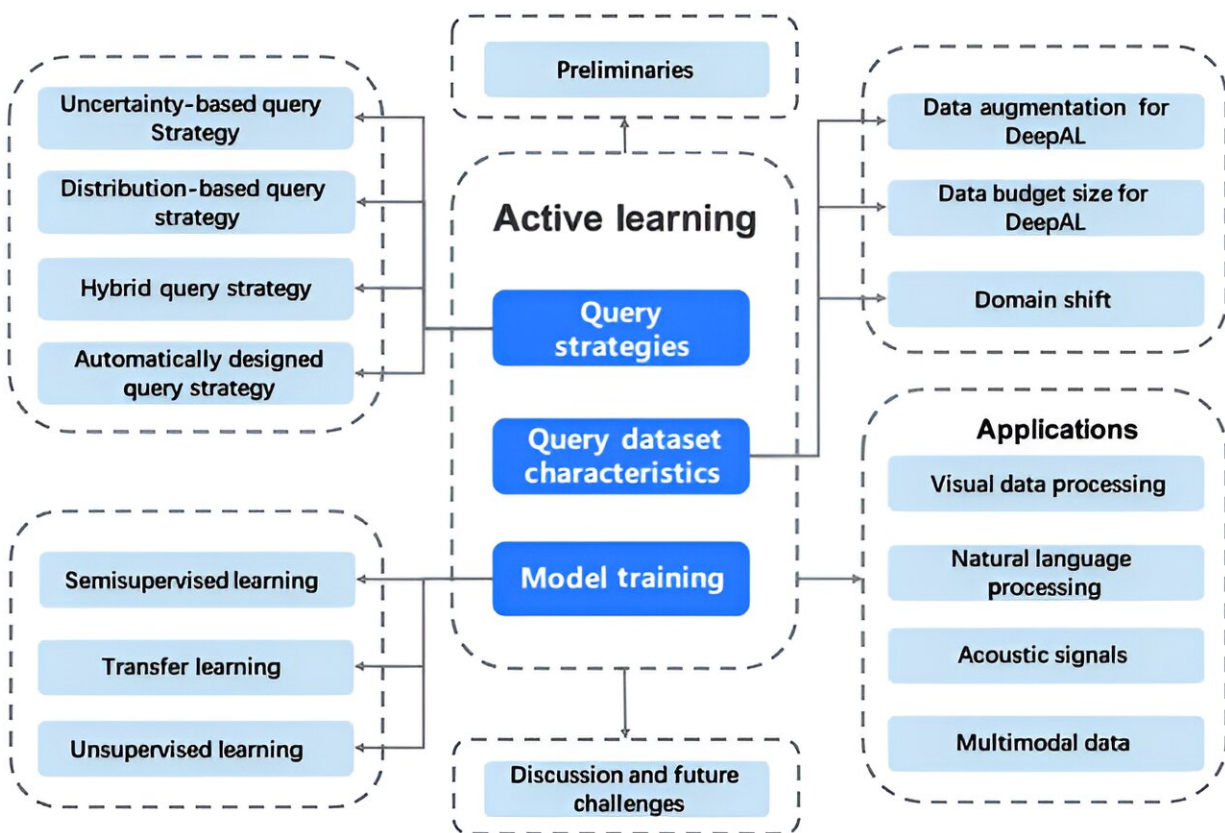# When deep learning meets active learning in the era of foundation models

December 1 2023

Schematic structure of the deep active learning survey, which covers key concepts, applications, and future challenges. Credit: *Intelligent Computing* (2023). DOI: 10.34133/icomputing.0058

A Chinese research team wrote a review article on deep active learning,

an increasingly popular method of combining active learning with deep learning for sample selection in the training of neural networks for artificial intelligence tasks. It was published in *Intelligent Computing*.

Given that research on deep active learning techniques in the context of foundation models is limited, this review offers some insights into this topic. It surveys existing deep active learning approaches, applications, and especially challenges "in the era of foundation models," concluding that it is necessary to develop customized deep active learning techniques for foundation models.

Recently, the success of foundation models has called attention to the data-intensive nature of artificial intelligence. Foundation models are generally built with deep learning technologies and trained on massive labeled datasets. Only with accurate data labeling or annotation can models make accurate predictions and adapt to various downstream tasks. However, producing such data is laborious, difficult, and expensive.

This is where deep active learning comes in. Using active learning to train deep learning models can effectively reduce the heavy labeling work because active learning only selects and labels the most valuable samples. As a result, deep active learning can smooth out the learning process and bring down the cost, contributing to "resource-efficient data that are robustly labeled."

In line with the structure of a typical active learning framework, which includes query data, query strategy, human annotation, and model training in a cycle, deep active learning approaches relate to query strategies, query dataset characteristics, and model training.

Effective query strategies are the key to selecting the most valuable samples for data annotation. Active learning query strategies generally

come in three categories: membership query synthesis, stream-based sampling, and pool-based sampling; this categorization is based on the flow of unlabeled samples to the data annotator.

For deep active learning algorithms, on the other hand, there are four kinds of strategies: uncertainty-based, distribution-based, hybrid, and automatically designed.

Uncertainty-based strategies identify the samples with the highest uncertainty, distribution-based strategies focus on the underlying structure of the data to identify representative samples, and hybrid strategies use a combination of uncertainty-based and distribution-based selection metrics; all three types are designed manually and, thus, are not easily adapted for deep learning models and are over-dependent on human expertise. These problems can be addressed by automatically designed strategies that use meta-learning or reinforcement learning.

Query strategies must be tailored to various dataset characteristics, such as the size, budget, and distribution of the query dataset, which is a subset carefully selected from a larger dataset to be labeled. Specifically, data augmentation is often used to improve the diversity and quantity of labeled training data; different strategies should be adopted to suit different budget sizes and to cope with the mismatch between the distributions of labeled and unlabeled data—that is, the so-called domain shift problem.

For model training, the authors discussed how to combine deep active learning with current data-heavy mainstream methods, including supervised training, semi-supervised learning, transfer learning, and unsupervised learning, to achieve optimal model performance.

The applications of deep active learning across various scenarios were then introduced, especially those involving costly, time-consuming data

collection and annotation. As the authors observed, deep active learning has been used to process not only single-modal data such as visual data, natural language, and acoustic signals but also abundant multi-modal data.

However, the authors also pointed out that the majority of current deep active learning methods concentrate on task-specific models instead of comprehensive, data-intensive foundation models.

To better integrate deep active learning into foundation models and maximize joint performance, several key challenges in training and refining foundation models need to be addressed, including data quality evaluation, active finetuning, efficient interaction between data selection and annotation, and the development of an efficient machine learning operations system.

Provided by Intelligent Computing