PAC-Bayesian Contrastive Unsupervised Representation Learning

Kento Nozawa

Pascal Germain

Benjamin Guedj

東京大学 THE UNIVERSITY OF TOKYO

RIKEN

UNIVERSITÉ LAVAL

Inria

UCL

Paper: http://auai.org/uai2020/proceedings/24_main_paper.pdf

Code: https://github.com/nzw0301/pb-contrastive

Contrastive unsupervised representation learning (CURL)

Goal: learn a good feature extractor $\mathbf{f}$, e.g. DNNs.

$\mathbf{x}$

similar $\mathbf{x}^+$

dissimilar $\mathbf{x}^-$

Contrastive unsupervised representation learning (CURL)

Goal: learn a good feature extractor $\mathbf{f}$, e.g. DNNs.

$\mathbf{x}$  $\rightarrow$ $\mathbf{f}$

similar $\mathbf{x}^+$  $\rightarrow$ $\mathbf{f}$

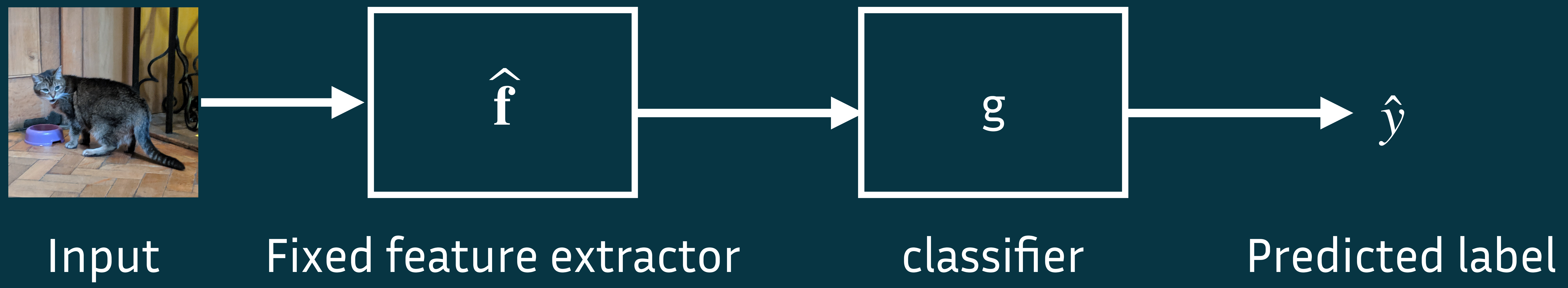
dissimilar $\mathbf{x}^-$  $\rightarrow$ $\mathbf{f}$

Contrastive Loss:
$\ell[\mathbf{f}(\mathbf{x}) \cdot (\mathbf{f}(\mathbf{x}^+) - \mathbf{f}(\mathbf{x}^-))]$

$\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}^+)$ are similar;

$\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}^-)$ are dissimilar.

Learnt representation works for supervised tasks



Why does CURL perform well?

The first theoretical guarantees for CURL (Arora et al. 2019)

Informal bound: $L_{\text{sup}}(\widehat{\mathbf{f}}) \le \alpha L_{\text{un}}(\mathbf{f}) + \underbrace{\mathcal{O}\left(\mathcal{R}(\mathcal{F}), \delta\right)}_{\text{Complexity}} \qquad \forall \mathbf{f} \in \mathcal{F}$

- α: Constant
- $\mathcal{R}$: Rademacher complexity of function class $\mathcal{F}$.
- δ: Confidence of PAC learning

Finding a good representation $\widehat{\mathbf{f}}$ guarantees to generalise well.

S. Arora et al. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *ICML*, 2019.

Our contributions

- We show PAC-Bayes bounds for CURL and derive new algorithms by minimising the bounds.
  - We replace the Rademacher complexity term from Arora et al. (2019) with a Kullback–Leibler divergence term, which is easier to compute in general.
  - The PAC-Bayes bound directly suggests a (theory driven) learning algorithm.

- We also show a PAC-Bayes bound for non-iid contrastive data.
  - The iid assumption seems unrealistic in many settings and is unlikely to hold with contrastive datasets.

General PAC-Bayes

- Q: Posterior. Probability distribution over a function class $\mathscr{F}$. It ***can*** depend on training data.
- P: Prior. Probability distribution over a function class $\mathscr{F}$. It ***cannot*** depend on training data.
- $R(Q) = \mathbb{E}_{(\mathbf{x},y)\sim\mathscr{D}}\mathbb{E}_{f\sim Q}\ell(y, f(\mathbf{x}))$: Expected risk of Q on test data
- $\widehat{R}(Q) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{f\sim Q}\ell(y_i, f(\mathbf{x_i}))$: Expected risk of Q on train data

Informal bound:

$$R(Q) \leq \alpha\widehat{R}(Q) + \underbrace{\mathcal{O}\left(\mathrm{KL}(Q\|P), \delta\right)}_{\text{Complexity}} \qquad \forall Q \text{ over } \mathscr{F}, \text{ w.h.p. } 1-\delta$$

The first PAC-Bayesian generalisation bound for CURL

Informal bound:

$$L_{\mathsf{sup}}(Q) \le \alpha \widehat{L}_{\mathsf{un}}(Q) + \underbrace{O\left(\mathrm{KL}(Q \| P), \delta\right)}_{\text{Complexity}} \qquad \forall Q \text{ over } \mathscr{F}, \text{ w.h.p. } 1 - \delta$$

- The complexity term is easier to compute than Rademacher one.
- Since all terms in the right-hand side are explicit or easy to approximate, we can minimise the bound directly.

Learning algorithms & Experiments

- Minimising $\widehat{L}_{\text{un}}(Q) + \mathcal{O}\left(\mathrm{KL}(Q \| P), \delta\right)$ w.r.t. Q.
  - P and Q are multivariate Gaussians with diagonal covariance.
  - We optimise Q's mean and covariance by using SGD (Dziugaite and Roy. 2017).
    - Approximate $\widehat{L}_{\text{un}}$ by sampling weights of neural networks from Q.

- Evaluation procedures:
  - **Learning**: Q on contrastive unsupervised data.
  - **Evaluation**: test 0-1 risk on supervised data by using centroid classifier.

Dziugaite and Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *UAI*, 2017.

Experimental results: Supervised performance & bound

	supervised		Arora et al. (2019)		PAC-Bayes based methods: parameter selection by validation set		PAC-Bayes based methods: parameter selection by PAC-Bayes bound	
	μ	μ-5	μ	μ-5	μ	μ-5	μ	μ-5
CIFAR-100								
AVG-2 risk ↓	0.086	0.125	0.106	0.144	**0.100**	**0.128**	0.246	0.292
TOP-5 risk ↓	0.422	0.540	0.471	0.574	**0.460**	**0.548**	0.766	0.806
Contrastive test risk $R_{\mathrm{un}}(Q)$ ↓	–		–		**0.197**		0.327	
PAC-Bayes upper bound ↓	–		–		0.718		**0.437**	
AUSLAN								
AVG-2 risk ↓	0.198	0.249	**0.144**	**0.167**	0.147	0.171	0.174	0.209
TOP-5 risk ↓	0.643	0.759	**0.433**	0.518	0.435	**0.509**	0.494	0.616
Contrastive test risk $R_{\mathrm{un}}(Q)$ ↓	–		–		**0.185**		0.220	
PAC-Bayes upper bound ↓	–		–		0.417		**0.361**	

- **AVG-2 risk**: averaged 0-1 risk over all combination of two classes in supervised data.
- **PAC-Bayes bound**: computed on the stochastic neural networks.

Conclusion

- We provide the first PAC-Bayes generalisation bounds for CURL.
  - This allows to derive new algorithms by directly optimising the bound.

- More results in the paper:
  - General PAC-Bayes bound for multiple dissimilar samples.
  - Bounds and learning algorithm for the non-iid case.

Paper: http://auai.org/uai2020/proceedings/24_main_paper.pdf

Code: https://github.com/nzw0301/pb-contrastive