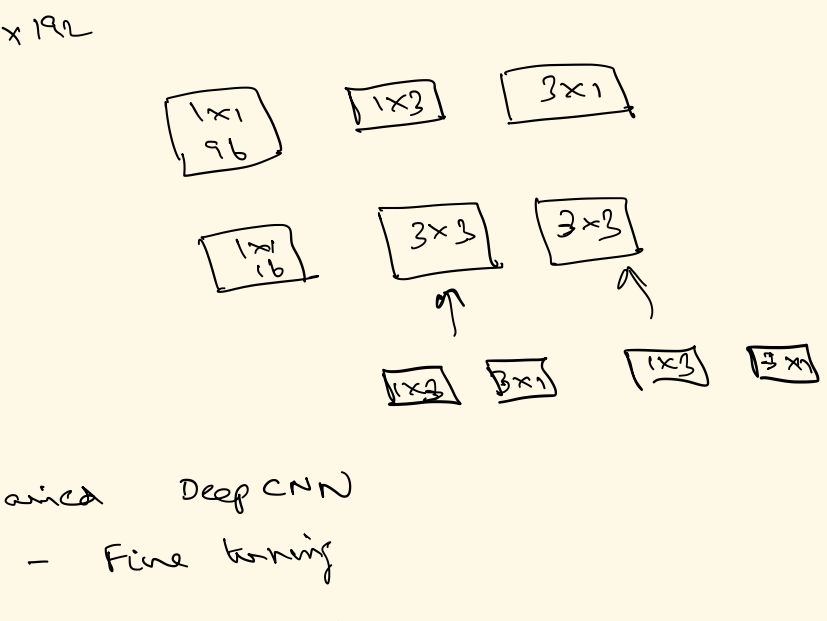
Mean pooling : $r = \frac{1}{4}$

Max pooling : There is only one node with maximum output → $\begin{cases} r = 1, & \text{for node with maximum output} \\ = 0, & \text{for other nodes} \end{cases}$

There are K nodes with maximum output $\begin{cases} r = \frac{1}{K}, & \text{for every node with maximum output} \\ = 0, & \text{for other nodes} \end{cases}$

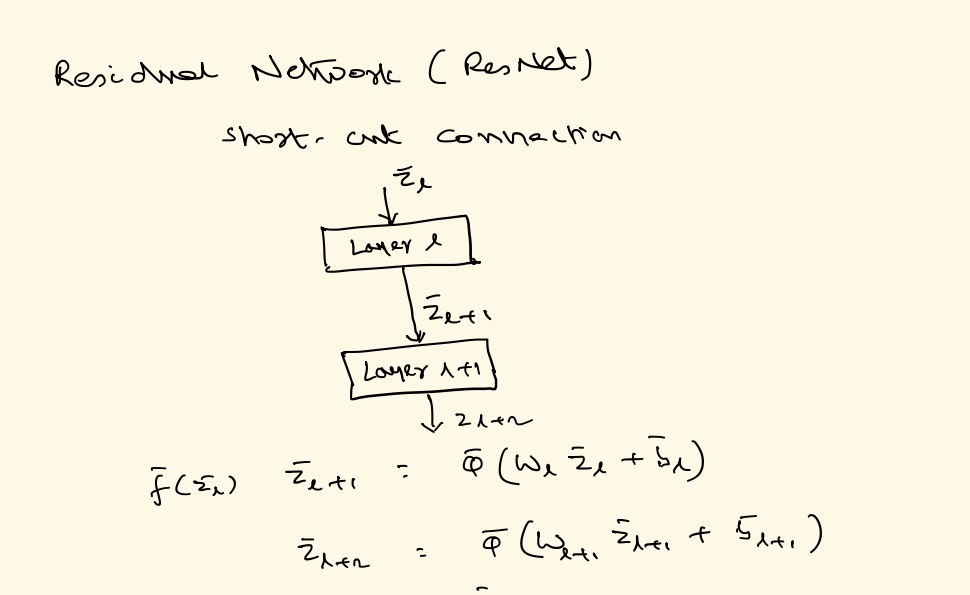
Kernel factorization

- Reduce number of parameters



5×5 Kernel
3×3 kernel
S=1
3×3 kernel
25 parameters.
9 parameters
9 parameters
18 parameters
3×3 kernel
1×3 kernel.
3×1 kernel
9 parameters
3 parameters
3 parameters
6 parameters

Inception module



28×28×192
1×1 96
3×3
1×1 16
5×5
1×1 96
1×3
3×1
1×1 16
3×3
3×3
1×3
3×1
1×3
3×1

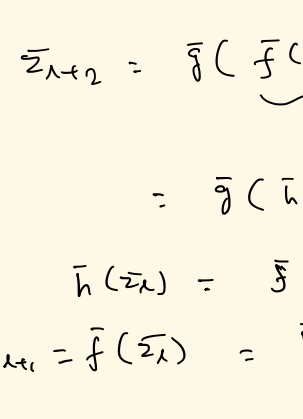
Pre-trained Deep CNN

- Fine tuning

Transfer Learning

Residual Network (ResNet)

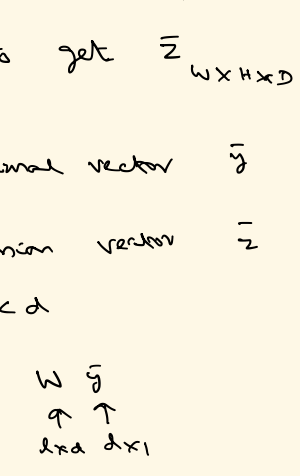
short-cut connection



$\bar{z}_l$
Layer l
$\bar{z}_{l+1}$
Layer $l+1$
z_{l+2}

$\bar{f}(\bar{z}_l)$ $\bar{z}_{l+1} = \bar{\Phi}(W_l\bar{z}_l + \bar{b}_l)$

$\bar{z}_{l+2} = \bar{\Phi}(W_{l+1}\bar{z}_{l+1} + \bar{b}_{l+1})$



$\bar{z}_l$
Layer l
$\bar{z}_{l+1}$
$\bar{f}(\bar{z}_l)$
Layer $l+1$
$\bar{z}_{l+2}$
Short-cut connection

$\bar{g}(\;)$

$\bar{z}_{l+2} = \bar{g}(\underbrace{\bar{f}(\bar{z}_l) + \bar{z}_l}_{\bar{h}(\bar{z}_l)})$

$= \bar{g}(\bar{h}(\bar{z}_l))$

$\bar{h}(\bar{z}_l) = \bar{f}(\bar{z}_l) + \bar{z}_l$

$\bar{z}_{l+1} = \bar{f}(\bar{z}_l) = \underbrace{\bar{h}(\bar{z}_l) - \bar{z}_l}_{\text{Residual}}$

Aggregation Method :

Final volume : $W \times H \times D$ ←

Flatten the volume to get $\bar{z}_{W \times H \times D}$ ←

d-dimensional vector $\bar{y}$

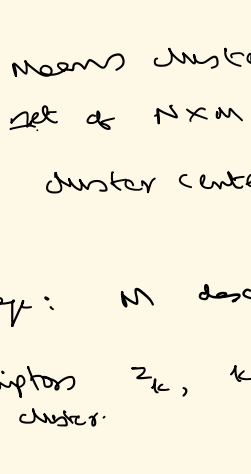
l-dimension vector $\bar{z}$

$l < d$

$\underset{l \times 1}{\bar{z}} = \underset{l \times d}{W}\, \underset{d \times 1}{\bar{y}}$

$\underbrace{W \times H}_{M} \times D \longrightarrow \bar{z}$

M vectors of dimension D



H W FM_1
H W FM_2
H W FM_D

M Local descriptors

- Dimension of local descriptor is D

Bag-of-words approach :

Text : Vocabulary of size : D

Elements of $\bar{z}$: Frequency of occurrence of a word

M local descriptors for each of N images in a training data set.

- Total no. of local descriptors : $N \times M$

K-Means clustering on set of $N \times M$ local descriptors.

K cluster centers : $\underbrace{\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_K}_{\text{Codebook}}$

Image : M descriptors ←

No. of descriptors in kth cluster z_k, $k = 1, 2, \ldots, K$

Image : $\bar{z} = [z_k]_{k=1}^{K}$

$W \times H \times D \longrightarrow K \times 1$

Vector of Locally Aggregated Descriptors (VLAD)

M Descriptors of an image : $\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_M\}$

$\bar{x}_i$: belongs to a cluster k

Membership value :

Hard assignment $a_{ki} = 1$, if $\bar{x}_i$ belongs to kth cluster

$= 0$, for all other clusters.

$\bar{z}_k = \sum_{i=1}^{M} a_{ki}(\bar{x}_i - \bar{\mu}_k)$

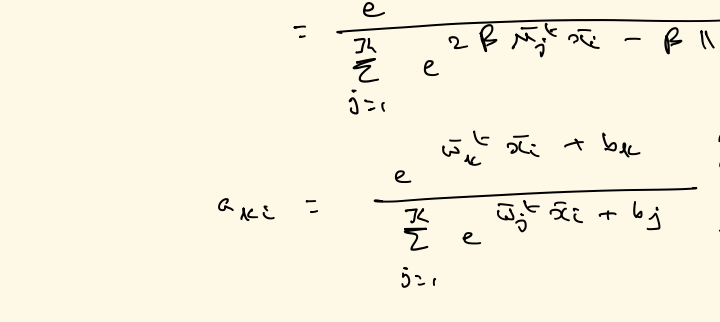
$\underset{K \times D}{\bar{z}} = \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \\ \vdots \\ \bar{z}_K \end{bmatrix}$

Soft assignment :

$a_{ki} = \frac{e^{-\beta \|\bar{x}_i - \bar{\mu}_k\|^2}}{\sum_{j=1}^{K} e^{-\beta \|\bar{x}_i - \bar{\mu}_j\|^2}}$

$\bar{z}_k = \sum_{i=1}^{M} a_{ki}(\bar{x}_i - \bar{\mu}_k)$

VLAD : $\underset{K \times D}{\bar{z}} = \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \\ \vdots \\ \bar{z}_K \end{bmatrix}$



Encoder
Decoder
Image
CNN
$\bar{z}$
RNN
Text

End-to-End training

NetVLAD

$$a_{ki} = \frac{e^{-\beta \|\bar{x}_i - \bar{\mu}_k\|^2}}{\sum_{j=1}^{K} e^{-\beta \|\bar{x}_i - \bar{\mu}_j\|^2}}$$

$$= \frac{e^{-\beta (\bar{x}_i - \bar{\mu}_k)^t(\bar{x}_i - \bar{\mu}_k)}}{\sum_{j=1}^{K} e^{-\beta (\bar{x}_i - \bar{\mu}_j)^t(\bar{x}_i - \bar{\mu}_j)}}$$

$$= \frac{e^{-\beta [\|\bar{x}_i\|^2 + \|\bar{\mu}_k\|^2 - 2\bar{x}_i^t\bar{\mu}_k]}}{\sum_{j=1}^{K} e^{-\beta [\|\bar{x}_i\|^2 + \|\bar{\mu}_j\|^2 - 2\bar{x}_i^t\bar{\mu}_j]}}$$

$$= \frac{e^{2\beta\bar{\mu}_k^t\bar{x}_i - \beta\|\bar{\mu}_k\|^2}}{\sum_{j=1}^{K} e^{2\beta\bar{\mu}_j^t\bar{x}_i - \beta\|\bar{\mu}_j\|^2}}$$

$$a_{ki} = \left.\frac{e^{\bar{w}_k^t\bar{x}_i + b_k}}{\sum_{j=1}^{K} e^{\bar{w}_j^t\bar{x}_i + b_j}}\right\}$$

$\bar{w}_k = 2\beta\bar{\mu}_k$ $b_k = -\beta\|\bar{\mu}_k\|^2$

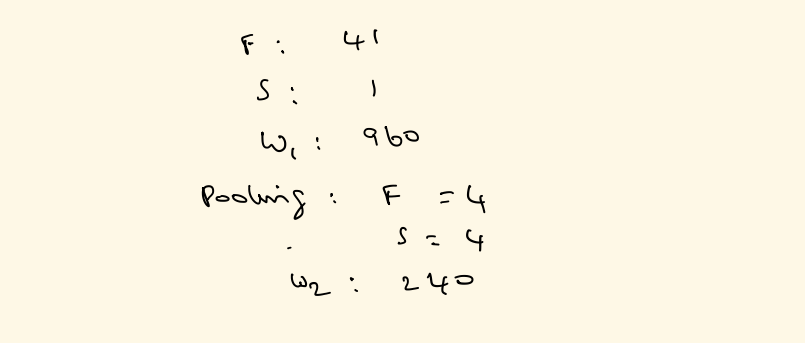
$\|\bar{w}_k\|^2 = 4\beta^2\|\bar{\mu}_k\|^2$

$b_k = -\beta\|\bar{\mu}_k\|^2$

$\frac{\|\bar{w}_k\|^2}{b_k} = -4\beta$

$\beta = -\frac{1}{4}\frac{\|\bar{w}_k\|^2}{b_k}$

$\bar{\mu}_k = \frac{1}{2\beta}\bar{w}_k$



Final volume of CNN
W×H×D
1×1×K Layer
Soft max
VLAD
$\bar{z}$
$\bar{w}_k, b_k$
a_{ki}
$\bar{\mu}_k$
NetVLAD Layer

2-dimensional CNNs.

- Images

1-dimensional CNNs.

- Audio
- 1-d kernel of size F
- Stride : S

1-d convolution:

Input : W samples.

Output of CL Feature map : W_1

$W_1 = \frac{W - F}{S} + 1$

W : 1000 samples

F : 41

S : 1

W_1 : 960

Pooling : F = 4

S = 4

W_2 : 240

3-dimensional CNNs for Video

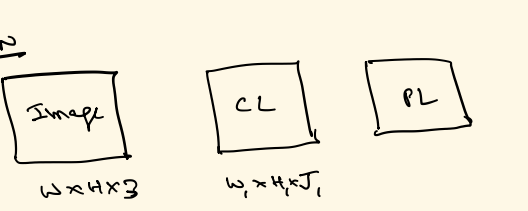
Video clip : T frames

$t = 1, 2, \ldots, T$

Frame : color image

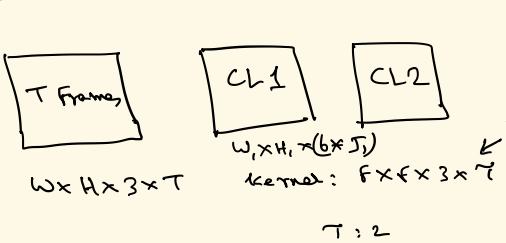
$W \times H \times T \times \underset{\uparrow\ \text{color channels}}{3}$

2-D CNN



Spatial Convolution
Spatial stride
Image
W×H×3
CL
$W_1 \times H_1 \times J_1$
Kernel : F×F×3
PL

3-D CNN



Spatiotemporal convolution
T Frames
W×H×3×T
CL1
CL2
$W_1 \times H_1 \times (6 \times J_1)$
Kernel : F×F×3×τ
Temporal depth of Kernel

τ : 2

Temporal stride S_t

T : 7 $\tau = 2$ $S_t = 1$

6 pairs of consecutive frames

6 FMs Kernel 1

No. of kernels : J_1

Total no. of FMs in CL : $6 \times J_1$

Depth of CL1 : $W_1 \times H_1 \times (6 \times J_1)$

Convolution kernel and pooling kernel : Temporal depth

I3D : Inflated 3-D CNN

- Two-streams of input
  1. Sequence of frames
  2. Optical flow features

C3D : 3-D CNN

Human action recognition

CNN for Image segmentation : U-Net

CNN for object detection : Faster Region CNN (Faster R-CNN)