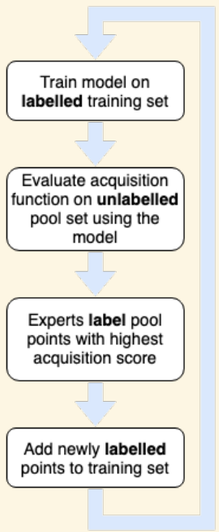
A Practical & Unified Notation for Information-Theoretic Quantities in ML

Andreas Kirsch, Yarin Gal, OATML, University of Oxford

Bayesian: Active Learning



The Bayesian model parameters Ω come with **prior distribution** $p(\omega)$, and the **training set** $\mathscr{D}^{train}$ induces a **posterior distribution** $p(\omega \mid \mathscr{D}^{train})$.

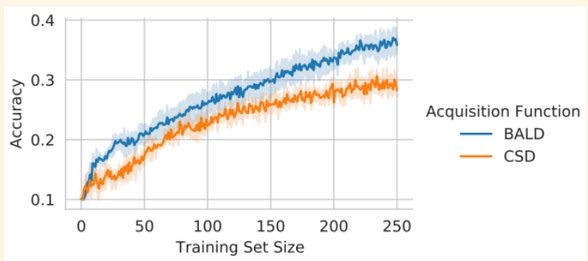
The **predictive distributio**n for a sample x is obtained by marginalizing: $p(y \mid x, \mathscr{D}^{train}) = \mathbb{E}_{p(\omega \mid \mathscr{D}^{train})}\left[p(y \mid x, \omega)\right]$.

The training set is extended with selected points $\{x_i^{acq}\}$ from the unlabeled **pool set**, which are then labeled by an **oracle**.

To decide which points to acquire, an **acquisition function** $a(\{x_i\}_i, p(\omega \mid \mathscr{D}^{train}))$ jointly scores unlabeled candidates from the pool set. The highest scoring set of samples of a predetermined **acquisition batch size** is acquired in each acquisition round.

Experiment Results

On CIFAR-10 after removing noisy labels.



Limitations & Plans

We had to remove samples with high predictive entropy (based on a LeNet-5 ensemble) to obtain good results. CSD does not seem very robust towards outliers, which makes sense given the definition of information gain: a wrong label will have higher information gain that one that was expected.

The method also failed on CIFAR-10 so far.

Regarding the notation, we want to expand it with additional examples and use-cases.

The Notation

We start with common and well-known definitions:

Definition 2.1. *Let Shannon's information content* $\mathrm{h}(\cdot)$*, cross-entropy* $\mathrm{H}(\cdot \parallel \cdot)$*, entropy* $\mathrm{H}(\cdot)$*, and KL divergence* $\mathrm{D_{KL}}(\cdot \parallel \cdot)$ *be defined for a probability distribution* p *and non-negative function* q *as:*

$$\begin{aligned}
\mathrm{h}(q) &:= -\ln q \\
\mathrm{H}(\mathrm{p}(X) \parallel \mathrm{q}(X)) &:= \mathbb{E}_{\mathrm{p}(x)}\, \mathrm{h}(\mathrm{q}(x)) \\
\mathrm{H}(\mathrm{p}(X)) &:= \mathrm{H}(\mathrm{p}(X) \parallel \mathrm{p}(X)) \\
\mathrm{D_{KL}}(\mathrm{p}(X) \parallel \mathrm{q}(X)) &:= \mathrm{H}(\mathrm{p}(X) \parallel \mathrm{q}(X)) - \mathrm{H}(\mathrm{p}(X)),
\end{aligned}$$

where we use the random variable (upper-case X*) to make clear which random variable the expectation is over.*

We can canonically extend the definitions to tie random variables to specific outcomes, e.g. X = x:

Definition 2.4. *Given random variables* X *and* Y *and outcome* y*, we define:*

$$\begin{aligned}
\mathrm{H}[y] &:= \mathrm{h}(\mathrm{p}(y)) \\
\mathrm{H}[X, y] &:= \mathbb{E}_{\mathrm{p}(x|y)}\, \mathrm{H}[x, y] = \mathbb{E}_{\mathrm{p}(x|y)}\, \mathrm{h}(\mathrm{p}(x, y)) \\
\mathrm{H}[X \mid y] &:= \mathbb{E}_{\mathrm{p}(x|y)}\, \mathrm{H}[x \mid y] = \mathbb{E}_{\mathrm{p}(x|y)}\, \mathrm{h}(\mathrm{p}(x \mid y)) \\
&= \mathrm{H}[X, y] - \mathrm{H}[y] \\
\mathrm{H}[y \mid X] &:= \mathbb{E}_{\mathrm{p}(x|y)}\, \mathrm{H}[y \mid x] = \mathbb{E}_{\mathrm{p}(x|y)}\, \mathrm{h}(\mathrm{p}(y \mid x)) \\
&\neq \mathrm{H}[X, y] - \mathrm{H}[X],
\end{aligned}$$

where we have shortened $Y = y$ *to* y*.*

Finally, we can define the mutual information using an one-sided definition and transfer it to mixed mutual information terms, which allows us to unify information gain and surprise:

Definition 2.5. *For random variables* X *and* Y *and outcomes* x *and* y *respectively, the point-wise mutual information* $\mathrm{I}[x; y]$*, the mutual information* $\mathrm{I}[X; Y]$*, the* information gain $\mathrm{I}[X; y]$*, and the* surprise $\mathrm{I}[y; X]$ *are:* @

$$\begin{aligned}
\mathrm{I}[x; y] &:= \mathrm{H}[x] - \mathrm{H}[x \mid y] = \mathrm{h}\left(\frac{\mathrm{p}(x)\,\mathrm{p}(y)}{\mathrm{p}(x, y)}\right) \\
\mathrm{I}[X; Y] &:= \mathrm{H}[X] - \mathrm{H}[X \mid Y] = \mathbb{E}_{\mathrm{p}(x,y)}\, \mathrm{I}[x; y] \\
\mathrm{I}[X; y] &:= \mathrm{H}[X] - \mathrm{H}[X \mid y] \\
\mathrm{I}[y; X] &:= \mathrm{H}[y] - \mathrm{H}[y \mid X] = \mathbb{E}_{\mathrm{p}(x|y)}\, \mathrm{I}[x; y].
\end{aligned}$$

Bayesian Core-Sets

BALD is a well-known acquisition function in active learning, which computes the expected information gain $I[\Omega; Y \mid x, \mathscr{D}^{train}]$ for the the Bayesian model parameters given the predictions for a candidate sample.

This is an approximation of the information gain if we knew the label, which is presented in more detail in the paper. Using the introduced notation, the information gain is $I[\Omega; y \mid x, \mathscr{D}^{train}]$.

We can apply this to the core-set problem which consists of selecting the most informative samples given the labels and examine how well the active learning approach works for core-sets. We call the Bayesian acquisition function **CSD (core-set by disagreement)** following BALD (Bayesian Active Learning by Disagreement).

Evaluating the Information Gain for CSD.

We show how to compute the information for the special case of an MC dropout model with dropout rate $\frac{1}{2}$. Computing the information gain for complex models is not trivial as its canonical expansion requires an explicit model. However, most available Bayesian neural networks, like MC dropout models, only provide implicit models which we can only sample from but which do not provide an easy way to compute their density. Moreover, to compute $H[\Omega \mid y, x, \mathscr{D}^{train}]$ naively we would have to perform a Bayesian inference step.

However, in this special case, we find that the information gain is directly approximated by the surprise $I[y; \Omega \mid x, \mathscr{D}^{train}]$, which is straightforward to compute.

Experiment Results

On MNIST after removing noisy labels.

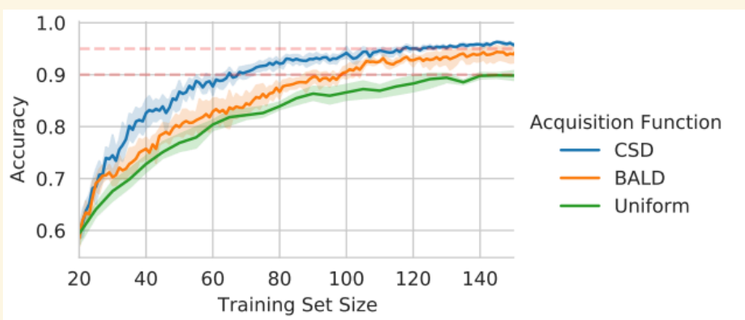


Figure 1. CSD vs BALD vs uniform acquisition on MNIST after ambiguous and mislabeled training samples have been dropped from the training set. CSD requires only 58 samples to reach 90% accuracy compared to 91 samples for BALD. 5 trials each.