Hierarchical Attentive Recurrent Tracking(HART)

(Adam et al. 2017)

Presented by Jiin Seo

February 2, 2018

Outline

1. Hierarchical Attention

2. Loss

Outline

1. Hierarchical Attention

2. Loss

1. Hierarchical Attention

HART (Adam et al. 2017)

- Goal : Single object tracking in videos

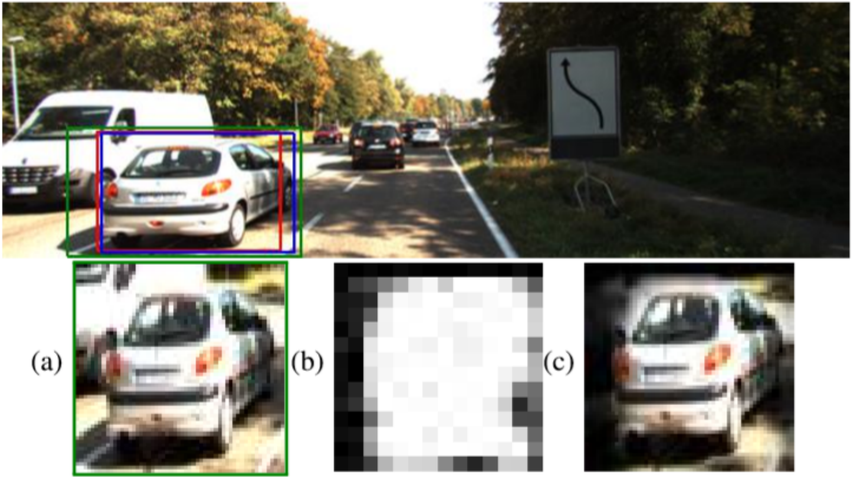

Figure: (a) attention glimpse (b) appearance attention (c)suppressing distractors

1. Hierarchical Attention

Architecture

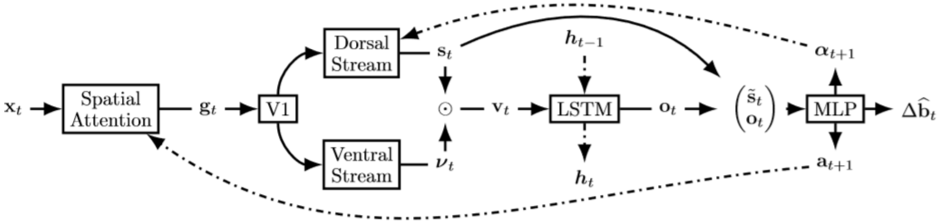

Figure: The architecture of HART

1. Hierarchical Attention

Spatial Attention

- Input image : $\mathbf{x}_t \in \mathbb{R}^{H \times W}$
- $\mathbf{A}_t^x \in \mathbb{R}^{w \times W}$, $\mathbf{A}_t^y \in \mathbb{R}^{h \times H}$
  : Each matix contains one Gaussian per row.
- The attention glimpse : $\mathbf{g}_t = \mathbf{A}_t^y \mathbf{x}_t (\mathbf{A}_t^x)^T, \ (\in \mathbb{R}^{h \times w})$

1. Hierarchical Attention

Appearance Attention

- $V1 : \mathbb{R}^{h \times w} \rightarrow \mathbb{R}^{h_v \times w_v \times c_v}$
- Dorsal stream computes foreground/background segmentation $\mathbf{s}_t$ using DFN.
- Ventral stream extracts appearance-based features $\boldsymbol{\nu}_t$ using CNN.

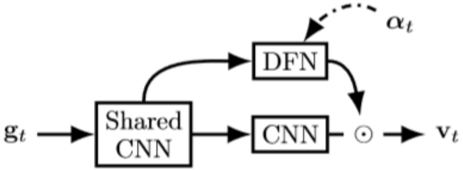

Figure: The architecture of the Appearance Attention

- Outputs of both stream combined as

$$\mathbf{v}_t = MLP(vec(\boldsymbol{\nu}_t \odot \mathbf{s}_t))$$

2. Hierarchical Attention

State Estimation

- Equations

$$
\begin{aligned}
\mathbf{o}_t, \mathbf{h}_t &= LSTM(\mathbf{v}_t, \mathbf{h}_{t-1}), \\
\boldsymbol{\alpha}_{t+1}, \Delta\mathbf{a}_{t+1}, \Delta\hat{\mathbf{b}}_t &= MLP(\mathbf{o}_t, vec(\mathbf{s}_t)), \\
\mathbf{a}_{t+1} &= \mathbf{a}_t + tanh(c)\Delta\mathbf{a}_{t+1}, \\
\hat{\mathbf{b}}_t &= \mathbf{a}_t + \Delta\hat{\mathbf{b}}_t
\end{aligned}
$$

Outline

2. Loss

Loss of HART

- Loss :

$$\mathcal{L}_{HART}(\mathcal{D}, \theta) = \lambda_t \mathcal{L}_t(\mathcal{D}, \theta) + \lambda_s \mathcal{L}_s(\mathcal{D}, \theta) + \lambda_a \mathcal{L}_a(\mathcal{D}, \theta) + R(\boldsymbol{\lambda}) + \beta R(\mathcal{D}, \theta),$$

with dataset $\mathcal{D} = \{(\mathbf{x}_{1:T}, \mathbf{b}_{1:T})^i\}_{i=1}^M$ and network parameter θ

2. Loss

Tracking Loss

- Tracking Loss term is based on IoU of the predicted bounding box w.r.t the ground truth.

$$\mathcal{L}_t(\mathcal{D}, \theta) = \mathrm{E}_{p(\hat{\mathbf{b}}_{1:T} | \mathbf{x}_{1:T}, \mathbf{b}_1)}[-log\ IoU(\hat{\mathbf{b}}_t, \mathbf{b}_t)],$$

- IoU(Intersection-over-Union)

$$IoU(a, b) = \frac{a \cap b}{a \cup b} = \frac{area\ of\ overlap}{area\ of\ union}$$

2. Loss

Spatial Attention Loss

$$\mathcal{L}_s(\mathcal{D}, \theta) = \mathrm{E}_{p(\mathbf{a}_{1:T}|\mathbf{x}_{1:T}, \mathbf{b}_1)}[-log(\frac{\mathbf{a}_t \cap \mathbf{b}_t}{area(\mathbf{b}_t)}) - log(1 - IoU(\mathbf{a}_t, \mathbf{x}_t))],$$

- The first term constrains the predicted attention to cover the bounding box.
- The second term prevents it from becoming too large.

2. Loss

Appearance Attention Loss

$$\mathcal{L}_a(\mathcal{D}, \theta) = \mathrm{E}_{p(\mathbf{a}_{1:T}, \mathbf{s}_{1:T} | \mathbf{x}_{1:T}, \mathbf{b}_1)}[\mathbf{H}(\tau(\mathbf{a}_t, \mathbf{b}_t), \mathbf{s}_t)],$$

, where $\mathbf{H}(p, q) = -\sum_z p(z) log \; q(z)$

$$\begin{aligned} \tau(\mathbf{a}_t, \mathbf{b}_t) &: \mathbb{R}^4 \times \mathbb{R}^4 \rightarrow \{0, 1\}^{h_v \times w_v} \\ &= \begin{cases} 1 & \text{where the bounding box overlaps with glimpse} \\ 0 & \text{o.w.} \end{cases} \end{aligned}$$

2. Loss

Regularisation

- We apply the L_2 regularisation to the parameters θ

$$\mathbf{R}(\mathcal{D}, \theta) = \frac{1}{2} \parallel \theta \parallel_2^2 + \frac{1}{2} \parallel \mathrm{E}_{p(\boldsymbol{\alpha}_{1:T} | \mathbf{x}_{1:T}, \mathbf{b}_1)}[\Psi_t \mid \boldsymbol{\alpha}_t] \parallel_2^2$$

Adaptive Loss Weights

- We learn the loss weighting $\boldsymbol{\lambda}$ to avoid hyper-parameter tuniing.

$$\mathbf{R}(\boldsymbol{\lambda}) = -\sum_i log(\boldsymbol{\lambda}_i^{-1})$$

, where $\boldsymbol{\lambda} = \{\lambda_t, \lambda_s, \lambda_a\}$,