Bayesian Generative Active Deep Learning

Toan Tran\textsuperscript{1} \hspace{1cm} Thanh-Toan Do\textsuperscript{2} \hspace{1cm} Ian Reid\textsuperscript{1} \hspace{1cm} Gustavo Carneiro\textsuperscript{1}

\textsuperscript{1}University of Adelaide, Australia \hspace{1cm} \textsuperscript{2}University of Liverpool

Wed Jun 12th 04:30 – 04:35 PM @ Room 201
Deep learning (DL)

- The dominant machine learning methodology [Huang et al., 2017, Rajkomar et al., 2018].
- **Issue:** significant human effort for the labeling; considerable computational resources for large-scale training process [Sun et al., 2017].

How to address these training issues? Two popular approaches:

1. (Pool-based) active learning (AL): Challenging to be applied in DL: AL may overfit the (small) informative training sets
2. Data augmentation (DA): the generation of new samples is done without regarding their informativeness $\implies$ the training process takes longer than necessary and relatively ineffective
Main goals of this paper

- Propose a novel Bayesian generative active deep learning method

- Targets the augmentation of the labeled data set with informative generated samples

- **Key technical contribution:** theoretically and empirically show the informativeness of this generated sample.

**Figure 1:** Our proposed method depicted by VAE-ACGAN model
Generative adversarial active learning (GAAL) [Zhu and Bento, 2017]

- Relies on an optimization problem to generate new informative samples
- Can generate rich representative training data with the assumptions:
  - GAN model has been pre-trained, and
  - The optimization during generation is solved efficiently

Comparison between our proposed method and GAAL [Zhu and Bento, 2017]

<table>
<thead>
<tr>
<th></th>
<th>GAAL [Zhu and Bento, 2017]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>acquisition function</td>
<td>simple (binary classifier)</td>
<td>more effective (deep models)</td>
</tr>
<tr>
<td>training of the generator ($G$) and classifier ($C$)</td>
<td>2-stage</td>
<td>$G$ and $C$ are jointly trained</td>
</tr>
<tr>
<td></td>
<td>GAN model is pre-trained</td>
<td>Allowing them to “co-evolve”</td>
</tr>
<tr>
<td>classification results</td>
<td>not competitive enough</td>
<td></td>
</tr>
</tbody>
</table>
Main technical contribution:
Combining BALD and BDA for generating new labeled samples that are informative for the training process.

Initial labeled data: \( \mathcal{D} = \{ (x_i, y_i) \}_{i=1}^N \), where \( x_i \in X \subseteq \mathbb{R}^d \) is the data sample labeled with \( y_i \in C = \{1, 2, \ldots, C\} \) (\( C = \# \text{ classes} \)).

Bayesian active learning by disagreement (BALD) scheme [Gal et al., 2017, Houlsby et al., 2011]:

The most informative sample \( x^* \) is selected from the (unlabeled) pool data \( \mathcal{D}_{\text{pool}} \) by [Houlsby et al., 2011]:

\[
x^* = \arg \max_{x \in \mathcal{D}_{\text{pool}}} a(x, \mathcal{M}),
\]

where the acquisition function \( a(x, \mathcal{M}) \) is estimated by the Monte Carlo (MC) dropout method [Gal et al., 2017]

\[
a(x, \mathcal{M}) \approx -\sum_c \left( \frac{1}{T} \sum_t \hat{p}_c \right) \log \left( \frac{1}{T} \sum_t \hat{p}_c \right) + \frac{1}{T} \sum_{c,t} \hat{p}_c \log \hat{p}_c,
\]

where \( T \) is the number of dropout iterations, \( \hat{p}_t = [\hat{p}_{t1}, \ldots, \hat{p}_{tC}] = \text{softmax}(f(x; \theta^t)) \), with \( f \) is the network function parameterized by \( \theta^t \sim p(\theta|\mathcal{D}) \) at the \( t \)-th iteration.
The generated sample $\mathbf{x}'$:

$$\mathbf{x}' = g(e(\mathbf{x}^*)),$$  \hspace{1cm} (3)

where a variational autoencoder (VAE) [Kingma and Welling, 2013] contains an encoder $e(.)$ and a decoder $g(.)$.

VAE training: minimizing the “reconstruction loss”, where if # training iterations is sufficiently large, we have:

$$\|\mathbf{x}' - \mathbf{x}^*\| < \varepsilon,$$  \hspace{1cm} (4)

$\varepsilon > 0$ (arbitrarily small) – see Fig. 3.

$\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{x}^*, \mathbf{y}^*), (\mathbf{x}', \mathbf{y}^*)\}$, which are used for the next training iteration.

Figure 3: Reduction of $\|\mathbf{x}' - \mathbf{x}^*\|$ as the training progresses of the VAE model.
Key Question

How about the "information content" of the generated sample $x'$ measured by $a(x', M)$?

Proposition 2.1

Assuming that there exists the gradient of the acquisition function $a(x, M)$ with respect to the variable $x$, namely $\nabla_x a(x, M)$, and that $x^*$ is an interior point of $D_{\text{pool}}$, then $a(x', M) \approx a(x^*, M)$ (i.e., the absolute difference between these values are within a certain range). Consequently, the sample $x'$ generated from the most informative sample $x^*$ by (3) is also informative.
Figure 4: Network architecture of our proposed model.
Classification performance measured by the top-1 accuracy as a function of the number of acquisition iterations and the percentage of training samples.

Our proposed algorithm, active learning using “information-preserving” data augmentation (AL w. VAEACGAN) is compared with:

- Active learning using BDA (AL w. ACGAN)
- BALD [Gal et al., 2017] without using data augmentation (AL without DA),
- BDA [Tran et al., 2017] without active learning (BDA) (using full and partial training sets)
- Random selection

Benchmark data sets: MNIST [LeCun et al., 1998], CIFAR-10, CIFAR-100 [Krizhevsky et al., 2012], and SVHN [Netzer et al., 2011].

Baseline classifiers: ResNet18 [He et al., 2016a] and ResNet18pa [He et al., 2016b]
Figure 5: Training and classification performance of the proposed Bayesian generative active learning (AL w. VAEACGAN) compared to other methods. This performance is measured as a function of the number of acquisition iterations and respective percentage of samples from the original training set used for modeling.
Table I: Mean ± standard deviation of the classification accuracy on MNIST, CIFAR-10, and CIFAR-100 after 150 iterations over 3 runs

<table>
<thead>
<tr>
<th></th>
<th>AL w. VAEACGAN</th>
<th>AL w. ACGAN</th>
<th>AL w. PMDA</th>
<th>AL without DA</th>
<th>BDA (partial training)</th>
<th>Random selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST Resnet18</td>
<td>99.53 ± 0.05</td>
<td>99.45 ± 0.02</td>
<td>99.37 ± 0.15</td>
<td>99.33 ± 0.10</td>
<td>99.33 ± 0.04</td>
<td>99.00 ± 0.13</td>
</tr>
<tr>
<td>Resnet18PA</td>
<td>99.68 ± 0.08</td>
<td>99.57 ± 0.07</td>
<td>99.49 ± 0.09</td>
<td>99.35 ± 0.11</td>
<td>99.35 ± 0.07</td>
<td>99.20 ± 0.12</td>
</tr>
<tr>
<td>CIFAR-10 Resnet18</td>
<td>87.63 ± 0.11</td>
<td>86.80 ± 0.45</td>
<td>82.17 ± 0.35</td>
<td>79.72 ± 0.19</td>
<td>85.08 ± 0.31</td>
<td>77.29 ± 0.23</td>
</tr>
<tr>
<td>Resnet18PA</td>
<td>91.13 ± 0.10</td>
<td>90.70 ± 0.24</td>
<td>87.70 ± 0.39</td>
<td>85.51 ± 0.21</td>
<td>86.90 ± 0.27</td>
<td>80.69 ± 0.19</td>
</tr>
<tr>
<td>CIFAR-100 Resnet18</td>
<td>68.05 ± 0.17</td>
<td>66.50 ± 0.63</td>
<td>55.24 ± 0.57</td>
<td>50.57 ± 0.20</td>
<td>65.76 ± 0.40</td>
<td>49.67 ± 0.52</td>
</tr>
<tr>
<td>Resnet18PA</td>
<td>69.69 ± 0.13</td>
<td>67.79 ± 0.76</td>
<td>59.67 ± 0.60</td>
<td>55.82 ± 0.31</td>
<td>65.79 ± 0.51</td>
<td>54.77 ± 0.29</td>
</tr>
</tbody>
</table>

Figure 6: Images generated by our proposed (AL w. VAEACGAN) approach for each data set.

Toan Tran (University of Adelaide)
Consistently shows that our proposed approach outperforms other methods.

The classification results by AL w. VAEACGAN are statistically significant w.r.t. BDA (partial training) on all those data sets, and w.r.t. AL w. ACGAN on CIFAR-\{10, 100\} for both models (i.e., \( p \leq .05 \), two-sample t-test).

Proposed a novel Bayesian generative active deep learning approach that consistently shows to be more effective than data augmentation and active learning in several classification problems.

Our proposed approach is (active learning) model-agnostic, it therefore can be combined with several currently state-of-the-art active learning methods [Ducoffe and Precioso, 2018, Gissin and Shalev-Shwartz, 2018, Sener and Savarese, 2018].

Future work

- To investigate how to generate samples directly using complex acquisition functions
- To work on the efficiency of our proposed method (its empirical computational cost is slightly higher than BDA [Tran et al., 2017] and BALD [Gal et al., 2017, Houlsby et al., 2011]).
- Imbalanced data sets
Poster presentation:
Wed Jun 12th 06:30–09:00 PM
@ Pacific Ballroom #263


Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. 2018.


