Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Doctoral Consortium Track

Learning from Multimedia Data with Incomplete Information

Renshuai Tao
Beihang University
rstao@buaa.edu.cn

Abstract

Traditional deep learning methods are based on the condition that the data is of high-quality, which means the data information is highly available. However, data in these scenes often have the characteristics of large background noise, lack of sample content, small target, serious occlusion and small number of samples. The application of related tasks in real open scenarios is very important, so it is urgent to make full use of these incomplete information data accurately.

1 Background

With the development of computing power and deep learning algorithms, we can process and apply millions or even hundreds of millions of large-scale data to train robust models. As a result, traditional deep learning methods are based on the condition that the data is of high-quality, i.e., the data information is highly available [Wei *et al.*, 2020]. However, in the real open environment, real visual scenes such as X-ray prohibited items detection, anomaly detection, medical image analysis and other real visual scenes are more complex. Unlike the traditional computing tasks of visual data, the data in these scenes often have the characteristics of large background noise, lack of sample content, small target, serious occlusion and small number of samples. We call this kind of data as "Data of Incomplete Information". The application of related tasks in real open scenarios is very important, so it is urgent to make full use of these incomplete information data accurately.

For example, in the security inspection scene, these items in the X-ray images have the problem of noisy data. As the density if the crowd density increases in public transportation hubs, security inspection has become more and more important in protecting public safety. X-ray scanners, which are adopted usually to scan the luggage and generate the complex X-ray images, play an important role in security inspection scenario. However, security inspectors struggle to accurately detect the prohibited items after a long time highlyconcentrating work, which may cause severe danger to the public. Therefore, it is imperative to develop a rapid, accurate and automatic detection method.

Regarding models, traditional CNN-based models trained through common detection datasets fail to achieve satisfactory performance in this scenario because that different from natural images with simple visual information, X-ray images are characterized by the lacking of strong identification properties and containing heavy noises. This urgently requires researchers to make breakthroughs in models. The effective utilization of X-ray images can provide ideas for learning from multimedia data of incomplete information.

2 Our Studies

In neurobiology, lateral inhibition disables the spreading of action potentials from excited neurons to neighboring neurons in the lateral direction. We mimic this mechanism by designing a bidirectional propagation architecture to adaptively filter the noisy information generated by the neighboring regions of the prohibited items. Also, lateral inhibition creates contrast in stimulation that allows increased sensory perception, so we activate the boundary by intensifying it from four directions and aggregating them into a whole shape.

Therefore, inspired by the mechanism that lateral suppression by neighboring neurons in the same layer making the network more efficient, we propose the Lateral Inhibition Module (LIM). In this section, we will introduce the two core sub-modules, namely Bidirectional Propagation and Boundary Activation, in Section 2.1 and Section 2.2, respectively.

2.1 Bidirectional Propagation

To disable the spreading of noisy information of neighboring regions, we mimic this mechanism by designing the bidirectional propagation architecture. Moreover, we add a dense mechanism to enhance the ability of BP to choose proper information to propagate.

As shown in Figure 1, for the dense top-down pathway on the left of Bidirectional Propagation, up-sampling spatially coarser but semantically stronger feature maps from higher pyramid levels hallucinates higher resolution features. These feature maps are enhanced by the corresponding feature maps from the convolutional layers via lateral connections. Each lateral connection merges feature maps of the same spatial size from the convolutional layer and the top-down pathway. The feature map of low convolutional layer is of lower-level semantics, but its activation is more accurately localized as

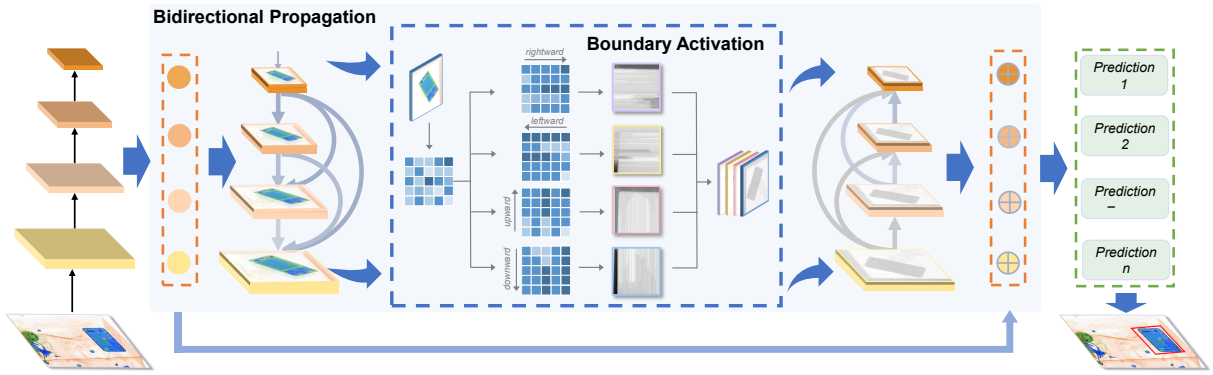

Figure 1: The network structure of Lateral Inhibition Module (LIM). Bidirectional Propagation filters noisy information to suppress the influence from neighbor regions to object regions and Boundary Activation activates the boundary as the identification property, respectively.

it was sub-sampled fewer times. Further, we construct the dense connections to ensure maximum information flow.

Specifically, to preserve the feed-forward nature, $\mathcal{F}^l(\mathbf{x})$ obtains additional inputs from the feature maps $\mathcal{F}^{l+1}(\mathbf{x}), \cdots, \mathcal{F}^L(\mathbf{x})$ of all preceding layers and passes on its own feature-maps to the feature maps $\mathcal{F}^{l-1}(\mathbf{x}), \cdots, \mathcal{F}^1(\mathbf{x})$ of all subsequent layers. Figure 1 illustrates this layout schematically. We define $\mathcal{U}^m(\cdot)$ as the up-sampling operation (2^m times) and $\mathcal{V}(\cdot)$ as a 1×1 convolutional layer to reduce channel dimensions. The process can be formulated as follows:

$$\mathbf{A}^l = \mathcal{V}\left(\mathcal{F}^l(\mathbf{x})\right) + \sum_{m=1}^{L-l} \mathcal{U}^m\left(\mathbf{A}^{l+m}\right), \tag{1}$$

where $\mathbf{A}^l$ refers to the feature map generated of the m-th layer of the Single-directional Propagation (the left part of Bidirectional Propagation).

as the Figure 1 illustrated, suppose the input feature map $\mathbf{B}^l$ refers to the feature map generated in Eq. (4) (Boundary Aggregation will be introduced in the following section). Similar to the previous definition, $\mathcal{D}^m(\cdot)$ is the down-sampling operation (2^m times). This process can be formulated as follows:

$$\mathbf{C}_{\mathrm{t}}^l = \mathcal{V}\left(\mathbf{B}^l\right) + \sum_{m=1}^{l-1} \mathcal{D}^m\left(\mathbf{C}_{\mathrm{t}}^{l-m}\right), \tag{2}$$

$$\mathbf{C}^l = \mathbf{C}_{\mathrm{t}}^l + \mathcal{F}^l(\mathbf{x}), \tag{3}$$

where $\mathbf{C}_{\mathrm{t}}^l$ refers to the output of l-th layer of the boundary-enhanced bottom-up pathway and $\mathbf{C}^l$ refers to the feature map generated of the l-th layer of Bidirectional Propagation. Finally, we convey the output of MuBo $\mathbf{C}^l$ to the following prediction layers.

2.2 Boundary Activation

To mimic the mechanism that lateral inhibition creates contrast in stimulation that allows increased sensory perception, we activate the boundary information by intensifying it from four directions inside the feature maps outputted by each layer and aggregating them into a whole shape.

Motivated by the schematic diagram, we design the module Boundary Activation to perceive the sudden changes of boundary and its surroundings. Suppose we want to capture the left boundary of the object in the feature map $\mathbf{A}^l \in R^{H \times W \times C}$ (the output of left part of Bidirectional Propagation). $\mathbf{A}_c^l$ donates the c-th channel of $\mathbf{A}^l$. Further, $\mathbf{A}_{ijc}^l$ refers to the location (i, j) of the feature map $\mathbf{A}_c^l$. To determine whether there is a sudden change between a position and the left of the point, the right-most point $\mathbf{A}_{iWc}^l$ traverses to the left. The process of perceiving the left boundary can be formulated as Eq. (4).

$$\mathbf{B}_{ijc}^l = \begin{cases} \mathbf{A}_{iWc}^l & \text{if } j = W, \\ \max\left\{\mathbf{A}_{ijc}^l, \mathbf{A}_{i(j+1)c}^l, \ldots, \mathbf{A}_{iWc}^l\right\} & \text{otherwise,} \end{cases} \tag{4}$$

where the $\mathbf{B}_{ijc}^l$ refers to the location (i, j) of c-th channel of the feature map $\mathbf{B}^l$ after Boundary Activation.

3 Discussion

Our long-term research interest is to investigate much more strategies to learn from multimedia data with incomplete information (noisy data). We believe that these works can provide exciting ways out of the dilemma that satisfactory performance relies heavily on high-quality dataset, whose construction is time-consuming and labour-intensive.

Now we are attempting to improve the accuracy of novel models for prohibited items detection in noisy X-ray images. Aiming to perform more analysis about this scenario, we also try to construct such datasets, which have been rare studied.

References

[Wei *et al.*, 2020] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 138–146, 2020.