Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Perspective:

On visual understanding

Yunhe PAN[1,2]

[1]*Institute of Artificial Intelligence, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

[2]*Zhejiang Lab, Hangzhou, China*

E-mail: panyh@zju.edu.cn

Received Aug. 16, 2021; Revision accepted Sept. 13, 2021; Crosschecked Sept. 14, 2021; Published online Sept. 16, 2021

https://doi.org/10.1631/FITEE.2130000

1 Problems and development in the field of visual recognition

From the beginning of artificial intelligence (AI), pattern recognition has been an important aspect of the field. In recent years, the maturity of deep neural networks (DNNs) has significantly improved the accuracy of visual recognition. DNN has been widely used in applications such as medical image classification, vehicle identification, and facial recognition, and has thus promoted the development of the AI industry to a climax. However, there are currently critical defects in visual recognition based on DNN technology. For example, these networks usually require a very large amount of labeled training data, and have weak cross-domain transferability and task generalization. Their learning and reasoning processes are still hard to understand, which leads to unexplainable predictions. These challenges present an obstacle to the development of AI research and application.

If we look at the current visual recognition technology from a larger and broader perspective, we can find that the above defects are fundamental, because the currently used DNN model needs to be trained with a large amount of labeled visual data, and then used in the process of visual recognition. In essence, it is a classification process based on data statistics and pattern matching (Krizhevsky et al., 2017), so it is heavily dependent on training sample distribution. However, to have interpretability and transferability, visual classification is not good enough, while visual understanding becomes indispensable.

2 Three-step model of visual understanding

Visual recognition is not equivalent to visual understanding. We propose that there are three steps in visual understanding, of which classification is only the first. After classification, one proceeds to the second step: visual parsing. In the process of visual parsing, the components of the visual object and their structural relationship are further identified and compared. Identification involves finding components and structures in visual data that correspond to the components and structures of known visual concepts. Parsing verifies the correctness of the classification results and establishes the structure of visual object data. After completing visual parsing, one proceeds to the third step: visual simulation. In this step, predictive motion simulation and operations including causal reasoning are carried out on the structure of the visual objects to judge the rationality of meeting physical constraints in reality, so as to verify the previous recognition and parsing results.

Prof. Yunhe PAN
Editor-in-Chief

We can take a picture of a cat as an example to

© Zhejiang University Press 2021

illustrate the modeling process of visual understanding. The process is as follows:

1. Recognition: It is a cat. Extract the visual concept of the cat and proceed to the next step; otherwise, stop here.

2. Parsing: Based on the structure contained in the visual concept, identify whether the cat's head, body, feet, tail, and their relationships are suitable for the cat concept. If not, return to step 1 for re-identification; if yes, proceed to the next step.

3. Simulation: Simulate various activities of the cat to investigate whether the cat's activities in various environments can be completed reasonably. If not, return to step 2; if yes, proceed to the next step.

4. End visual understanding: Incorporate the processed structured data into the knowledge about cats.

3 Characteristics of the three-step visual understanding model

To further understand the above-mentioned three-step visual understanding model, we will further discuss some of its characteristics:

1. The key step in visual understanding is visual parsing. This is an identification of the components contained in the object according to a conceptual structure based on the visual concept (Pan, 2019), obtained by visual recognition. Parsing a visual object, in order from top to bottom, is a process of identifying and constructing visual data from the root of the concept tree to the branches and leaves.

2. Human visual parsing tasks are often aimed only at the main components of concepts. The main components have existing, commonly used names. For subsidiary parts that have not been described in language, such as the area between the cheekbones and chin of the face, only experts specialized in anatomy (such as doctors or artists) have professional concepts and memories. Therefore, visual parsing is a cross-media (Yang et al., 2008) process that incorporates multiple knowledge (Pan, 2020b) including vision and language.

3. Visual knowledge (Pan, 2019) is essential for visual parsing and visual simulation, because the visual concept structure provides a reliable source for component identification and comparison. Parents and teachers play a large role in establishing visual knowledge. When they say to a child, "Look, this is a kitten. Kittens have pointed ears, round eyes, long whiskers, and four short legs. When they run fast and leap high, they can catch a mouse," they are guiding children in constructing basic visual knowledge in their long-term memory.

4. Visual data that have been understood have actually been structured to form visual knowledge. Such visual knowledge can easily be incorporated into long-term memory. For example, when one sees a cat whose head is very small, or whose fur color and markings are unusual, or who has a particular gait, this information may be included in one's "cat" memory by expanding the concept of "cat" (Pan, 2019). The category of visual concepts is very important, and its extent reflects the general degree of knowledge. In fact, it is not always useful to collect a large amount of sample data to train a DNN model. However, the more widely distributed and balanced the data are within a concept category, the better, because the robustness and generalization ability of the model trained based on such sample data are stronger.

5. The learned visual information can naturally be explained, because it has deep structural cognition; it can also be used for transfer learning because the semantic concepts have cross-media relevance. This semantic information can clearly indicate the reasonable direction of transferable recognition.

4 Advancing visual recognition to visual understanding

Visual understanding is important, because it can potentially work with visual knowledge (Pan, 2019) and multiple knowledge representation (Pan, 2020b) to open a new door to AI research. Visual understanding involves not only in-depth visual recognition, but also thorough learning and application of visual knowledge (Pan, 2020a). AI researchers have been studying visual recognition for more than half a century. Speech recognition, a research task started in parallel with visual recognition, moved on to analysis of words, sentences, and paragraphs quite early, and has successfully developed human-computer dialogue and machine translation, setting a well-known milestone. Therefore, we suggest that it is necessary to advance visual recognition to visual understanding,

and that this is an appropriate time to target this deeper visual intelligence behavior.

Acknowledgements

I am grateful to Profs. Yueting ZHUANG, Fei WU, Weidong GENG, Yi YANG, Lingyun SUN, and Siliang TANG for providing valuable suggestions.

References

Krizhevsky A, Sutskever I, Hinton GE, 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM*, 60(6):84-90. https://doi.org/10.1145/3065386

Pan YH, 2019. On visual knowledge. *Front Inform Technol Electron Eng*, 20(8):1021-1025. https://doi.org/10.1631/FITEE.1910001

Pan YH, 2020a. Miniaturized five fundamental issues about visual knowledge. *Front Inform Technol Electron Eng*, early access. https://doi.org/10.1631/FITEE.2040000

Pan YH, 2020b. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3):216-217. https://doi.org/10.1016/j.eng.2019.12.011

Yang Y, Zhuang YT, Wu F, et al., 2008. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Trans Multim*, 10(3):437-446. https://doi.org/10.1109/TMM.2008.917359