Multimodal Fusion in Fully-connected Architecture

Weiwen Chen, Shengliang Cai

Outline

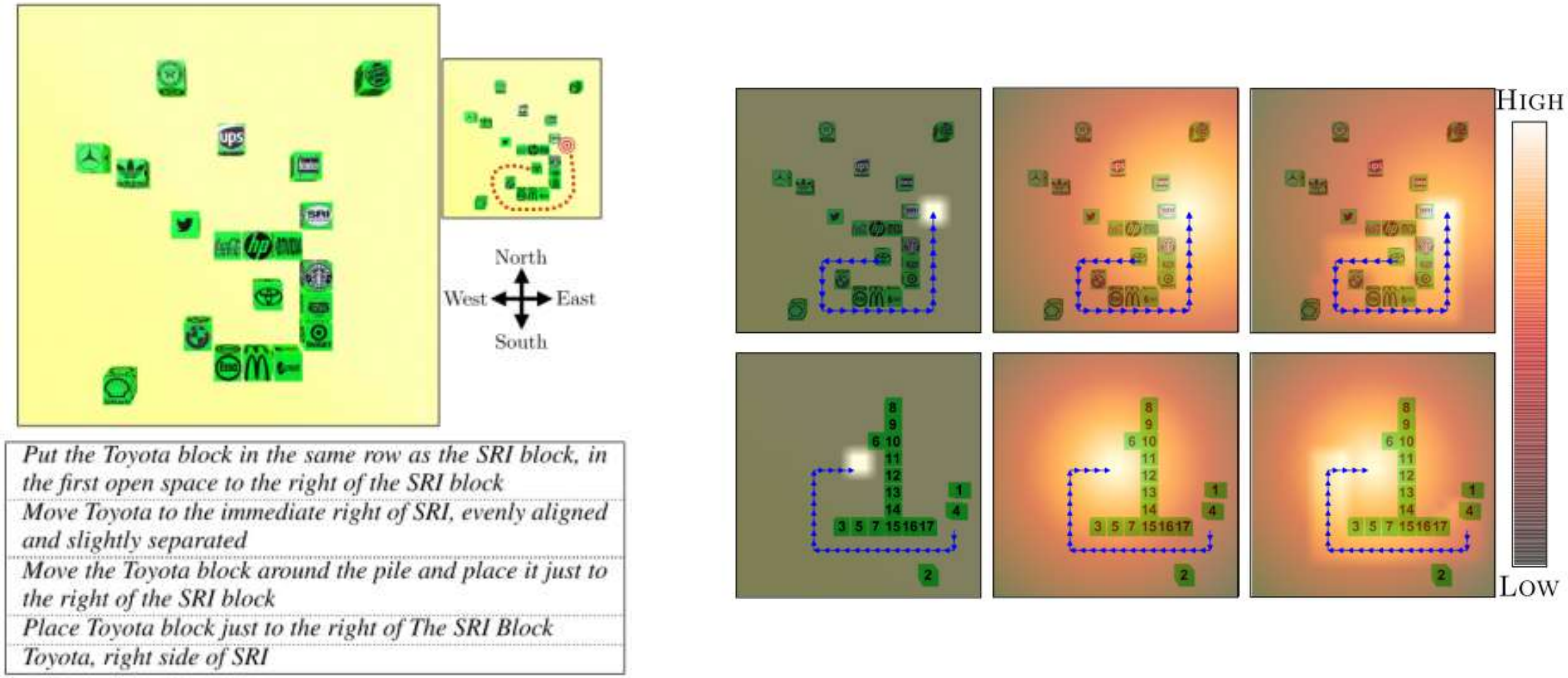
- Mapping
- NiN
- SENet
- Biliner Pooling

- Idea

Mapping

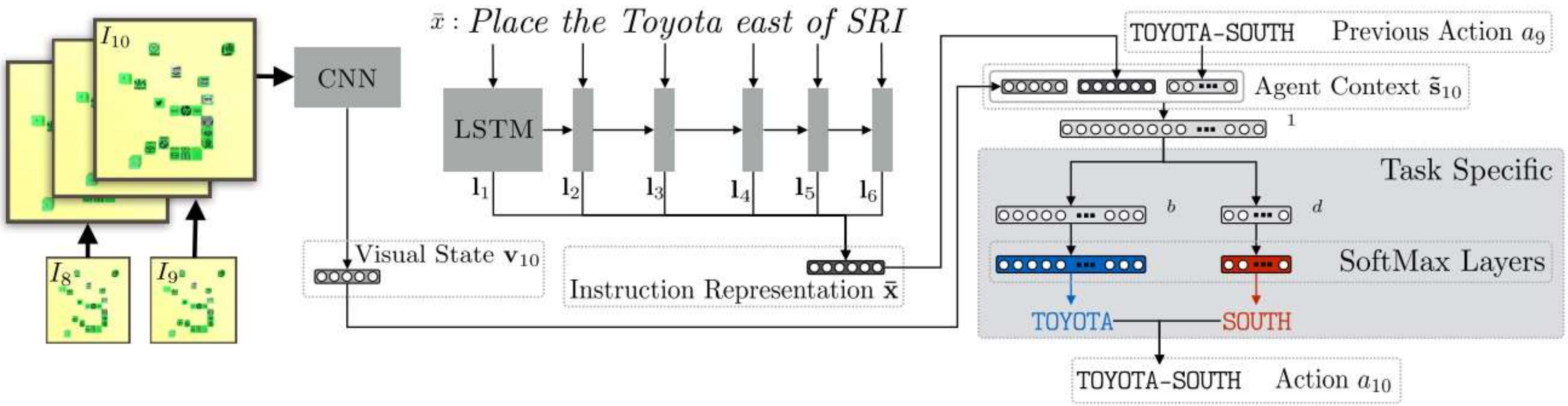
Instructions and Visual Observations to Actions

Mapping



Misra, D., Langford, J., & Artzi, Y. (2017). Mapping instructions and visual observations to actions with reinforcement learning. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1004–1015.

Mapping



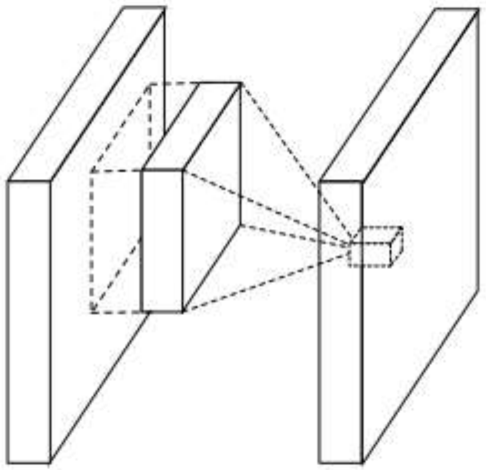
Misra, D., Langford, J., & Artzi, Y. (2017). Mapping instructions and visual observations to actions with reinforcement learning. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1004–1015.

Network in Network

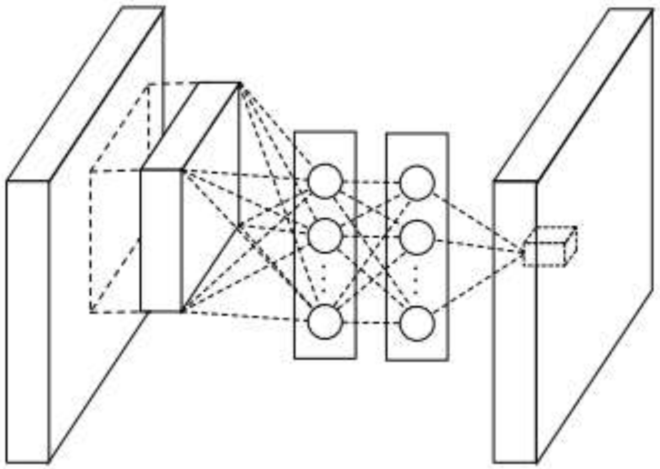
Mapping Instructions and Visual Observations to Actions

Network in Network mlpconv Layer

多智体人工智能实验室
Multi-Agent Artificial Intelligence Laboratory



(a) Linear convolution layer



(b) Mlpconv layer

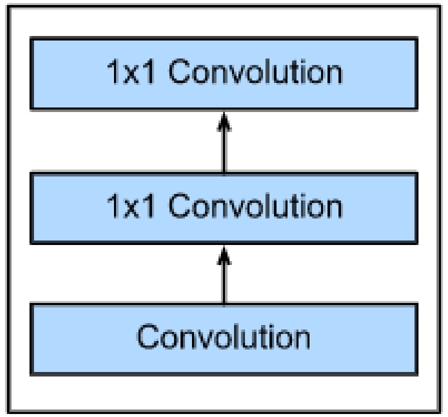
$$
\begin{aligned}
f^1_{i,j,k_1} &= \max({w^1_{k_1}}^T x_{i,j} + b_{k_1}, 0). \\
&\vdots \\
f^n_{i,j,k_n} &= \max({w^n_{k_n}}^T f^{n-1}_{i,j} + b_{k_n}, 0).
\end{aligned}
$$

Lin, M., Chen, Q., & Yan, S. (2013). Network In Network (paper). *ArXiv Preprint*, 10. http://arxiv.org/abs/1312.4400

Network in Network mlpconv Layer

- *n* is the number of layers in the multilayer perceptron. Rectified linear unit is used as the activation function in the multilayer perceptron.
- The above structure **allows complex and learnable interactions of cross channel information.**
- It is **equivalent to a convolution layer with 1×1 convolution kernel**.

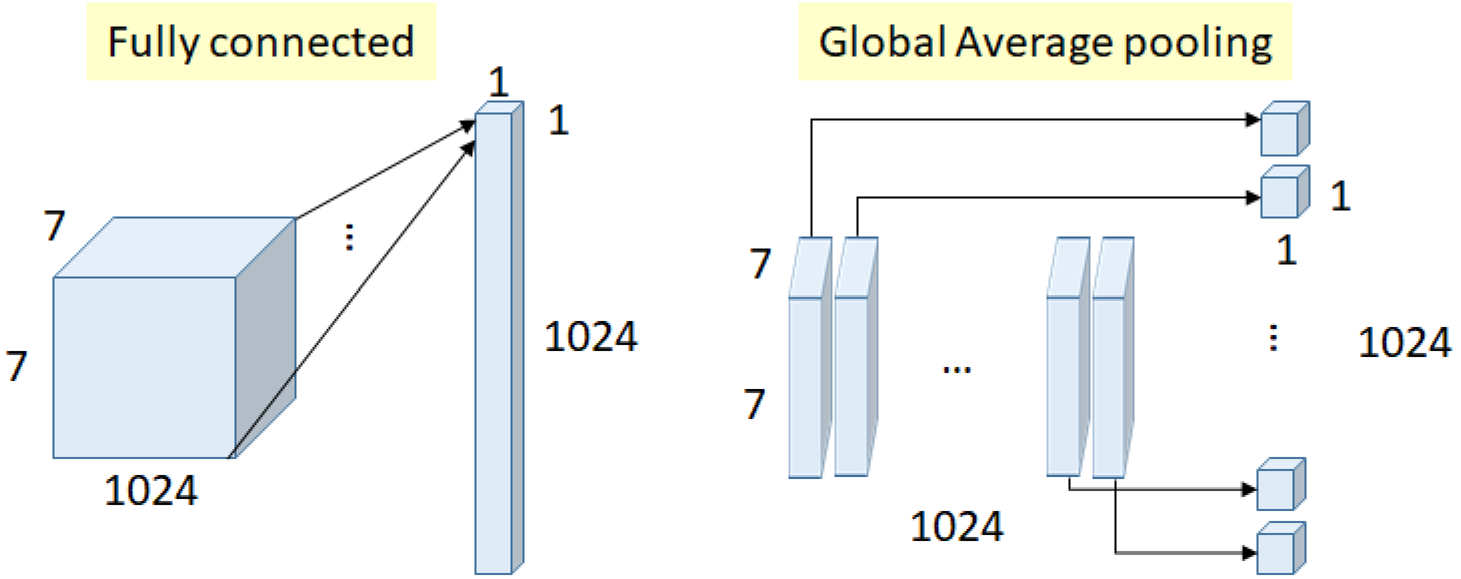
NiN block



Lin, M., Chen, Q., & Yan, S. (2013). Network In Network (paper). *ArXiv Preprint*, 10. http://arxiv.org/abs/1312.4400

Network in Network Global Average Pooling

- Fully Connected Layer – overfitting



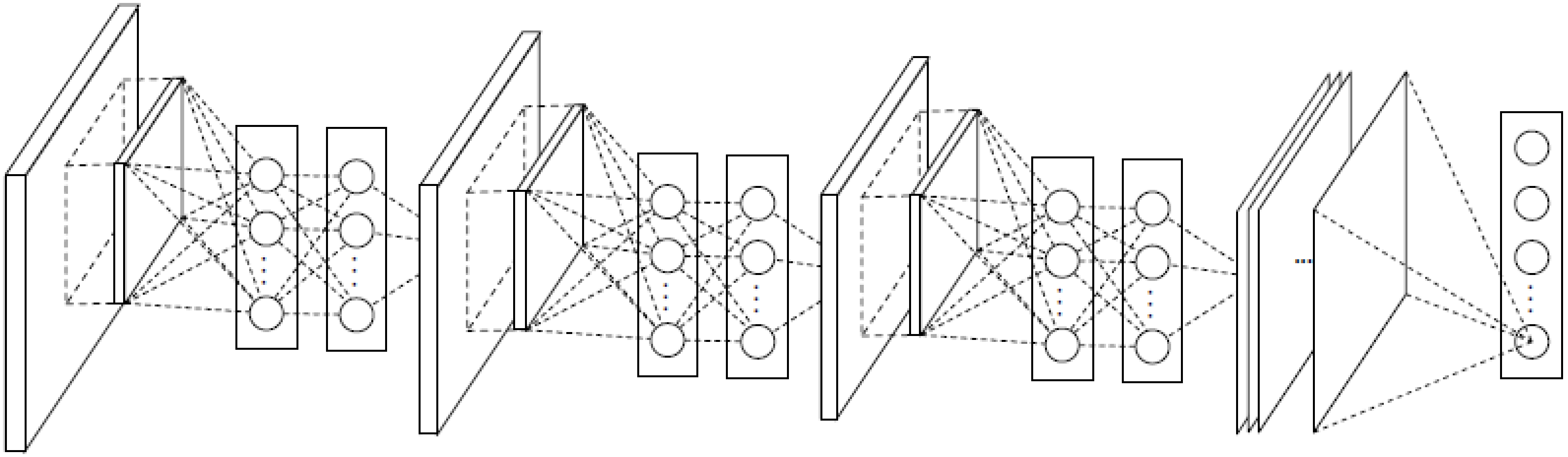
https://towardsdatascience.com/review-nin-network-in-network-image-classification-69e271e499ee

Network in Network Global Average Pooling

- **take the average of each feature map, and the resulting vector is fed directly into the softmax layer.**
- One advantage is that it is more native to the convolution structure by **enforcing correspondences between feature maps and categories.**
- Another advantage is that there is **no parameter** to optimize in the global average pooling thus **overfitting is avoided at this layer.**
- Furthermore, global average pooling sums out the spatial information, thus it is **more robust to spatial translations of the input**.

Lin, M., Chen, Q., & Yan, S. (2013). Network In Network (paper). *ArXiv Preprint*, 10. http://arxiv.org/abs/1312.4400

Network in Network



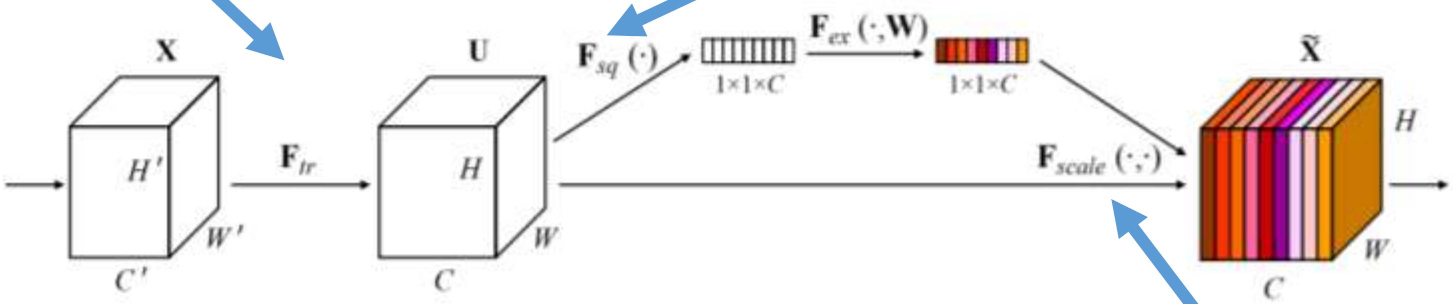
SENet

SENet

多智体人工智能实验室
Multi-Agent Artificial Intelligence Laboratory

$$\mathbf{u}_c = \mathbf{v}_c * \mathbf{X} = \sum_{s=1}^{C'} \mathbf{v}_c^s * \mathbf{x}^s. \qquad (1)$$

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \qquad (2)$$



$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})), \qquad (3)$$

$$\widetilde{\mathbf{x}}_c = \mathbf{F}_{scale}(\mathbf{u}_c, s_c) = s_c\,\mathbf{u}_c, \qquad (4)$$

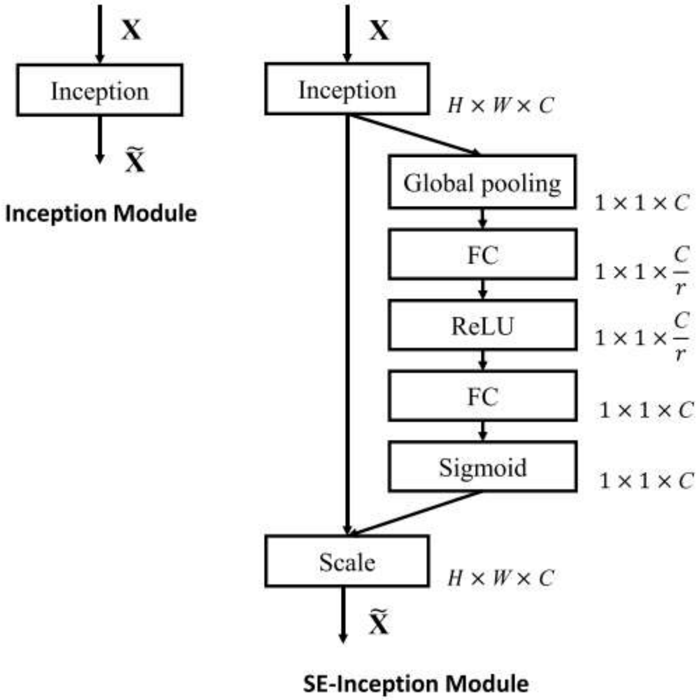


Fig. 2. The schema of the original Inception module (left) and the SE-Inception module (right).

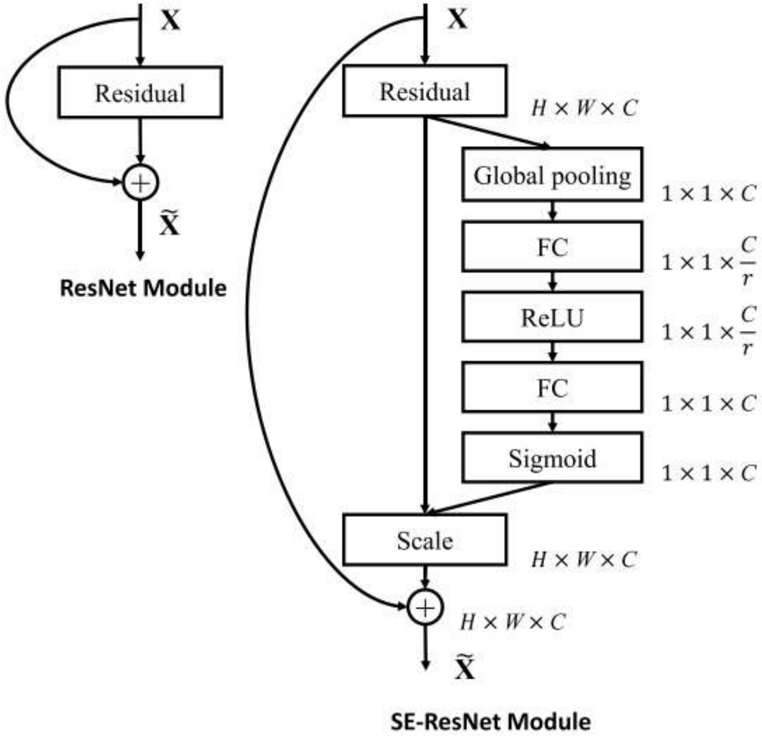


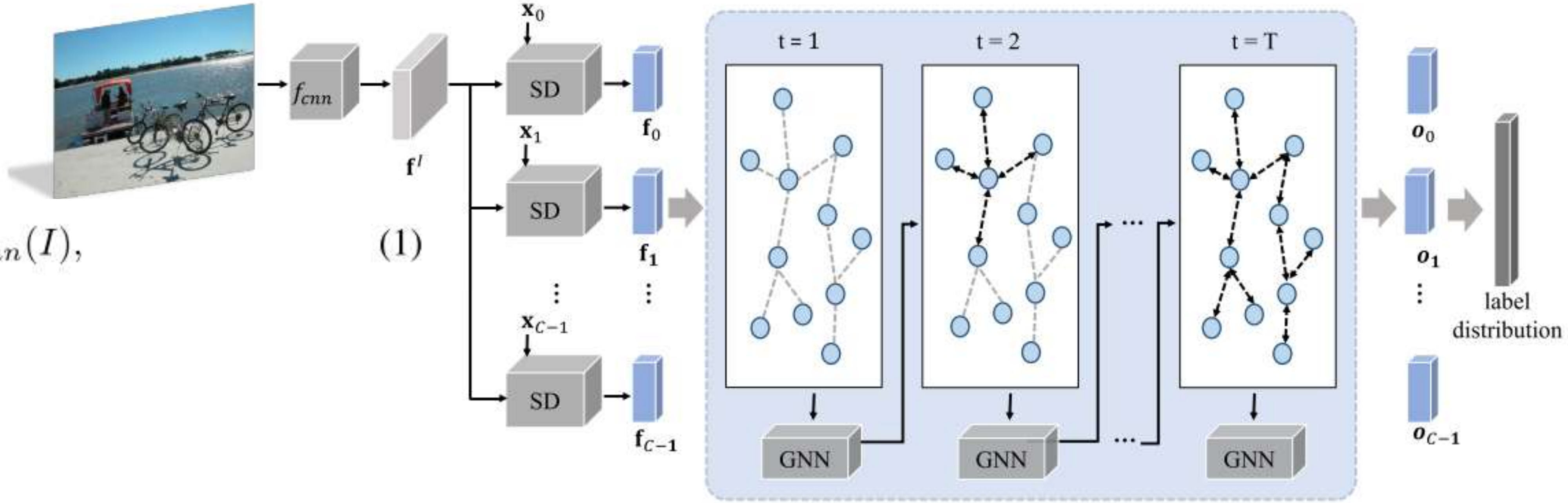
Fig. 3. The schema of the original Residual module (left) and the SE-ResNet module (right).

Bilinear Pooling

Bilinear Pooling

$$\mathbf{x}_c = f_g(w_c), \quad (2)$$

$$\mathbf{f}^I = f_{cnn}(I), \quad (1)$$

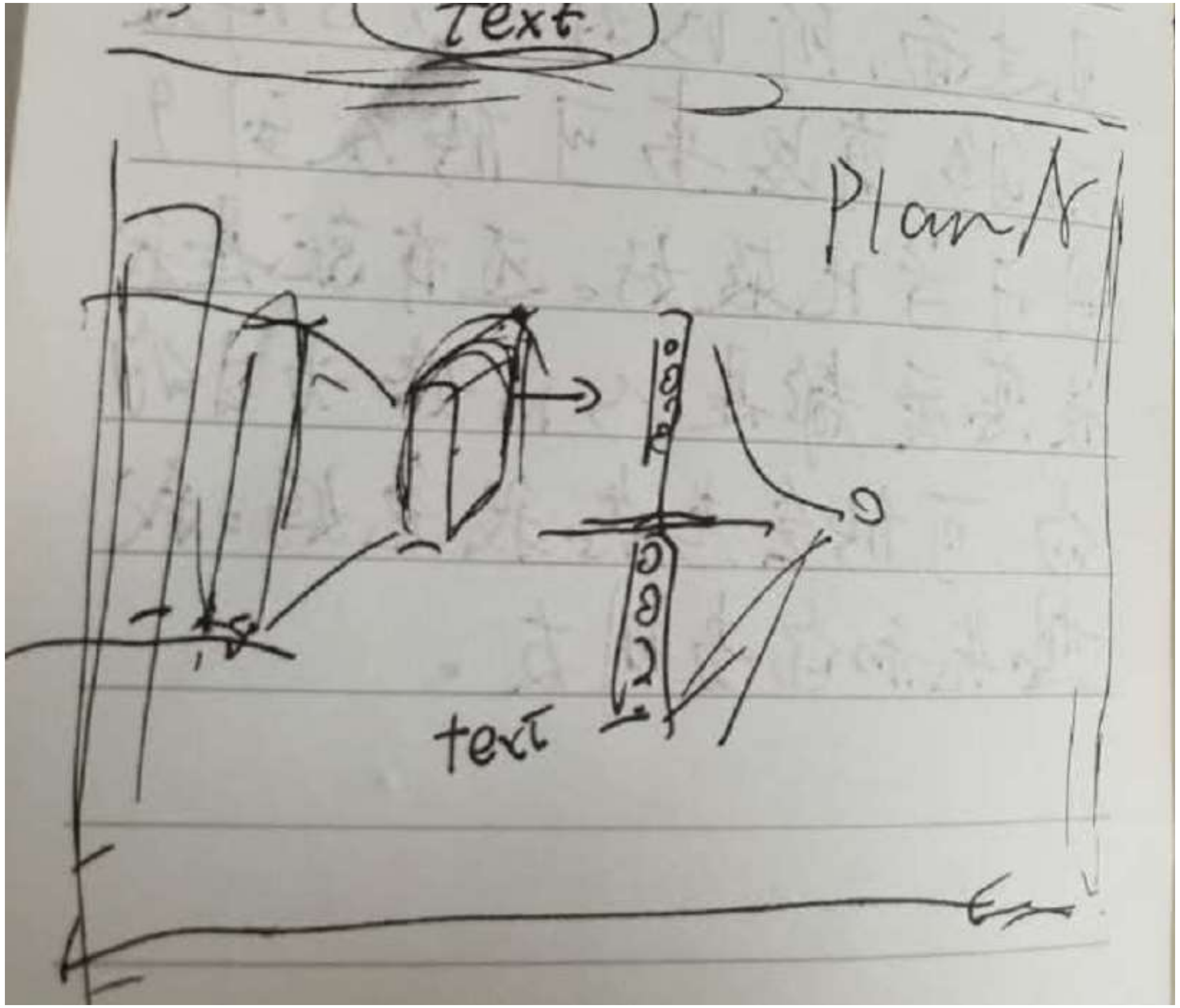


$$\tilde{\mathbf{f}}^I_{c,wh} = \mathbf{P}^T \left(\tanh\left((\mathbf{U}^T\mathbf{f}^I_{wh}) \odot (\mathbf{V}^T\mathbf{x}_c)\right)\right) + \mathbf{b}, \quad (3)$$

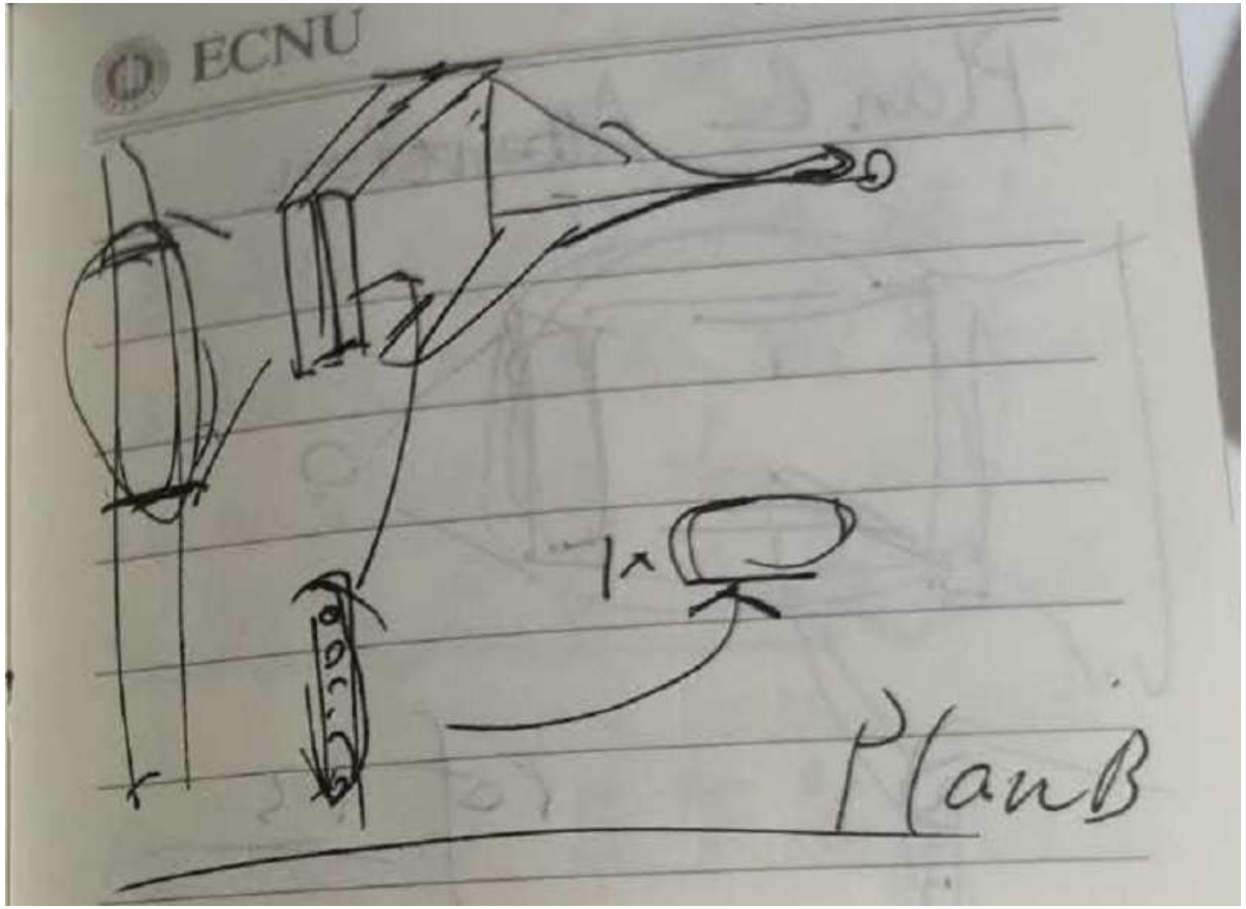
$$\tilde{a}_{c,wh} = f_a(\tilde{\mathbf{f}}^I_{c,wh}). \quad (4)$$

Our ideas

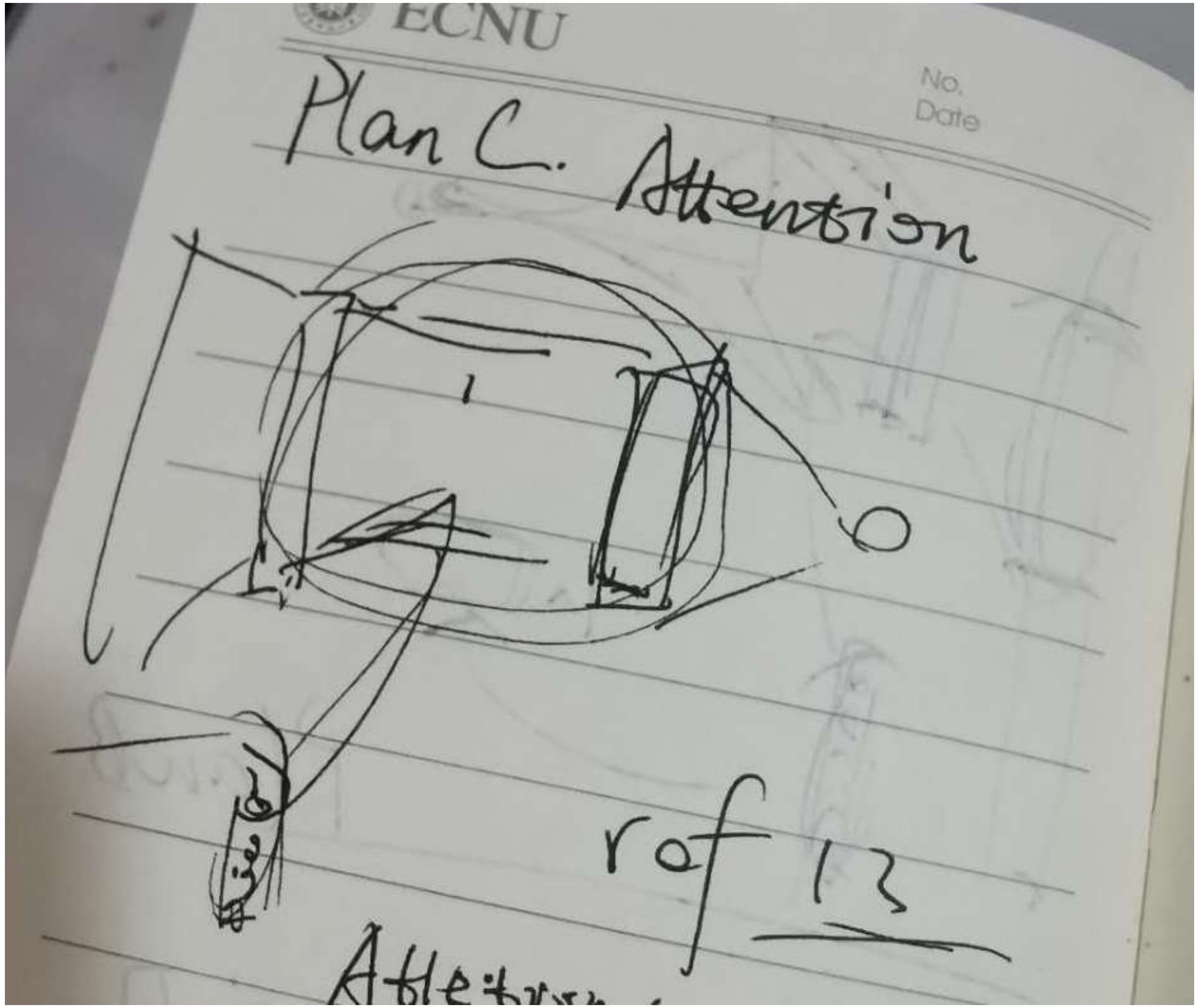
Local Fully-connected



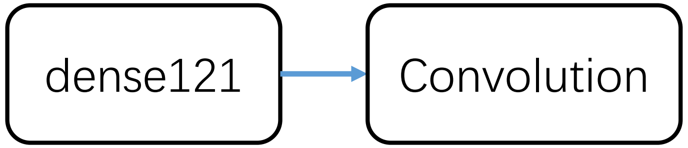
Text -> Convolution

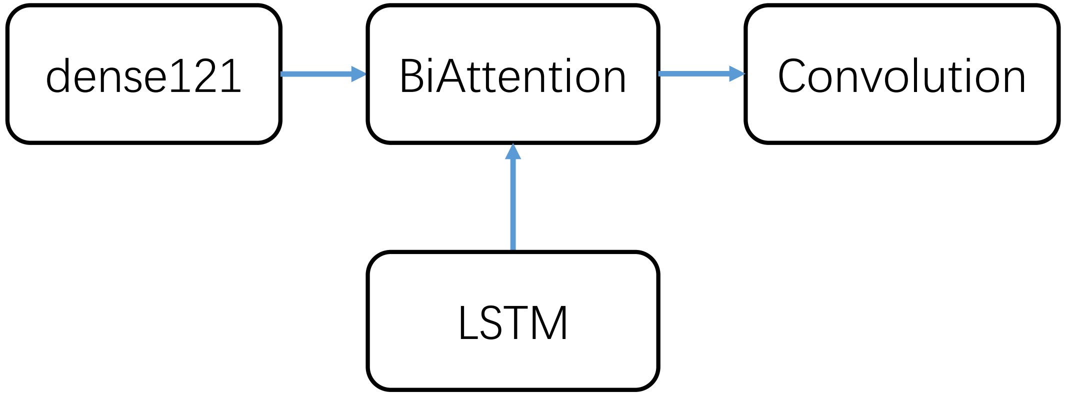


Plan C



With Bilinear Pooling





Thank you

- End