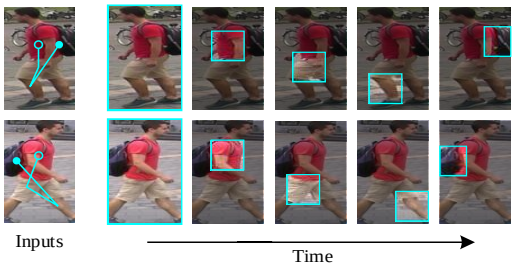
Recurrent Deep Attention Network for Person Re-Identification

Changhao Wang[1], Jun Zhou[2] , Xianfei Duan[2], Guanwen Zhang[1]*, Wei Zhou[1]
1 School of Electronics and Information, Northwestern Polytechnical University, Xi' an, China
2 CNPC logging Co.,Ltd, China

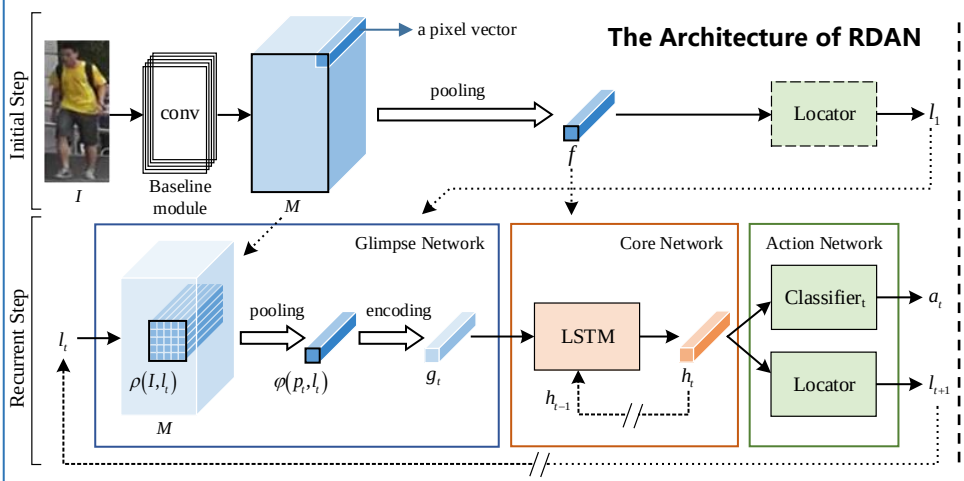
I. Introduction

- Person re-identification
- Attention selection
- Reinforcement learning
- Triplet-based reward

We propose a Recurrent Deep Attention Network (RDAN) that embeds convolutional architecture in a recurrent attention model and is able to select attention progressively.



II. The Proposed RDAN



- A Baseline Module (extracts deep feature)
- A Glimpse Module (generates local observation)
- A Core Network (combines temporal information)

Optimized by supervised learning

$$L_{id} = -\sum_{t=1}^{T}\sum_{i=1}^{N} 1\{i = y\}\log P_t^i$$

$$L_{tri} = \sum_{t=1}^{T} \max\left(\left(d_t^{a,p} - d_t^{a,n} + \alpha\right), 0\right)$$

- T Classifiers (for classification)
- A Locator (selects the location)

Optimized by reinforcement learning

$$J(\theta) = \mathbb{E}_\pi[R_{tri}] = \mathbb{E}_\pi[\sum_{t=1}^{T} r_t^{gc}]$$

$$r_t^{gc} = (d_t^{a,n} - d_t^{a,p}) - (d_{t-1}^{a,n} - d_{t-1}^{a,p})$$

At the initial step, baseline module transfers the input image I into a convolutional feature map M, and outputs the global feature vector f. The locator then takes f as input to select initial location where the glimpse network focuses on at the beginning.

During recurrent step, glimpse network extracts a glimpse around location l_t on M and produces the representation g_t of glimpse. The core network wraps g_t and h_{t-1} as inputs to output current hidden state h_t. Action network predicts identity of I and selects next location l_{t+1} based on h_t.

During deploying stage for person re-id, We refer to the **hidden state** $\boldsymbol{h_t}$ of core network at the last time step to indicate the identity of individual (i.e. final feature vector).

Where:

$$d_t^{a,p} = \|\boldsymbol{h}_t^a - \boldsymbol{h}_t^n\|_2 , d_t^{a,n} = \|\boldsymbol{h}_t^a - \boldsymbol{h}_t^n\|_2$$

III. Experiments

Comparison with the methods based on part-level features and attention mechanism

Methods	Market-1501		DukeMTMC-reID		CUHK03-NP(Detected)		CUHK03-NP(Labeled)	
	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)
PCB	92.3	77.4	81.9	65.3	61.3	54.2	-	-
PCB + RPP	93.8	81.6	83.3	69.2	63.7	57.5	-	-
HA-CNN	91.2	75.7	80.5	63.8	41.7	39.6	44.4	41.0
Mancs	93.1	82.3	84.9	71.8	65.5	60.5	69.0	63.9
CASN(IDE)	92.0	78.0	84.5	67.0	57.4	50.7	58.9	52.2
CASN(PCB)	94.4	82.8	87.7	73.7	71.5	64.4	73.7	68.0
RDAN	94.6	85.4	88.0	75.2	69.5	64.5	74.2	69.4

(a) (b) (c)

We connect the location produced by the proposed model to visualize the **attention selection process**, as shown in right figure.