School of Informatics

Deep Probabilistic Active Learning for Reliable Modeling in Image Recognition

B231968
January 2023

Abstract

Active learning and skeptical learning are effective techniques for reducing annotating costs by improving query strategy. Those implement human-in-the-loop systems and cover a broad range of research topics where high-quality annotated large datasets are unavailable. However, those techniques are limited in representing high-dimensional spatial image features and complex relationships between a small amount of data annotated by the user and uncertainty over unseen data. In this article, I review the emerging interactive learning approaches, with the dual goal of helping readers better understand the existing techniques and stimulating new work in the burgeoning area of deep learning with probabilistic circuits for interactive learning.

Date: Sunday 22nd January, 2023

Supervisor: Georgios Papoudakis

1 Introduction

Interactive learning is the machine learning method that uses human-in-the-loop systems to predict outputs with less-supervision. Active Learning (AL) and Skeptical Learning (SKL) are the primary types of interactive learning methods used in probabilistic approaches for recognizing images from raw image inputs via selected image feature subsets. However, these methods have their limitations in representing high-dimensional spatial information and complex relationships between a small amount of data labeled by user and uncertainty over unseen data. Inspired by the successful application of Bayesian Convolutional Neural Networks (BCNNs) to active learning and by the hierarchically structured architecture of neural networks, deep learning and probabilistic circuits techniques have also been applied successfully to interactive learning, as reported in recent literature [1] [2]. This paper reviews these emerging interactive learning approaches, with the dual goal of helping readers better understand the existing techniques and stimulating new work in the burgeoning area of deep learning with probabilistic circuits for interactive learning.

AL with human-in-the-loop systems covers a wide range of research topics in interactive learning where a large number of high-quality annotated datasets by human experts are unavailable, such as medical image recognition (diagnose lesions from images), speech recognition (predict speakers from input speech), and information extraction (classify classes from documents). These topics have the common goal of selecting the most informative uncertain samples and differ in the forms of inputs. Although modeling of those is equally significant for interactive learning tasks, this review focuses on image data modeling techniques for AL and SKL.

SKL is built on AL to address its limitation by asking users to check or fix labels. The probabilistic approaches using AL and SKL have been shown to produce highly reliable and relatively accurate predictions with a selected small subset of data [3][4][5][6]. However, the representation ability of image features to model uncertainty is, of course, limited compared to deep learning techniques. Inadequate uncertainty modeling is one of the main reasons for this deficiency [1]. Take kernel-based active learning, for example. This approach typically uses kernel functions to represent the space of image features given raw image data [3][4]. Based on the representation, an uncertainty measure and/or an information density measure are used to model uncertainty.

Given the success of applying deep learning to various vision tasks, the approach has been applied to uncertainty modeling in AL image classification. Specifically, BCNN is a computer vision model that integrates Bayesian inference in deep learning architecture. As an inference model, BCNNs can be used for AL to perform much better in terms of accuracy and the number of queried labels than the classical models (e.g., kernel-based AL models) [1].

Since early 2010, probabilistic circuits (PCs) have emerged as a new field of machine-learning research and have attracted the attention of many probabilistic modeling researchers. PCs refer to a class of expressive and tractable machine-learning techniques that exploit deep computational graphs of large structured output spaces for measuring model uncertainty with exact computation of a query class (e.g., marginals, MAP inference, expectations, etc.) [7]. Depending on the degree of structural constraints, different families of PCs (e.g., Conditional Randomized Interactive Skeptical Probabilistic circuits (CRISPs) [2], Sum-Product Networks (SPNs) [8], Probabilistic Sentential Decision Diagrams (PSDDs) [9], Cutset Networks (CNets) [10], Arithmetic Circuits (ACs) [11], etc.) have been intensively studied and explored by probabilistic modeling researchers in recent years. Especially, CRISP is a graphical model for incremental active and skeptical learning in the wild. As a high-capacity expressive model, CRISPs are expected to perform much more reliably and efficiently than conventional shallow structures.

Considering its relevance to interactive learning models, I include CRISPs as an example of deep probabilistic circuit models in this review.

This article first reviews the classical framework for interactive learning, including kernel-based active and skeptical learning, focusing on vision modeling and not on speech or text. I then analyze the limitations of those approaches. Also, I introduce the crucial techniques and models of (deep) neural networks relevant to interactive learning, including BCNNs and CRISPs. Subsequently, emerging active learning approaches using deep learning techniques for vision modeling are reviewed, with an analysis of their motivations and a description of their implementations. Finally, I discuss the remaining issues associated with current deep active learning methods with image data and suggest future directions in this field.

2 Literature Review

2.1 Conventional Interactive Learning using Probabilistic Models

2.1.1 Kernel-Based Active Learning

Active Learning (AL) emerged in the mid-1990s [12]. In this framework, the relationship between a small amount of data labeled by user and uncertainty over unseen data is modeled using probabilistic models. Support Vector Machines (SVM) with margin-based uncertainty and Gaussian Processes (GPs) with adaptive-based uncertainty are the commonly concentrated probabilistic models [1]. Those learning methods use kernels to express input features and select unlabeled examples that are hardest to classify to perform AL. This approach is known as kernel-based active learning. In a kernel-based AL, raw images or low dimensional features (such as SIFT features [13]) are fed to kernels and the model uncertainty is represented by Query Strategy Frameworks as an uncertainty measure [14]. Uncertainty Sampling and Expected Error Reduction are the most popular frameworks for deciding labels to query. Kernel-based interactive learning approach is able to transform the original feature space into a more effective one compared to other kernel metric learning approaches [15]. As a result, it dramatically reduces the number of labels required to train a model so cost and time.

Best-versus-Second Best (BvSB) [3] employs the SVM with uncertainty sampling-based method and feeds kernel matrices features to the model. It classifies multi-class images by SVM using class membership probability estimates for AL. In BvSB, an uncertainty measure generalizes margin-based uncertainty to the multi-class case so that active learning handles a large number of classes and large data sizes efficiently. BvSB is able to easily handle multi-class problems and work without knowledge of the number of classes, which could increase with time. Also, this one-versus-one SVM approach (a classifier trained for each pair of classes) makes the model computationally and interactively efficient compared to the conventional one-versus-all classification approach [16].

To estimate class membership probabilities of the unlabeled examples, the authors use the difference between the probability values of the two classes having the highest estimated probability value as a measure of uncertainty. It is essentially a comparison of the best guess and the second best guess of an example to be labeled and is more robust than relying on entropy score, which only considers the best guess. This uncertainty-sampling-based system typically fails to take information in the large amount of unlabeled instances into account and the ability to assess an instance is limited to the small set of labeled instances. This leads to the myopic decision problem, such as querying outliers that are not informative for a model to learn and

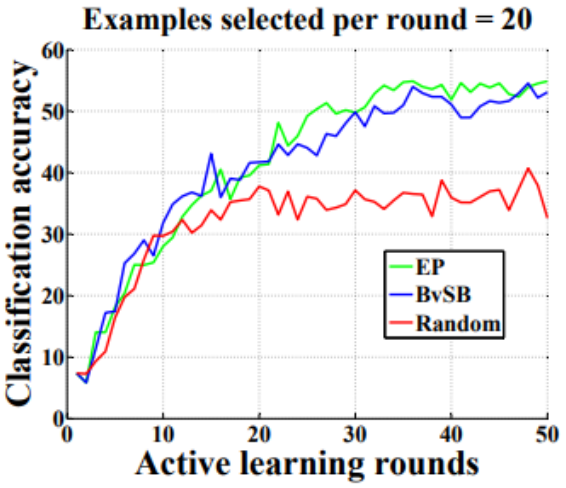

Figure 1: BvSB Active Learning results (blue line) on the 13 natural scene categories dataset [3].

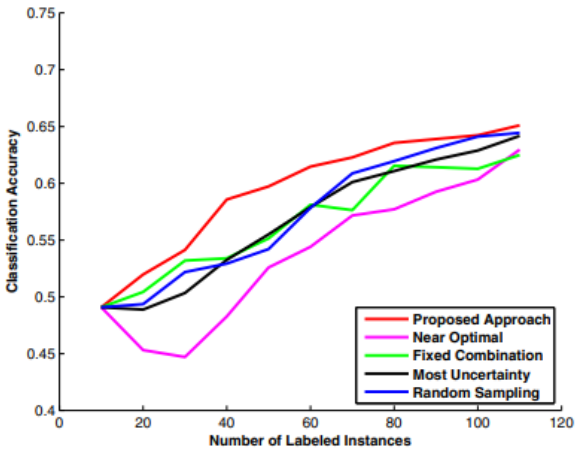

Figure 2: GPs Active Learning results (red line) on a randomly sampled 10-class subsets of the natural scene categories dataset [4].

information about the unlabeled data distribution will be ignored from the training process. To deal with this issue, an information-density-based measuring technique using GPs [4] is applied along with the uncertainty sampling to take both labeled and unlabeled instances into account, where uncertainty measure and and information density measure are represented by a combined framework. Specifically, the authors feeds dense SIFT features and compute the following three steps for uncertain label selection. 1. Uncertainty measure $f(\boldsymbol{x}_i)$ to capture the informativeness of labeled instances: Determine the conditional distribution $P(y|\boldsymbol{x}_i, \theta_L)$ for the set of all class values Y using this model to compute the measure, where y is the class value, $\boldsymbol{x}_i$ is the candidate instance, and θ_L is the classification model trained over the labeled set L. 2. Information density measure $d(\boldsymbol{x}_i)$ to approximate the unlabeled instances distribution with GPs: Produce σ_i^2 and $\sigma_{i|U_i}^2$, the covariance of $\boldsymbol{x}_i$ and of $\boldsymbol{x}_i|U_i$ respectively, using a valid Kernel function $\sigma_i^2 = K(\boldsymbol{x}_i, \boldsymbol{x}_i)$ to compute the measure, where U_i is the index set of unlabeled instances after removing i from U, such that $U_i = U - i$. 3. Combine 1 and 2 to integrate the strengths of both $f(\boldsymbol{x}_i)^\beta d(\boldsymbol{x}_i)^{1-\beta}$ where $0 \leq \beta \leq 1$ is a parameter that controls the degree of relative importance of the two measures. This combined measure is used to pick the best instance that is most uncertain to classify and very informative about the remaining unlabeled instances.

Figure 1 and 2 show the SVM model (BvSB) requires about 600 examples to achieve classification accuracy of 50% while the GPs model requires less than 20 examples for the similar performance. Note that the GPs model uses the randomly sampled 10-class subsets of Natural Scene Dataset while the SVM model uses whole Natural Scene Dataset. Although the number of classes are restricted from 13 to 10 classes in the GP model, the performance improvement is significant.

2.1.2 Skeptical Learning

Skeptical Learning (SKL) aims to retrieve clean label from an annotator by asking to reconsider feedback if the machine is confident that an obtained example is mislabeled. Different from the Active Learning (AL), which blindly accepts the annotator's supervision, SKL implements

a noise handling strategy for interactive learning. Some probabilistic approaches to SKL have been studied since 2019, such as random forests (RFs) [5], GPs [6], etc.

Skeptical Supervised Machine Learning (SSML) [5] as RF-based SKL is the initial introduction of the concept. SSML is able to minimize the impact of mislabeling due to untrustfulness of human annotators. In addition, this confidence-score-based approach makes a skeptical supervised learning strategy generalizable compared to the classic active learning approach. SSML applies Ground Knowledge (e.g., a professor's office is part of a department building) and Schematic Knowledge (e.g., department buildings are always inside the Univesity premises) for conflict resolution. It is to interpret user's answers and thus infer the meaning of different labels used by machine and users. The SSML system with a conflict resolution component means that the overarching system fixes examples mislabeled by a human annotator and runs with progressively higher confidence.

The SSML algorithm sequentially activates three (Train, Refine, and Regime) modes. In the Train mode, the system performs active learning and trains a predictor using input-output pair. The predictor computes the expected probability of contradicting the user using the predictor confidence score $c^p_{\hat{y}}$ in a predicted label $\hat{y}$ for input $\boldsymbol{x}$, the user confidence score c^u_y in a label provided by the user for that input, and the parameter θ. The predictor becomes confident to challenge the user when the expected probability exceeds the threshold. Then, Train mode is switched to Refine mode, and the system starts asking feedback for all inputs. When the expected probability of querying the user falls the threshold and the predictor becomes confident for the current labels, Refine model is switched to Regime mode, and the system backs to the active learning process and queries the user with a loose threshold.

An RF-based SKL system is generally sensitive to task-shifted noises where the query budget is limited or the F1 score attainment is tasked. This leads to over-confident problems, such as not querying the user after seeing examples from mostly one class and the issue that many aggressive queries will be sent from the active learning part to reach a tasked F1 score.

To overcome this issue, an incremental skeptical classification technique in the wild is built on Gaussian Processes (ISGP) [6] to classify real-world examples where the supervision is noisy and the number of classes grows over time. ISGP models the classification task using the GP posterior $P(f(x) \geq 0|x)$ by computing the CDF of a standard normal distribution $\Phi(\frac{\mu(x)}{\sigma(x)})$ using $\sigma(x) = \sqrt{k(x,x)}$, the mean function $\mu(x)$, and the covariance function $k(x,x)$ conditioned on (x,y). Building on the previous work, Incremental Multi-class GPs (IMGP) [17], ISGP combines the ℓ-th GP posterior $\mathrm{P}_\ell(1|\mathrm{x_t})$ with a soft-max to obtain a multi-class posterior to predict a label at t-th iteration $\hat{y_t} = \mathrm{argmax}_\ell \frac{1}{Z} \mathrm{expP}_\ell(1|\mathrm{x_t})$, where Z is a normalization factor $\Sigma_{\ell'} \mathrm{expP}_{\ell'}(1|\mathrm{x_t})$ and $\ell \in Y_{t-1}$.

This way, ISGP can avoid pathological behaviors by preventing the model from being over-confident in regions far away from the training set. Figure 3 depicts such a symptom: the F1 score of SSML (SRF) plateaus at roughly 70 iterations, while the one of ISGP keeps improving up to iteration 200. In addition, incremental updates (i.e., compute precision matrix Γ_{t+1} from Γ_t without any matrix inversion [17]) reduce cost from $O(t^3)$ to $O(t^2)$ per iteration, which addresses GPs limitation in scalability and makes the real-world experiment with increasing classes feasible.

Once $\hat{y_t}$ is predicted, ISGP decides whether to request the label of x_t while prioritizing more uncertain labels to save the labeling cost as the model improves. ISGP sends active queries to the user if $a_t = 1$ by sampling a_t from a Bernoulli distribution with the parameter $\alpha_t = P_{\hat{y_t}}(f(x_t) \leq 0|x_t)$. This randomized querying strategy prevents ISGP from trusting the model too much. Analogous to the above active learning part, ISGP next sends skeptical queries

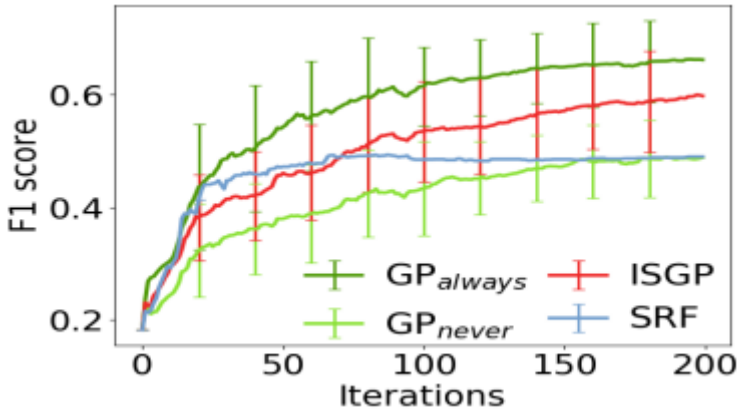

Figure 3: Results on location prediction. F1 score of ISGP [6], SSML (SRF) [5], and GP baselines [17] on a real-world and synthetic dataset.

to the user if $\gamma_t = 1$ by sampling γ_t from a Bernoulli distribution with the parameter $\gamma_t = P(f_{\hat{y}_t}(x_t) - f_{\tilde{y}_t}(x_t) \geq 0)$, where $f_{\hat{y}_t}(x_t)$ is the function value for the predicted label at a test point of x_t and $f_{\tilde{y}_t}(x_t)$ is the one for the annotated label by the user. Finally, ISGP updates the model using a fixed label y'_t replied to by the user and repeats the process.

This model's uncertainty-based approach using only two hyper-parameters in covariance function makes the model simple and skeptical query decision far more robust compared to thresholded-confidence-based approach in the original SKL formation [5].

2.2 Uncertainty Modeling using Deep Active Learning Techniques

Given the success of applying deep learning to various vision tasks, the approach has been applied to uncertainty modeling in interactive image classification to achieve better uncertainty representation. Studies in applied deep learning techniques to active learning (AL) had not been extensive until recently. From 2017 to 2021, several articles on the topic of interactive image recognition [1][18][19][20] have been published. They reported positive results that the deep learning techniques improved representation of model uncertainty and its performance. These deep learning approaches typically focus on data subset selection techniques, and I review one of those that uses the Bayesian framework in this section.

2.2.1 Uncertainty Representation using Deep Bayesian Model

In deep AL approach, the deep learning techniques are applied to the input-to-uncertainty representation step of interactive learning modeling for AL, i.e., to represent prediction uncertainty on high dimensional image data. One example of this approach is AL using Bayesian equivalent of convolutional neural networks (BCNNs [21]) for uncertainty modeling [1].

This work improves the conventional uncertainty modeling approach in kernel-based active learning. The authors achieve improvement in two aspects: First, they model spatial information in the input image captured by CNNs and uncertainty representation based on approximate variational inference using stochastic regularisation approaches such as dropout [21]. Second, they adopt BCNNs to combine recent advances in Bayesian deep learning into the active learning framework in a practical manner.

At the active learning stage using BCNNs, the authors place prior probability distributions over a set of model parameters $w = W_1..., W_L : w \sim p(w)$ (e.g., a standard Gaussian prior). They define the likelihood model $p(y = c|\boldsymbol{x}, \boldsymbol{w}) = \text{softmax}(\boldsymbol{f^w}(\boldsymbol{x}))$ for classification or a Gaussian likelihood for regression according to the prediction task. They train a model with dropout before every weight layer to approximate variational inference, i.e., minimises the Kullback-Leibler(KL) divergence $D_{KL}(p(\boldsymbol{w}|\mathcal{D}_{train})||q^*_\theta(\boldsymbol{w}))$, where $p(\boldsymbol{w}|\mathcal{D}_{train})$ is the true model posterior given a training set $\mathcal{D}_{train}$ and $q^*_\theta(\boldsymbol{w})$ is a distribution in a tractable family. Then, prediction is derived with sampling from the approximate posterior using Monte Carlo integration: $p(y = c|\boldsymbol{x}, \mathcal{D}_{train}) \approx \frac{1}{T}\sum_{t=1}^{T} p(y = c|\boldsymbol{x}, \hat{\boldsymbol{w}}_t)$ with $\hat{\boldsymbol{w}}_t \sim q^*_\theta(\boldsymbol{w})$, where $q_\theta(\boldsymbol{w})$ is the Dropout distribution [21]. They use the uncertainty information possessed by BCNNs with existing acquisition functions (querying strategy), then approximate an acquisition function using approximate distribution $q^*_\theta(\boldsymbol{w})$. In this approach, the approximated acquisition function results in a computationally tractable estimator.

The authors experiment with the number of acquisition functions and compare them with semi-supervised models, which use similar model structures, to prove the effectiveness of this BCNNs active learning approach. Table 1 shows some evaluation results where each line presents the image classification test error on MNIST given by a model. The percentage of Random (baseline) test error is 4.66%. The baseline model uses random acquisition function $a(\boldsymbol{x}) =$ unif(), sampling from a uniform distribution for label querying. They also construct models using different acquisition functions, BALD [22], Max Entropy [23], and Variation Ratios [24].

Table 1 illustlates that using BCNNs to model active learning achieved significantly better or similar performance to using semi-supervised methods. Considering that AL models have access to only the 1000 acquired images while semi-supervised ones have further access to the remaining images with no labels, the authors successfully demonstrate significant improvement on existing AL approaches.

Technique	Test error
Semi-supervised:	
Semi-sup. Embedding [25]	5.73%
Transductive SVM [25]	5.38%
MTC [26]	3.64%
DGN [27]	2.40%
Active learning with various acquisitions:	
Random (baseline)	4.66%
BALD	1.80%
Max Entropy	1.74%
Variation Ratios	1.64%

Table 1: Test error on MNIST. Active learning has access to only the 1000 acquired images. Semi-supervised further has access to the remaining images with no labels [1].

2.3 Probabilistic Circuits for Interactive Learning

Since early 2010, probabilistic circuits (PCs) have been applied to the modeling of expressive and tractable probability estimation, e.g., for density estimation [28][29]. Varying degrees of structural constraints lead to different families of probabilistic circuits, such as Sum-Product-

Networks (SPNs)[8] and Probabilistic Sentential Decision Diagrams (PSDDs)[9]. One significant advantage of PCs is their strong ability to measure uncertainty in large structured output spaces using reliable and efficient computation of conditional distributions. Considering that interactive learning is a classification task and its probabilistic modeling aims to describe the conditional distributions of continuous input features, I review one specialized model to integrate PCs into active and skeptical learning in this section.

2.3.1 Tractable Interactive Learning using Deep Probabilistic Circuits

A CRISPs (Conditional Randomized Interactive Skeptical Probabilistic circuits) is a class of tractable probabilistic models that can model incremental interactive learning in large structured output spaces using deep computational graphs [2]. InCRISPs, conditional circuits are composed of a distribution $p_{g(\boldsymbol{x})}(\boldsymbol{Y})$ implemented as a PC and a deep gating function $g(\boldsymbol{x})$ that outputs the parameters θ of the circuit, given an input $\boldsymbol{x}$. The conditional PC encodes dependencies over the labels, so the softmax/sigmoid last layer, which is generally intractable in deep neural networks, can be replaced with a learned gating function. Also, when CRISPs satisfy required structural properties [2], they guarantee the exact computational cost in time linear to the size of the encoded graph. It is beneficial the whole deep nets model can be trained end-to-end with exact computation. To learn CRISPs using GPU-accelerated libraries with a fast gradient-based optimizer and avoid learning computational graphs, the authors propose a randomized structured circuit, which leverages a randomized construction approach [30] [31].

CRISPs can be applied to model tractable uncertainty in active learning that involves interaction with human agents. In this context, the authors show the reliable computation of query label subsets by computing $\frac{1}{1-\alpha}\log(\sum_{\boldsymbol{Q}} p_\theta(\boldsymbol{Q}|\boldsymbol{x})^\alpha)$, the conditional Renyi entropy of a CRISP circuit for a subset of labels $\boldsymbol{Q} \subseteq \boldsymbol{Y}$, where all $\alpha > 0, \alpha \neq 1$, to measure uncertainty. Maximizing the conditional entropy is equivalent to deriving the maximally informative subset of labels $\boldsymbol{Q}^*$ annotated by a human agent, which is a significant factor in applications with large output spaces in the wild.

CRISPs can also be used to model suspiciousness on potentially mislabeled examples in skeptical learning. In this context, the authors show the tractable computation of the model's suspiciousness by tracking the margin between the user's annotation $\tilde{y}_t$ and the machine's prediction $\hat{\boldsymbol{y}}_t = \text{argmax}_{\boldsymbol{y}} p_\theta(\boldsymbol{y}|\boldsymbol{x}_t)$. For future work, the authors plan to evaluate CRISPs on active learning benchmarks for deep learning and structured-output prediction (SOP) tasks.

2.4 Comparison Among The Approaches

2.4.1 Acquisition Strategies and Label Querying/Selection

Table 2 summarizes the recently proposed uncertainty modeling approaches using classical and neural network techniques for AL and SKL. In uncertainty modeling for interactive learning, the acquisition process is strategy dependent. The same is true for the uncertainty modeling using (deep) neural network techniques. As shown in Table 2, a single label is typically queried in the acquisition process for classical models [1]-[4], while a subset of labels is selected for neural network models [5][7][17][18], as the architecture requires a set of input features to fit parameters jointly.

Gaussian Processes models for classical AL and SKL are adaptive-base where they update uncertainty and information density measures over samples per query [2][4]. Those typically

perform better than the SVM and RF models. On the other hand, recently proposed neural network models typically employ batch-AL, selected from a batch of unlabeled samples [17][18]. Various acquisition strategies for sample querying/selection have been used, as listed in Table 2. The review in the section 2.3 shows that probabilistic circuits are good at modeling the tractable distribution of high-dimensional output spaces with structured computational graphs. Thus, future approaches are expected to take this into account when modeling reliable probability estimation using deep learning techniques.

2.4.2 Model Structures and Active/Skeptical Learning

As shown in Table 2, different model structures have been adopted in interactive learning approaches. Various structures such as SVM, GPs, Bayesian CNNs, NNs, and DNNs were used to represent conditional PDFs and to achieve active learning [1][2][5][7][17][18]. On the other hand, RFs and GPs were used to represent the margin between the user's annotation and the machine's prediction to achieve skeptical learning [3][4]. Representation learning ability is a significant characteristic of a deep learning model. In BCNNs-based active learning [5], the experimental results in Table 1 show that the increased ability to learn image features by convolutional neural networks improves performance as it helps to measure more accurate uncertainty.

Model (structure)	**Classical**	**Query strategy**	Tractable
BvSB (SVM) [1]	AL	Margin-base	Yes
(GP) [2]	AL	Adaptive-base	Yes
SSML (RF) [3]	SKL	Confidence score	Yes
ISGP (GP) [4]	SKL	Adaptive-base	Yes
	(Deep) NN	**Subset selection strategy**	
(BCNN) [5]	AL	Approximate inference	No
BAIT (NN) [17]	AL	Fisher-based batch	Yes
CRISPs (DNN-PC) [7]	AL, SKL	Entropy reduction	Yes
BADGE (NN) [18]	AL	Gradient embeddings batch	No

Table 2: A summary of the proposed uncertainty modeling approaches using classical and (deep) neural network techniques for active and skeptical learning.

3 Summary & Conclusion

This review provides an overview of the emerging interactive learning approaches using (deep) neural network techniques. Compared with the conventional uncertainty modeling methods in image recognition based on the classical AL and SKL, deep learning models (e.g., BCNNs AL, CRISPs) are better at describing the complex relationships between a small amount of annotated data and uncertainty over unseen data, and therefore, improve the representation of model uncertainty and its performance. Various implementations of building active and skeptical learning models using classical and (deep) neural network methods for interactive image recognition in the current literature have been reviewed and compared. To facilitate a review of the area and to provide insights into the different approaches reported in the literature, I categorize them into three classes (classical, deep learning, and probabilistic circuits), describe and analyze each, and make connections systematically.

Despite the empirical successes of a range of deep learning methods in AL and SKL as reviewed in this article, there remain significant issues that need further research to exploit the intrinsic strength of deep learning techniques in the field. For example, current attempts have not achieved reliable results in prediction of labels using intractable deep Bayesian convnet models [5]. Considering the application in the sensitive area involving Human-in-the-Loop systems such as medical image recognition (diagnose lesions from images) and speech recognition, deep model structures specifically designed for reliable modeling with tractable uncertainty measurement and prediction may be necessary. Furthermore, few considerations have been made thus far in deep learning approaches to model interactive emotion recognition for a robot using AL/SKL and the subset of facial image features. I believe that a promising direction to pursue shortly is to apply the deep inference models with more reliable modeling abilities, such as Bayesian convnets with Probabilistic Circuits, as well as to apply those on emotion recognition to the AL and SKL tasks in the future.

References

[1] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *CoRR*, abs/1703.02910, 2017.

[2] Stefano Teso and Antonio Vergari. Efficient and reliable probabilistic interactive learning with structured outputs, 2022.

[3] Porikli Fatih Joshi, Ajay J and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 2372–2379. IEEE, 2009.

[4] Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 859–866. IEEE, 2013.

[5] Mattia Zeni, Wanyi Zhang, Enrico Bignotti, Andrea Passerini, and Fausto Giunchiglia. Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge. 2019.

[6] Andrea Bontempelli, Stefano Teso, Fausto Giunchiglia, and Andrea Passerini. Learning in the wild with incremental skeptical gaussian processes. In *IJCAI International Joint Conference on Artificial Intelligence*. IJCAI, 2020.

[7] Antonio Vergari. A systematic view of the literature on probabilistic circuits, computational graphs encoding tractable probability distributions.

[8] Hoifung Poon and Pedro M. Domingos. Sum-product networks: A new deep architecture. *CoRR*, abs/1202.3732, 2012.

[9] Doga Kisa, Guy Van den Broeck, Arthur Choi, and Adnan Darwiche. Probabilistic sentential decision diagrams. In *KR International Conference on Principles of Knowledge Representation and Reasoning*, KR'14, page 558–567. AAAI Press, 2014.

[10] Tahrima Rahman, Prasanna Kothalkar, and Vibhav Gogate. Cutset networks: A simple, tractable, and scalable approach for improving the accuracy of chow-liu trees. In *ECML/PKDD (2)*, volume 8725 of *Lecture Notes in Computer Science*, pages 630–645. Springer, 2014.

[11] Mark Chavira, Adnan Darwiche, and Manfred Jaeger. Compiling relational bayesian networks for exact inference. *Int. J. Approx. Reason.*, 42(1-2):4–20, 2006.

[12] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. 1996.

[13] Wikipedia. Scale-invariant feature transform. `http://en.wikipedia.org/w/index.php?title=Scale-invariant\%20feature\%20transform&oldid=1135086423`, 2023. [Online; accessed 23-January-2023].

[14] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[15] I. Daoudi and K. Idrissi. A kernel-based active learning strategy for content-based image retrieval. In *CBMI International Workshop on Content Based Multimedia Indexing*, pages 1–6. CBMI, 2010.

[16] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *IEEE International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[17] Erik Rodner Alexander Lutz and ¨Joachim Denzler. I want to know more – efficient multi-class incremental learning using gaussian processes. Pattern Recognition and Image Analysis, 2013.

[18] Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham M. Kakade. Gone fishing: Neural active learning with fisher embeddings. *CoRR*, abs/2106.09675, 2021.

[19] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *CoRR*, abs/1906.03671, 2019.

[20] Vishal Kaushal, Suraj Kothawade, Ganesh Ramakrishnan, Jeff A. Bilmes, and Rishabh K. Iyer. PRISM: A unified framework of parameterized submodular information measures for targeted data subset selection and summarization. *CoRR*, abs/2103.00128, 2021.

[21] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2015.

[22] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning, 2011.

[23] Claude Elwood Shannon. *A mathematical theory of communication*. Bell System Technical Journal, 1948.

[24] Linton G. Freeman. *Elementary applied statistics*. 1965.

[25] Jason Weston, Frédéric Ratle, and Ronan Collobert. Deep learning via semi-supervised embedding. In *ICML International Conference on Machine Learning*, page 1168–1175, New York, NY, USA, 2008. ICML.

[26] Salah Rifai, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The manifold tangent classifier. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[27] Diederik P. Kingma, Danilo Jimenez Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *CoRR*, abs/1406.5298, 2014.

[28] Robert Gens and Domingos Pedro. Learning the structure of sum-product networks. In Sanjoy Dasgupta and David McAllester, editors, *ICML International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 873–880, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[29] Yitao Liang, Jessa Bekker, and Guy Van den Broeck. Learning the structure of probabilistic sentential decision diagrams. In *UAI Conference on Uncertainty in Artificial Intelligence*. UAI, August 2017.

[30] Robert Peharz, Antonio Vergari, Karl Stelzner, Alejandro Molina, Martin Trapp, Kristian Kersting, and Zoubin Ghahramani. Probabilistic deep learning using random sum-product networks. *CoRR*, abs/1806.01910, 2018.

[31] Andy Shih and Stefano Ermon. Probabilistic circuits for variational inference in discrete graphical models. *CoRR*, abs/2010.11446, 2020.