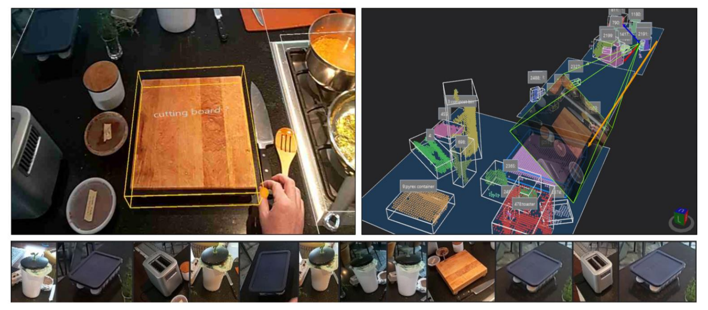
Streaming Active Learning with Deep Neural Networks

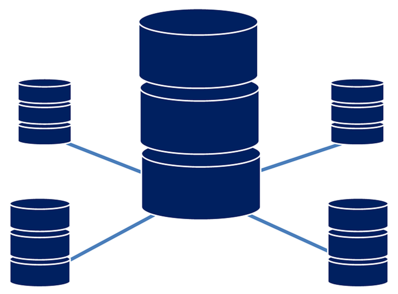
Akanksha Saran, Safoora Yousefi, Akshay Krishnamurthy, John Langford, Jordan T. Ash

Real-World Applications with Streaming Data Settings

In several real-world applications, data arrive in a stream and the total number of samples are unknown ahead of time.



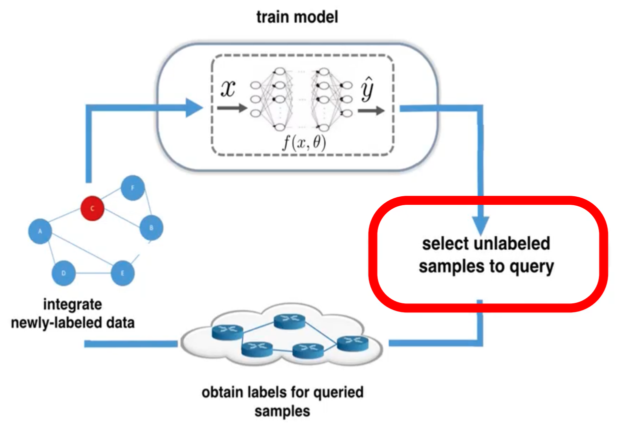
- Interaction-centric AR/VR applications such as continual object/activity learning in the wild
- Fixed datasets that are large, fractured and interacted via streaming, distributed data frameworks



How can we train deep neural networks in a data efficient manner for streaming applications?

Batch Active Learning for Deep Neural Networks

- Batch active learning or pool-based active learning for deep neural networks identifies a batch of k samples from an unlabeled data pool to be integrated into the training set.
- Popular approaches for batch active learning rely on samplers that require all unlabeled data to be simultaneously available.

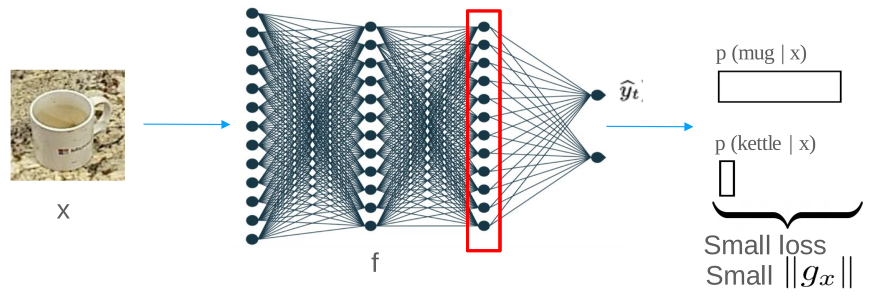


State-of-the-art non-streaming batch active learning method BADGE [1] trades off between the model's **uncertainty** about data labels and **diversity** of samples in the batch.

Representation: Hypothetical Gradient Embeddings

$$\hat{y}_t = \arg\max f(x_t;\theta)$$

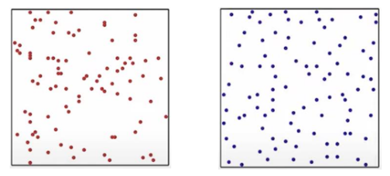
$$g(x_t) = \frac{\partial}{\partial \theta_L}\ell(f(x_t;\theta),\hat{y}_t)$$



Sampling: Volume Sampling

$$p_B \propto \det\left(\sum_{x\in B} g(x)g(x)^\top\right)$$

The determinant for volume sampling is large for a batch of high magnitude, linearly independent samples, encouraging diversity in the batch.



Streaming Batch Active Learning for Deep Neural Networks

For streaming batch active learning, it is desirable to approximate volume sampling with the following properties:

Committal: Select samples for querying as soon as they arrive in the stream

Equitable sampling: Distribute labeling queries evenly across the data stream to match a maximum query rate q

Approximate Volume Sampling

Sampling a batch of points

$$p_B \propto \det\left(\sum_{x\in B} g(x)g(x)^\top\right)$$

Sampling a single point

$$p_t \propto \det\left(\widehat{\Sigma}_t + \underbrace{g(x_t)g(x_t)^\top}\right)$$

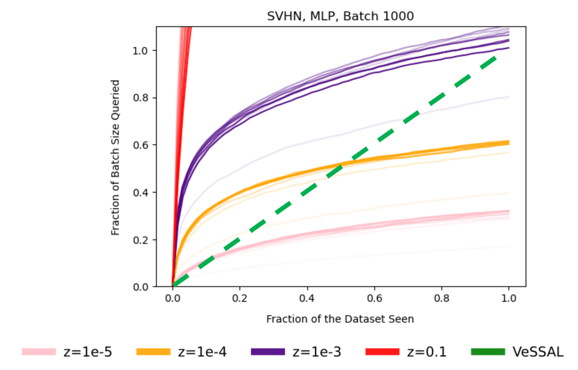
Covariance of points already chosen in the batch

Determinantal contribution of the current sample

$$\mathbb{E}_x[p_t] = \mathbb{E}_x\left[\boxed{z_t}\cdot \underbrace{g(x_t)^\top\widehat{\Sigma}_t^{-1}g(x_t)}_{\text{Elliptical Potential}}\right] = \boxed{q}$$

Scaling term

Query Rate



VeSSAL: VolumE Sampling for Streaming Active Learning

$$\begin{aligned}\mathbb{E}_x\left[z_t\cdot g(x)^\top\hat{\Sigma}_t^{-1}g(x)\right] &= z_t\cdot\mathbb{E}_x\left[\operatorname{tr}\left(g(x)^\top\hat{\Sigma}_t^{-1}g(x)\right)\right]\\ &= z_t\cdot\mathbb{E}_x\left[\operatorname{tr}\left(\hat{\Sigma}_t^{-1}g(x)g(x)^\top\right)\right]\\ &= z_t\cdot\operatorname{tr}\left(\hat{\Sigma}_t^{-1}\mathbb{E}_x\left[g(x)g(x)^\top\right]\right).\end{aligned}$$

VeSSAL (algebraically) autotunes the scaling term z_t by disentangling the gradient statistics $\mathbb{E}_x\left[g(x)g(x)^\top\right]$ from the constantly evolving $\hat{\Sigma}_t^{-1}$.

$$\mathbb{E}_x[p_t] = \mathbb{E}_x\left[z_t\cdot g(x_t)^\top\widehat{\Sigma}_t^{-1}g(x_t)\right] = q \quad [1]$$

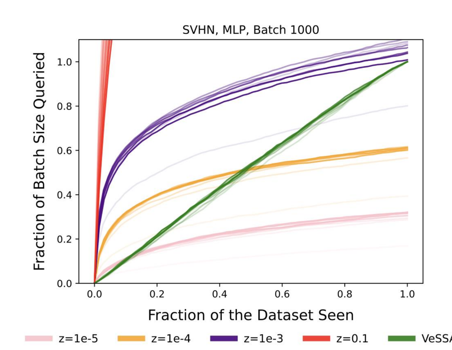
$$= z_t \operatorname{tr}\left(\widehat{\Sigma}_t^{-1}\mathbb{E}_x\left[\underbrace{g(x_t)g(x_t)^\top}\right]\right) \quad [2]$$

Inverse covariance of points already chosen in the batch

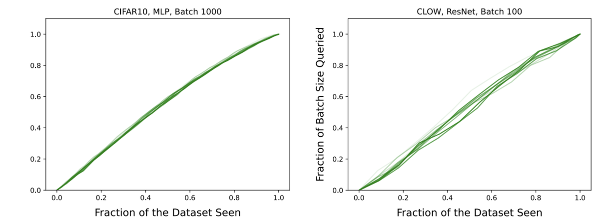
Covariance of all samples seen so far

$$\Rightarrow \quad p_t = \frac{q\cdot g(x_t)^\top\hat{\Sigma}_t^{-1}g(x_t)}{\operatorname{tr}\left(\frac{1}{t}\hat{\Sigma}_t^{-1}\sum_{i=1}^{t}g(x_i)g(x_i)^\top\right)}$$

Sampling Probability



$$q_t = \frac{k-|B_t|}{n-t}$$



Algorithm 1 Volume sampling for streaming active learning (VeSSAL)

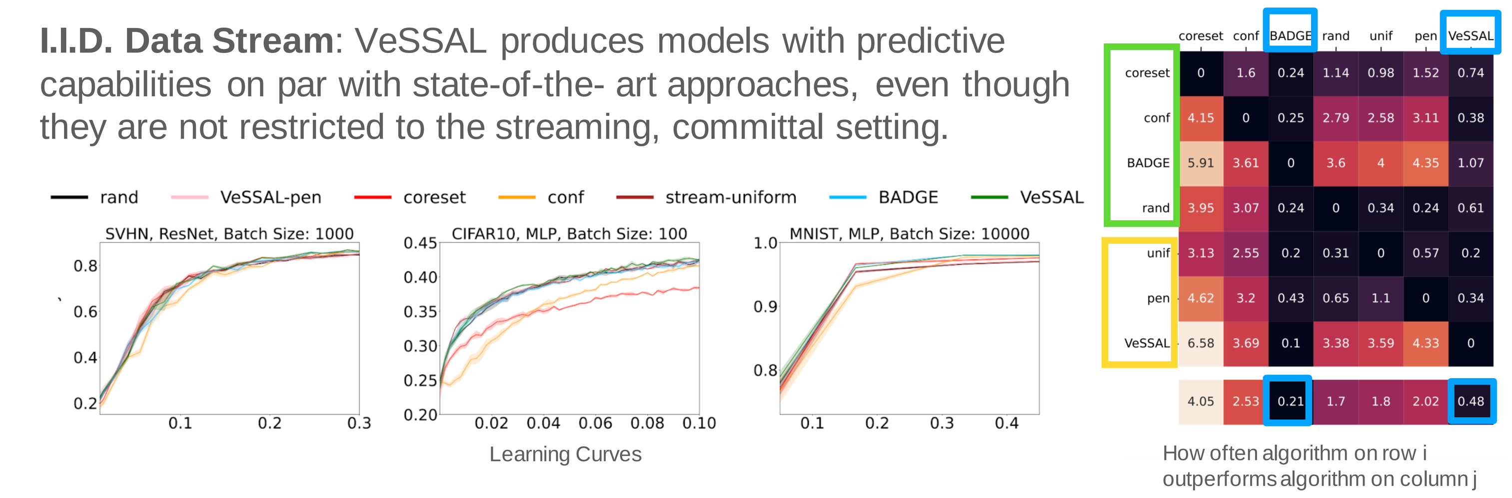
Require: Neural network $f(x;\theta)$, unlabeled stream of samples U, ideal sampling rate q

Initialize $t = 1$
Initialize $\hat{\Sigma}_0^{-1} = \lambda^{-1}I_d$ {regularized by λ for stability}
Initialize $A_0 = 0_{d,d}$ {covariance over all data}
Initialize $B = \emptyset$ {set of chosen samples}
for $x_t \in U$: **do**
  $A_t \leftarrow \frac{t-1}{t}A_{t-1} + \frac{1}{t}g(x_t)g(x_t)^\top$
  $p_t = q\cdot g(x_t)^\top\hat{\Sigma}_t^{-1}g(x_t)\operatorname{tr}(\hat{\Sigma}_t^{-1}A_t)^{-1}$
  with probability $\min(p_t, 1)$:
    Query label y_t for sample x_t
    $B \leftarrow B \cup (x_t, y_t)$
    $\hat{\Sigma}_{t+1}^{-1} \leftarrow \hat{\Sigma}_t^{-1} - \frac{\hat{\Sigma}_t^{-1}g(x_t)g(x_t)^\top\hat{\Sigma}_t^{-1}}{1+g(x_t)^\top\hat{\Sigma}_t^{-1}g(x_t)}$ {rank-1 Woodbury update}
  else:
    $\hat{\Sigma}_{t+1}^{-1} \leftarrow \hat{\Sigma}_t^{-1}$
  $t \leftarrow t + 1$
  return labeled batch B for retraining f
end for

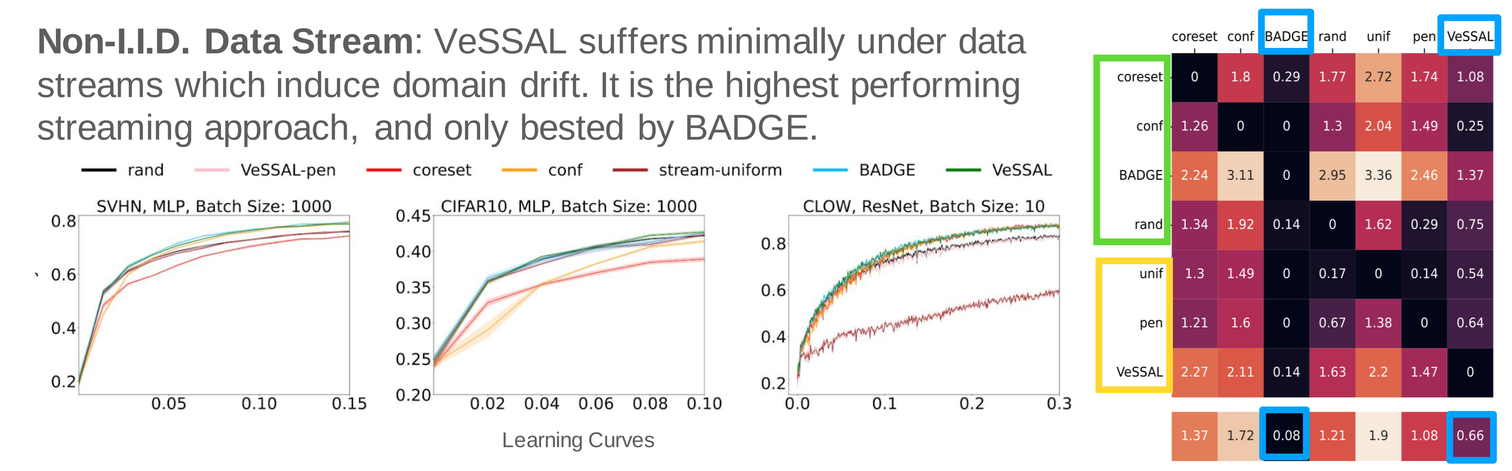
Results

We conduct experiments with 4 datasets x 3 batch sizes x 3 neural network architectures x 7 active learning algorithms (streaming and non-streaming).

I.I.D. Data Stream: VeSSAL produces models with predictive capabilities on par with state-of-the- art approaches, even though they are not restricted to the streaming, committal setting.

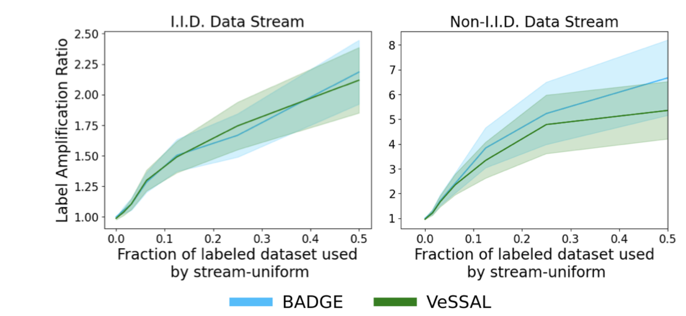


Non-I.I.D. Data Stream: VeSSAL suffers minimally under data streams which induce domain drift. It is the highest performing streaming approach, and only bested by BADGE.

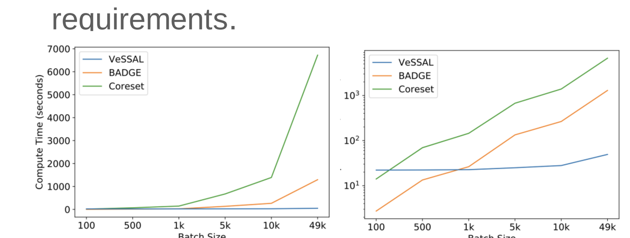


Predictive Power: VeSSAL delivers more predictive power (up to 5x) for the same labelling budget compared to uniform sampling in streaming settings.

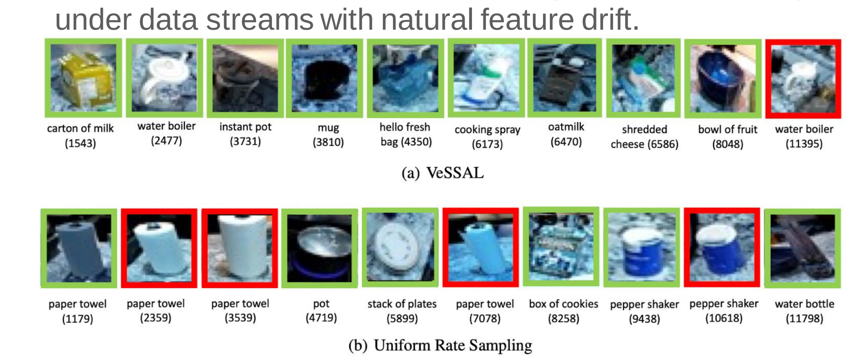
This is evaluated using the **Label Amplification Ratio** which is the number of samples used by a uniform sampling approach divided by the number of samples required by an active sampling approach to reach the same performance.



Compute Requirements: VeSSAL enjoys fixed run time with increasing batch sizes, while other non-streaming approaches have super-linear compute requirements.



Qualitative Results: VeSSAL samples diverse images under data streams with natural feature drift.



VeSSAL is a high-performing, hyperparameter free, computationally efficient, committal acquisition function that trades off between diversity & uncertainty from a stream of samples to match a desired query rate.

References

[1] Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, un- certain gradient lower bounds. *International Conference on Learning Representations*, 2020.

[2] Ash, J., Goel, S., Krishnamurthy, A., and Kakade, S. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems*, 34: 8927–8939, 2021.

[3] MacKay, D. J. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[4] Settles, B. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[5] Bohus, D., Andrist, S., Feniello, A., Saw, N., and Horvitz, E. Continual learning about objects in the wild: An interactive approach. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 476–486, 2022.