Yandex

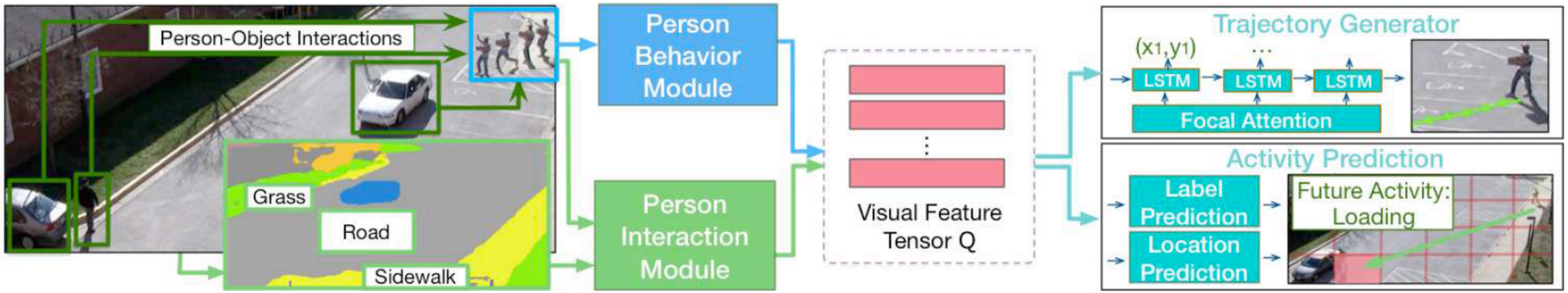
Reference papers for trajectory prediction

Rudenko Irina

Peeking into the Future: Predicting Future Person Activities and Locations in Videos

Carnegie Mellon University Google AI Stanford University

Модель:


Person-Object Interactions
Grass
Road
Sidewalk
Person Behavior Module
Person Interaction Module
Visual Feature Tensor Q
Trajectory Generator
(x1,y1)
...
LSTM
LSTM
LSTM
Focal Attention
Activity Prediction
Label Prediction
Location Prediction
Future Activity: Loading

Person behavior module:

- extracts visual information from the behavioral sequence of the person

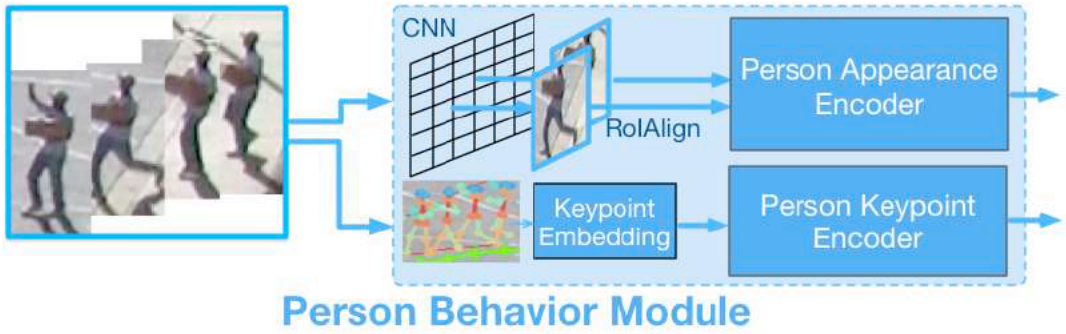


Figure 3. Person behavior module given a sequence of person frames. See Section 3.2.

- Uses a pre-trained object detection model with “RoIAlign” for person appearance features
- Averages the features along the spatial dimensions for each person and feeds them into an LSTM encoder
- Regional multi-person pose estimation model to extract person keypoint information
- Apply the linear transformation to embed the keypoint coordinates before feeding into the LSTM encoder.

Person interaction module:

- looks at the interaction between a person and their surroundings.

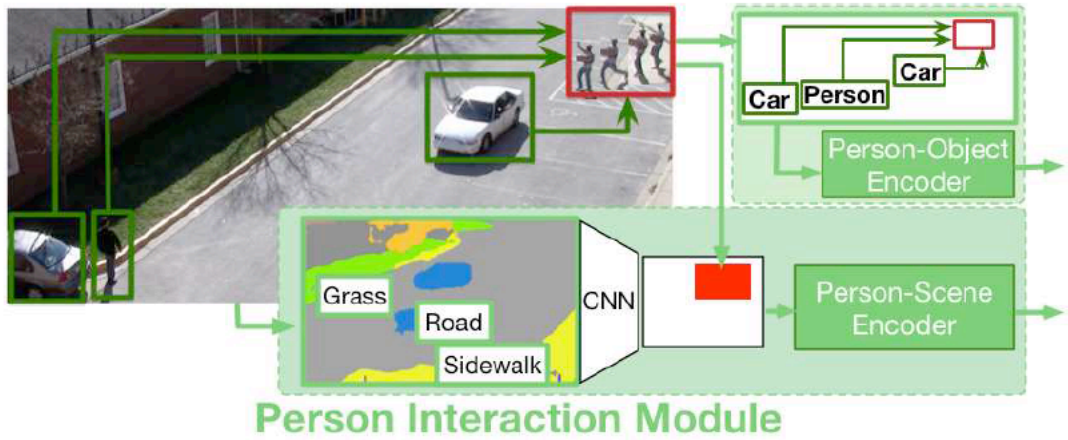


Figure 4. The person interaction module includes person-scene and person-objects modeling. See Section 3.3.

- Pre-trained scene segmentation model, 10 common scene classes (roads, sidewalks, etc.)
- Transform the integer tensor into binary masks and average along the temporal dimension
- Apply two convolutional layers on the mask feature with a stride of 2 to get the scene CNN features in two scales
- Pool the scene features at the person's current location from the convolution feature map and feed this into a LSTM encoder

Person interaction module:

- looks at the interaction between a person and their surroundings.

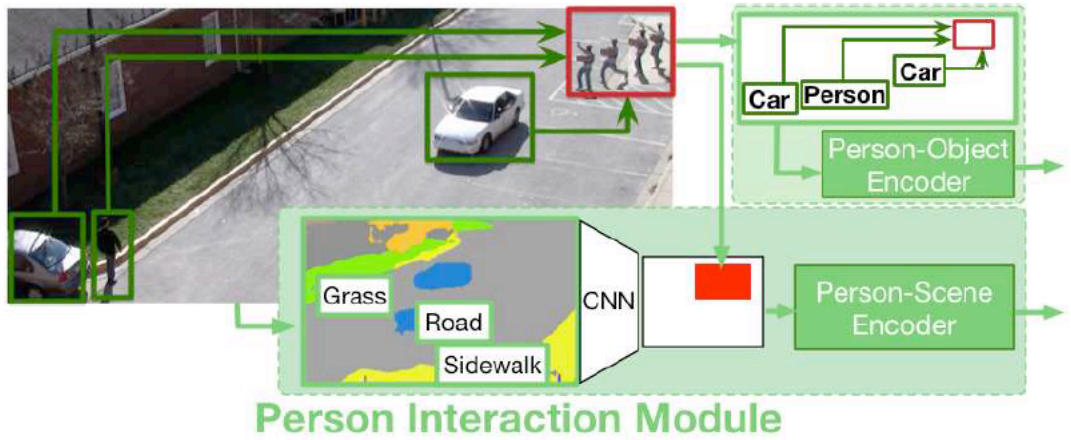


Figure 4. The person interaction module includes person-scene and person-objects modeling. See Section 3.3.

- Encode the geometric relation

$$\mathcal{G}_k = [\log(\frac{|x_b - x_k|}{w_b}), \log(\frac{|y_b - y_k|}{h_b}), \log(\frac{w_k}{w_b}), \log(\frac{h_k}{h_b})]$$

- One-hot encoding for the object type

- Embed all and feed this into a LSTM encoder

Trajectory embeddings:

- Extract trajectory embeddings.

$$e_{t-1} = \tanh\{W_e[x_{t-1}, y_{t-1}]\} + b_e \in \mathbb{R}^d$$

- Feed them into a LSTM encoder

Activity Prediction:

- Utilizes rich visual semantics to predict the future activity label and location for the person.

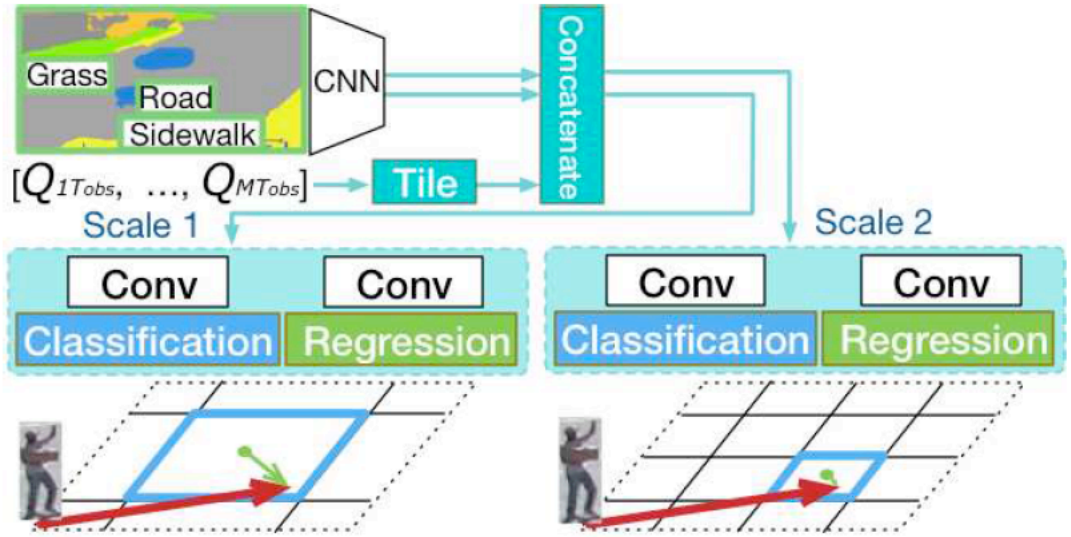


Figure 5. Activity location prediction with classification and regression on the multi-scale Manhattan Grid. See Section 3.5.

- Compute the future Na activity probabilities using the concatenated last hidden states of the encoders

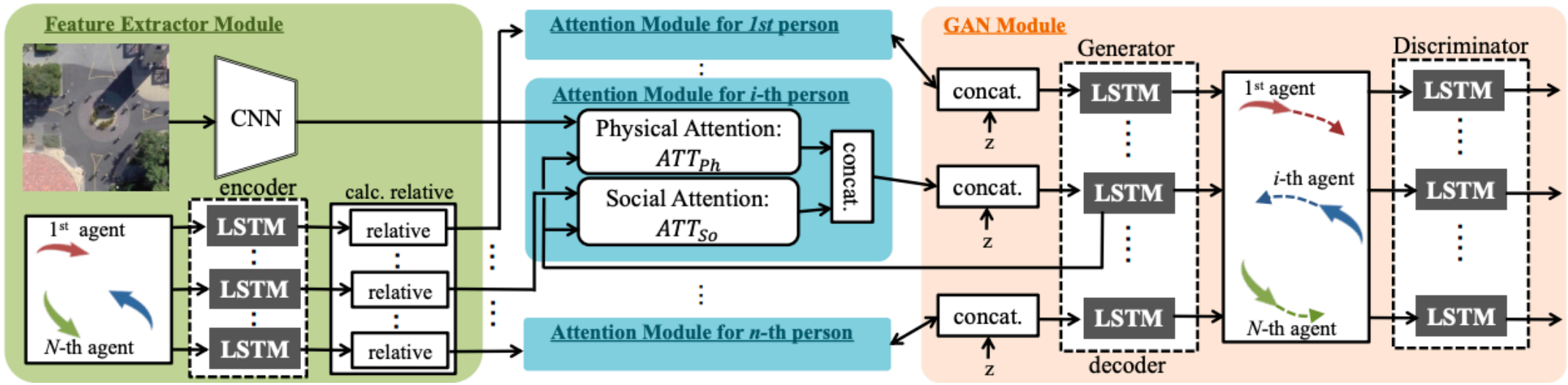
$$\text{cls}_{act} = \text{softmax}(W_a \cdot [Q_{1T_{obs}:}, \cdots, Q_{MT_{obs}:}])$$

- The future activity of a person could be multi-class

SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints

Stanford University University of Florida University of Adelaide

Модель:


Feature Extractor Module
CNN
encoder
calc. relative
1st agent
N-th agent
LSTM
LSTM
LSTM
relative
relative
relative
Attention Module for 1st person
Attention Module for i-th person
Physical Attention: ATT_{Ph}
Social Attention: ATT_{So}
concat.
Attention Module for n-th person
GAN Module
Generator
concat.
concat.
concat.
z
z
z
LSTM
LSTM
LSTM
decoder
1st agent
i-th agent
N-th agent
Discriminator
LSTM
LSTM
LSTM

Feature extractors

- Extract visual information from from the image I_t by VGGnet-19

$$V_{Ph}^t = CNN(I^t; W_{cnn})$$

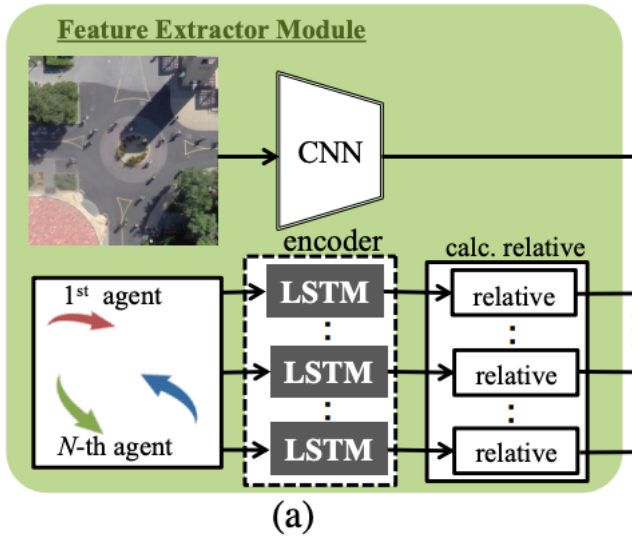
- Extract features from the past trajectory of all agents by LSTM

$$V_{en}^t(i) = LSTM_{en}(X_i^t, h_{en}^t(i); W_{en})$$

- Extract relative features

$$V_{So}^t(i) = \left(V_{en}^t(\pi_j) - V_{en}^t(i) \middle| \forall \pi_j \in [N] \backslash i\right)$$

where π_j is the index of the other agents sorted according to their distances to the target agent i



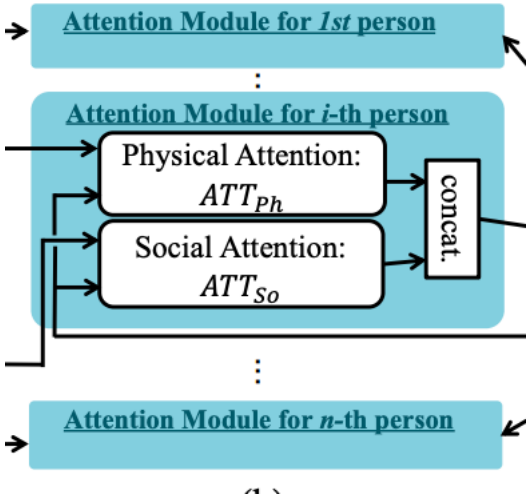
Attention Module

- Physical Attention: learns the spatial (physical) constraints in the scene from the training data

$$C_{Ph}^t(i) = ATT_{Ph}(V_{Ph}^t, h_{dec}^t(i); W_{Ph})$$

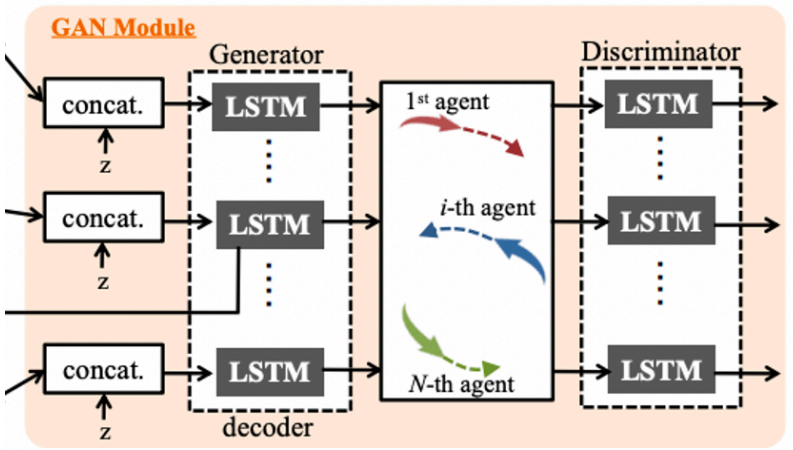
- Social Attention: highlights which other agents are most important to focus on when predicting the trajectory of the agent i

$$C_{So}^t(i) = ATT_{So}(V_{So}^t(i), h_{dec}^t(i); W_{So})$$



LSTM based Generative Adversarial Network

- LSTM generator from scene encoding + white noise vector
- LSTM discriminator
- Losses:



$$W^* = \underset{W}{\operatorname{argmin}} \quad \mathbb{E}_{i,\tau}[\mathcal{L}_{GAN}(\hat{L}_i^\tau, L_i^\tau) + \lambda \mathcal{L}_{L2}(\hat{Y}_i^{1:\tau}, Y_i^{1:\tau})],$$

$$\mathcal{L}_{GAN}(\hat{L}_i^\tau, L_i^\tau) =$$

$$\min_G \max_D \quad \mathbb{E}_{T_i^{1:\tau} \sim p(Y_i^{1:\tau})}[L_i^\tau \log \hat{L}_i^\tau] + \mathbb{E}_{T_i^{1:\tau} \sim p(\hat{Y}_i^{1:\tau})}[(1 - L_i^\tau)\log(1 - \hat{L}_i^\tau)],$$

$$\mathcal{L}_{L2}(\hat{Y}_i^\tau, Y_i^\tau) = ||\hat{Y}_i^\tau - Y_i^\tau||_2^2.$$