# Semantic Object Parsing with Graph LSTM

Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, Shuicheng Yan

Sun Yat-sen University, 360 AI Institue, National University of Singapore
Adobe Research

# Motivation

- Extends the traditional LSTMs from sequential and multi-dimensional data to general graph-structured data.

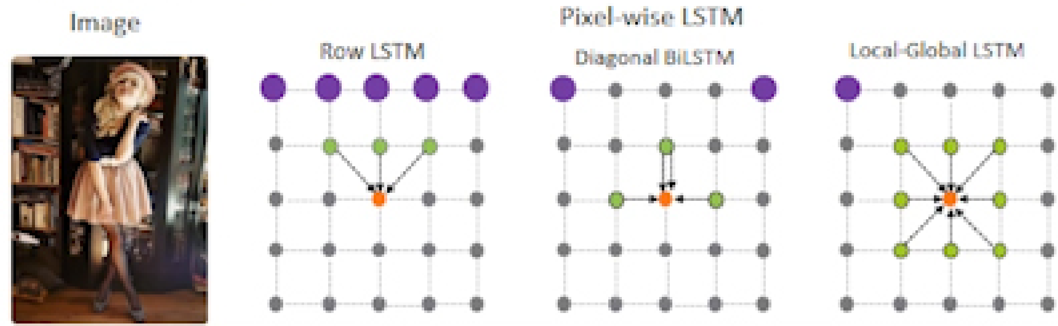
- Graph LSTM can handle the inference over an adaptive graph topology where different nodes are connected with different numbers of neighbors, depending on the local structures in the image.

- Effectively reduces redundant computational costs while better preserving object/part boundaries to facilitate global reasoning.

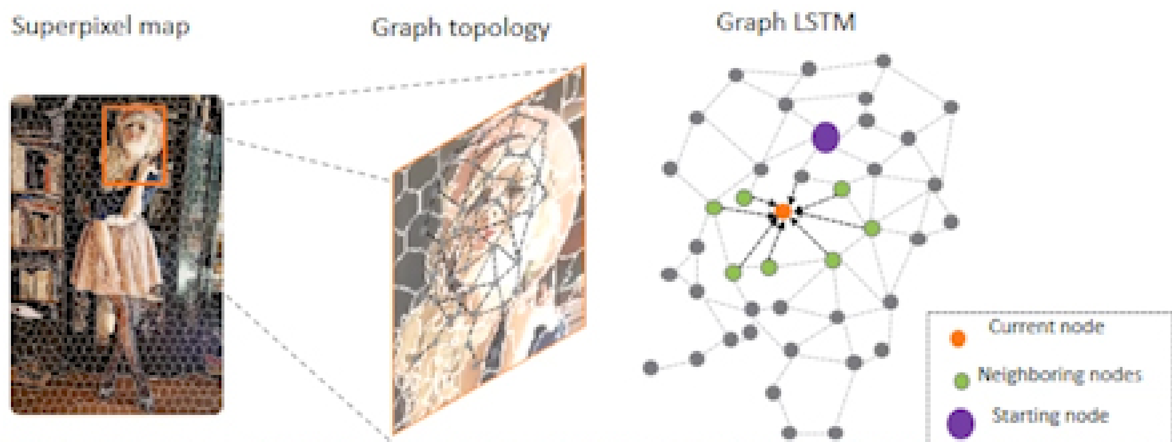
# Graph LSTM

- Generalize the LSTM for sequential data or multi-dimensional data to general graph-structured data

Traditional multi-dimensional LSTM:




Graph LSTM:



# The key contributions of Graph LSTM

- Confidence-driven Scheme

Graph LSTM specifies the adaptive starting node and node updating sequence for the information propagation.

- Averaged Hidden States for Neighboring Nodes.

Considering the adaptive graph topology for each image, the hidden states used for computing the LSTM gates of each node are obtained by averaging the hidden states of neighboring nodes.

- Adaptive Forget Gates

Graph LSTM configures different forget gates for different neighboring nodes in order to capture their distinguished semantic correlations.

# Graph LSTM Unit

- The hidden and memory states by Graph LSTM can be updated as follows:

$$
\begin{aligned}
g_i^u &= \delta(W^u \mathbf{f}_{i,t+1} + U^u \mathbf{h}_{i,t} + U^{un} \bar{\mathbf{h}}_{i,t} + b^u), \\
\bar{g}_{ij}^f &= \delta(W^f \mathbf{f}_{i,t+1} + U^{fn} \mathbf{h}_{j,t} + b^f), \\
g_i^f &= \delta(W^f \mathbf{f}_{i,t+1} + U^f \mathbf{h}_{i,t} + b^f), \\
g_i^o &= \delta(W^o \mathbf{f}_{i,t+1} + U^o \mathbf{h}_{i,t} + U^{on} \bar{\mathbf{h}}_{i,t} + b^o), \\
g_i^c &= \tanh(W^c \mathbf{f}_{i,t+1} + U^c \mathbf{h}_{i,t} + U^{cn} \bar{\mathbf{h}}_{i,t} + b^c),
\end{aligned}
$$

Adaptive forget gates

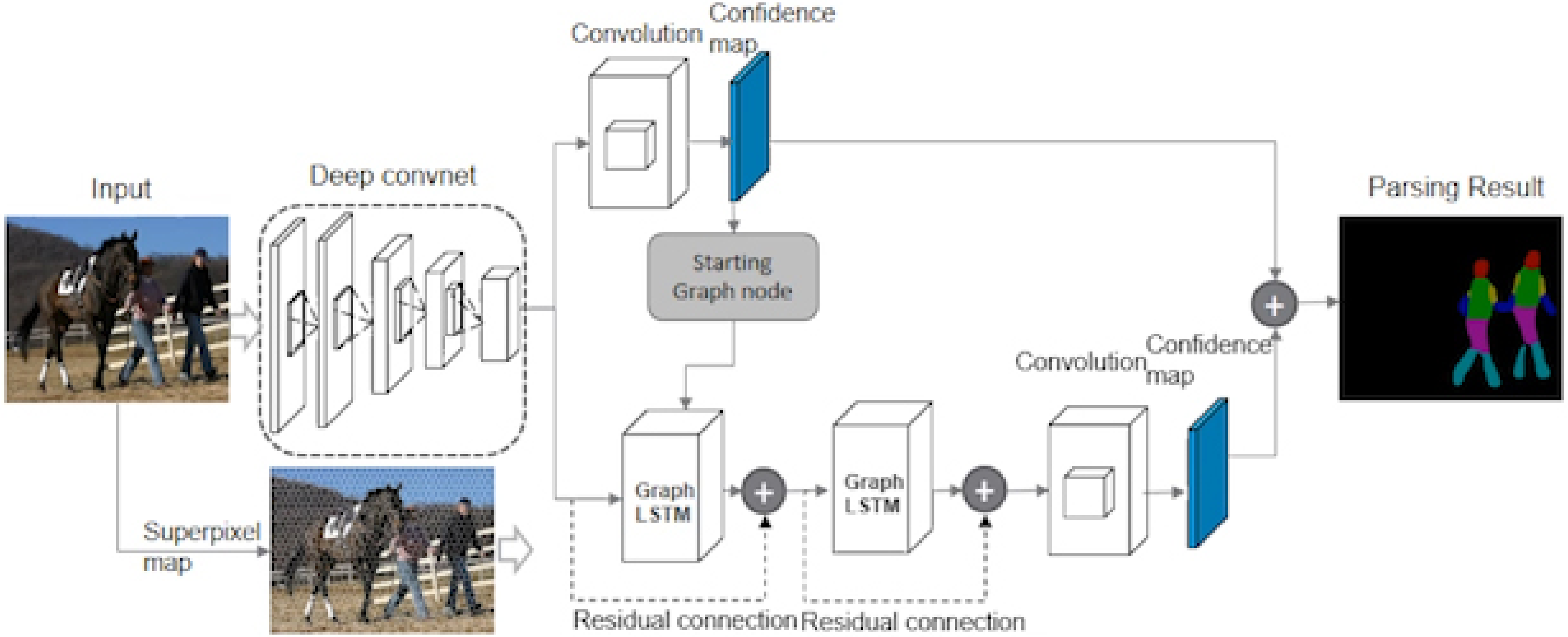
$$
\mathbf{m}_{i,t+1} = \frac{\sum_{j \in \mathcal{N}_\mathcal{G}(i)} (\mathbb{1}(q_j = 1) \bar{g}_{ij}^f \odot \mathbf{m}_{j,t+1} + \mathbb{1}(q_j = 0) \bar{g}_{ij}^f \odot \mathbf{m}_{j,t})}{|\mathcal{N}_\mathcal{G}(i)|} + g_i^f \odot \mathbf{m}_{i,t} + g_i^u \odot g_i^c,
$$

$$
\mathbf{h}_{i,t+1} = \tanh(g_i^o \odot \mathbf{m}_{i,t+1}).
$$

The memory states of the node are updated by combining the memory states of visited nodes and those of unvisited nodes by using the adaptive forget gates.

# Network Architecture for Semantic Object Parsing

- The Graph LSTM layers are stacked to sequentially update the hidden states of all super-pixel nodes.



# Experiments

- The graph LSTM obtains the state-of-art performances on four object parsing dataset.

## PASCAL-Person-Part

| Method | head | torso | u-arms | l-arms | u-legs | l-legs | Bkg | Avg |
|---|---|---|---|---|---|---|---|---|
| DeepLab-LargeFOV [15] | 78.09 | 54.02 | 37.29 | 36.85 | 33.73 | 29.61 | 92.85 | 51.78 |
| HAZN [12] | 80.79 | 59.11 | 43.05 | 42.76 | 38.99 | 34.46 | 93.59 | 56.11 |
| Attention [13] | - | - | - | - | - | - | - | 56.39 |
| LG-LSTM [21] | **82.72** | 60.99 | 45.40 | **47.76** | 42.33 | 37.96 | 88.63 | 57.97 |
| **Graph LSTM** | 82.69 | **62.68** | **46.88** | 47.71 | **45.66** | **40.93** | **94.59** | **60.16** |

## Horse-Cow Parsing

| Horse | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Bkg | head | body | leg | tail | Fg | IOU | Pix.Acc |
| SPS [26] | 79.14 | 47.64 | 69.74 | 38.85 | - | 68.63 | - | 81.45 |
| HC [36] | 85.71 | 57.30 | 77.88 | 51.93 | 37.10 | 78.84 | 61.98 | 87.18 |
| Joint [16] | 87.34 | 60.02 | 77.52 | 58.35 | 51.88 | 80.70 | 65.02 | 88.49 |
| LG-LSTM [21] | 89.64 | 66.89 | 84.20 | 60.88 | 42.06 | 82.50 | 68.73 | 90.92 |
| HAZN [12] | 90.87 | 70.73 | 84.45 | 63.59 | 51.16 | - | 72.16 | - |
| **Graph LSTM** | **91.73** | **72.89** | **86.34** | **69.04** | **53.76** | **87.51** | **74.75** | **92.76** |

| Cow | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Bkg | head | body | leg | tail | Fg | IOU | Pix.Acc |
| SPS [26] | 78.00 | 40.55 | 61.65 | 36.32 | - | 71.98 | - | 78.97 |
| HC [36] | 81.86 | 55.18 | 72.75 | 42.03 | 11.04 | 77.04 | 52.57 | 84.43 |
| Joint [16] | 85.68 | 58.04 | 76.04 | 51.12 | 15.00 | 82.63 | 57.18 | 87.00 |
| LG-LSTM [21] | 89.71 | 68.43 | 82.47 | 53.93 | 19.41 | 85.41 | 62.79 | 90.43 |
| HAZN [12] | 90.66 | **75.10** | 83.30 | 57.17 | 28.46 | - | 66.94 | - |
| **Graph LSTM** | **91.54** | 73.88 | **85.92** | **63.67** | **35.22** | **88.42** | **70.05** | **92.43** |

## ATR

| Method | Acc. | F.g. acc. | Avg. prec. | Avg. recall | Avg. F-1 score |
|---|---|---|---|---|---|
| Yamaguchi et al. [28] | 84.38 | 55.59 | 37.54 | 51.05 | 41.80 |
| PaperDoll [37] | 88.96 | 62.18 | 52.75 | 49.43 | 44.76 |
| M-CNN [41] | 89.57 | 73.98 | 64.56 | 65.17 | 62.81 |
| ATR [27] | 91.11 | 71.04 | 71.69 | 60.25 | 64.38 |
| Co-CNN [42] | 95.23 | 80.90 | 81.55 | 74.42 | 76.95 |
| Co-CNN (more) [42] | 96.02 | 83.57 | 84.95 | 77.66 | 80.14 |
| LG-LSTM [21] | 96.18 | 84.79 | 84.64 | 79.43 | 80.97 |
| LG-LSTM (more) [21] | 96.85 | 87.35 | 85.94 | 82.79 | 84.12 |
| CRFasRNN (more) [10] | 96.34 | 85.10 | 84.00 | 80.70 | 82.08 |
| Graph LSTM | 97.60 | 91.42 | 84.74 | 83.28 | 83.76 |
| **Graph LSTM (more)** | **97.99** | **93.06** | **88.81** | **87.80** | **88.20** |

## Fashionista

| Method | Acc. | F.g. acc. | Avg. prec. | Avg. recall | Avg. F-1 score |
|---|---|---|---|---|---|
| Yamaguchi et al. [28] | 87.87 | 58.85 | 51.04 | 48.05 | 42.87 |
| PaperDoll [37] | 89.98 | 65.66 | 54.87 | 51.16 | 46.80 |
| ATR [27] | 92.33 | 76.54 | 73.93 | 66.49 | 69.30 |
| Co-CNN [42] | 96.08 | 84.71 | 82.98 | 77.78 | 79.37 |
| Co-CNN (more) [42] | 97.06 | 89.15 | 87.83 | 81.73 | 83.78 |
| LG-LSTM [21] | 96.85 | 87.71 | 87.05 | 82.14 | 83.67 |
| LG-LSTM (more) [21] | 97.66 | 91.35 | 89.54 | 85.54 | 86.94 |
| Graph LSTM | 97.93 | 92.78 | 88.24 | 87.13 | 87.57 |
| **Graph LSTM (more)** | **98.14** | **93.75** | **90.15** | **89.46** | **89.75** |

# Discussions

- Graph LSTM vs locally fixed factorized LSTM

Using richer and adaptive local contexts (i.e., number of neighbors) to update the states of each pixel can lead to better parsing performance.

- Adaptive forget gates vs Identical forget gates

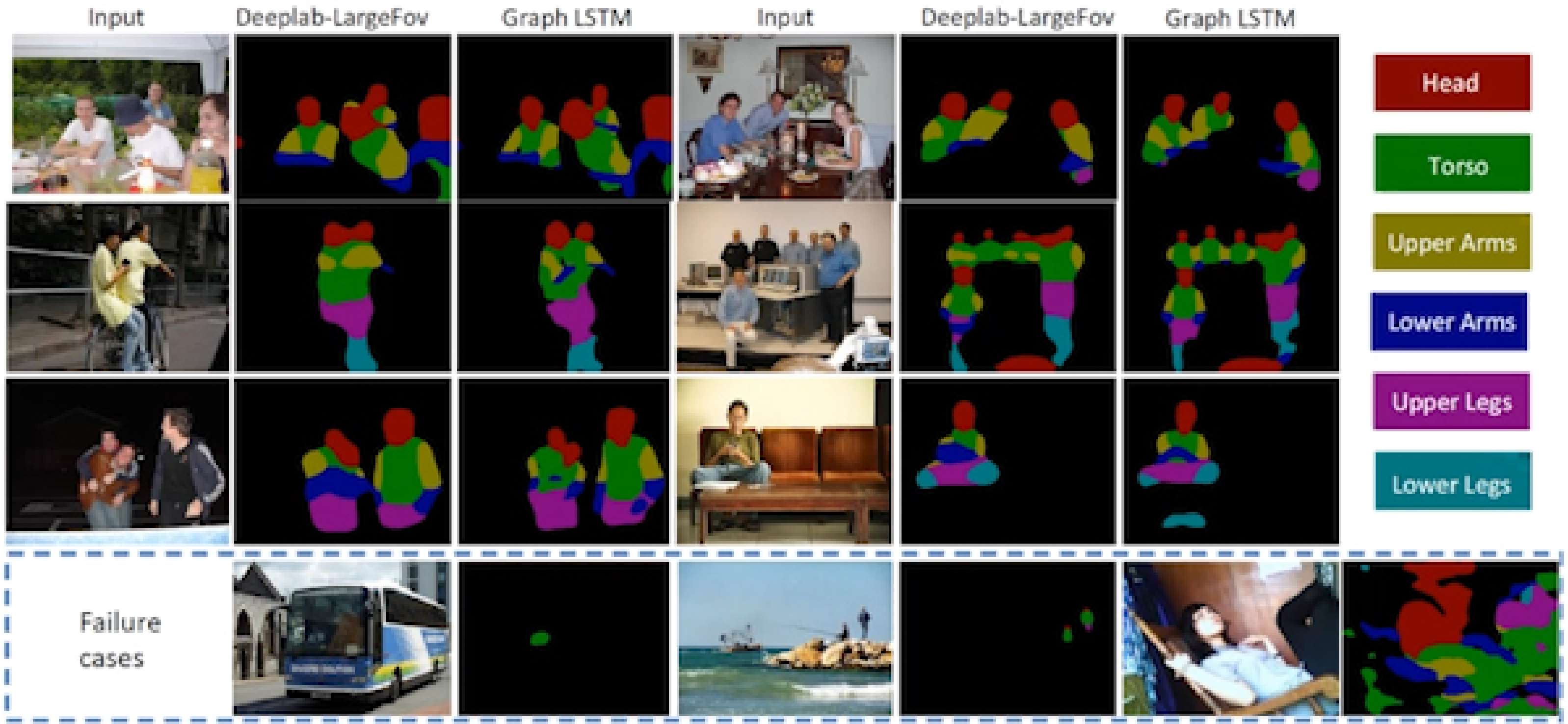
Diverse semantic correlations with local context can be considered and treated differently during the node updating.

- Confidence-driven node updating scheme

The features of superpixel nodes with higher foreground confidences embed more accurate semantic meanings and thus lead to more reliable global reasoning.

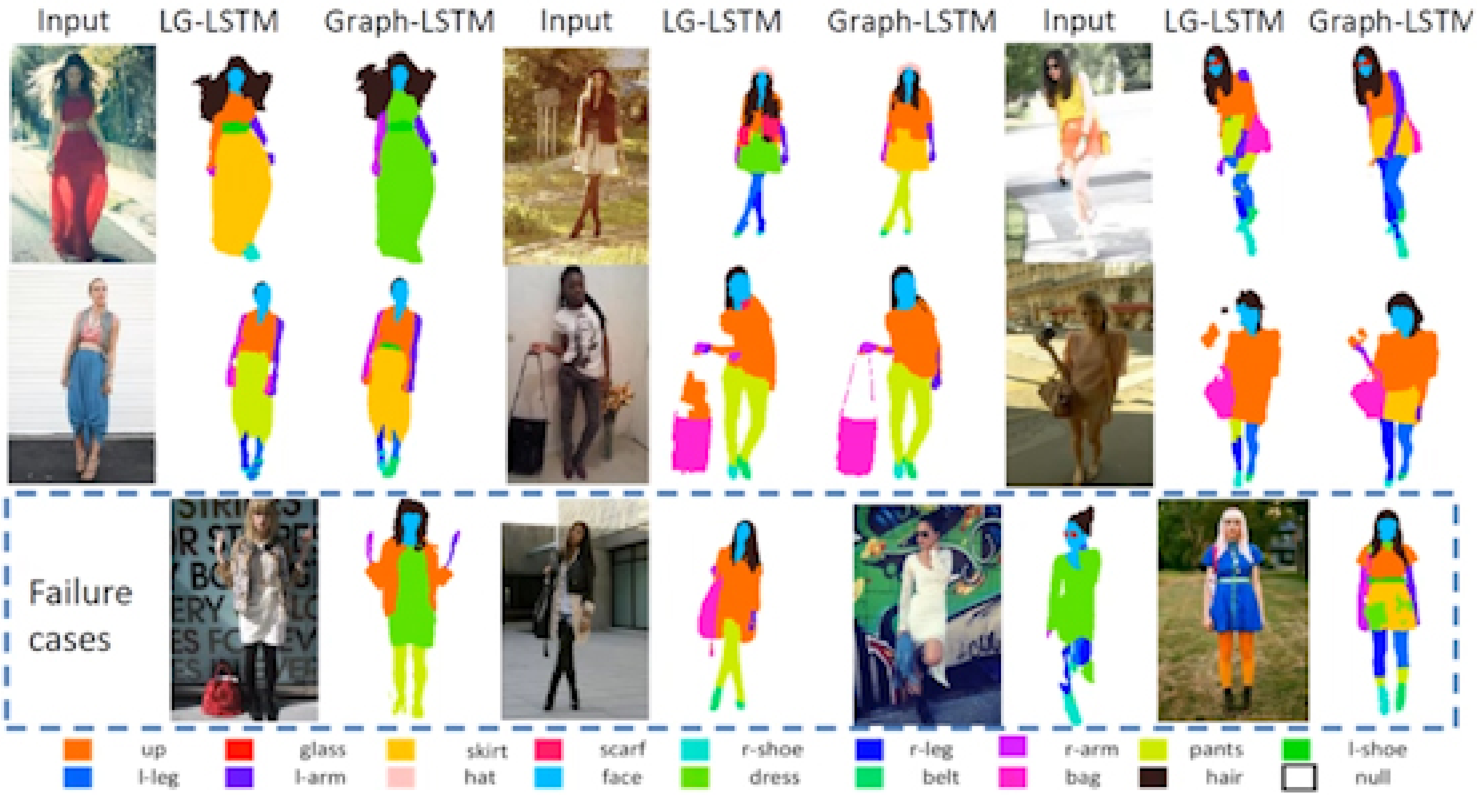
# Experiments

- Visual Comparison and failure cases on PASCAL-Person-Part dataset



# Experiments

- Visual comparison and failure cases on ATR dataset



# Semantic Object Parsing with Graph LSTM

Please stop by **Poster Session 1A: S-1A-08** for more details