Continual Learning through One-Class Classification using VAE

ICASSP 2020 paper by Felix Wiewel, Andreas Brendle and Bin Yang

Felix Wiewel

Institute of Signal Processing and System Theory

University of Stuttgart

03.30.2020

Table of Contents

Table of Contents

Continual Learning

- Goal
  - Train a DNN on sequence of tasks
- Restriction
  - Only data of most recent task available
- Challenges
  - Catastrophic forgetting [Fre99]
  - Knowledge transfer
  - Model size
- Methods
  - Regularization [ZPG17, KPR+17]
  - Structural [YYLH18]
  - Rehearsal [SLKK17, LP+17, vdVT18]
  - Bayesian [NLBT18]

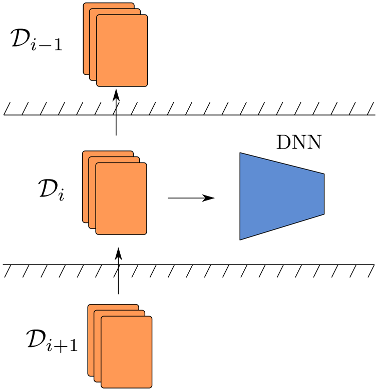

Figure: Training on a sequence of tasks

One-Class Classification

- One-Class Classifier [MH96]
  - Distinguish target class vs others
  - *Within-class* generalization
  - *Between-class* generalization
  - *Out-of-class* generalization
- Approaches
  - Density based
  - Boundary methods
  - Reconstruction based [AC15, XCZ+18, KKH18, SCKC16]

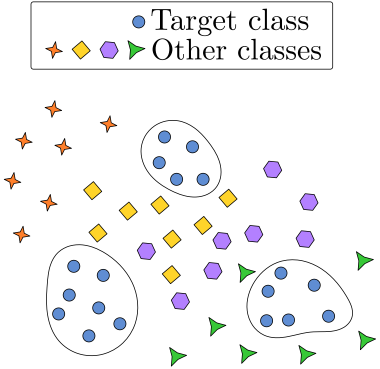

Figure: One-Class Classification example

Table of Contents

Proposed Method

- Basic idea
  - Interpret CL as series of OCC problems
  - Use generative model to build memory
  - Share latent space
- Realization
  - VAE for OCC of every class
  - Shared encoder
  - Generative replay using learned decoders

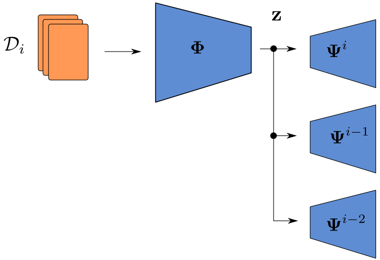

Figure: Proposed method structure

Proposed Method

- General formulation using
  - Shared encoder $\boldsymbol{\Phi}$
  - Class specific decoders $\boldsymbol{\Psi}^i$
  - Supervised loss $\mathcal{L}$
  - Regularization $\mathcal{R}$ against catastrophic forgetting

$$\max_{\boldsymbol{\Phi}, \boldsymbol{\Psi}^1, \dots \boldsymbol{\Psi}^M} \mathcal{L}(\boldsymbol{\Phi}, \boldsymbol{\Psi}^1, \dots \boldsymbol{\Psi}^M) + \mathcal{R}(\boldsymbol{\Phi}) \tag{1}$$

Proposed Method

- Formulation using VAE
  - Shared encoder $\boldsymbol{\Phi}$
  - Class specific decoders $\boldsymbol{\Psi}^i$
  - Evidence lower bound (ELBO), i.e. $\mathbb{E}_{p_{\mathbf{x}}}\left[\mathbb{E}_{q_{\boldsymbol{\Phi}}}\left[\ln p_{\boldsymbol{\Psi}}(\mathbf{x}|\mathbf{z})\right] - \mathrm{D}(q_{\boldsymbol{\Phi}} \parallel p_{\mathbf{z}})\right]$

$$\max_{\boldsymbol{\Phi},\boldsymbol{\Psi}^1,\ldots\boldsymbol{\Psi}^M} \sum_{m=1}^{M} \mathbb{E}_{p_{\mathbf{x}}^m}\left[\mathbb{E}_{q_{\boldsymbol{\Phi}}}\left[\ln p_{\boldsymbol{\Psi}^m}(\mathbf{x}|\mathbf{z})\right] - \mathrm{D}(q_{\boldsymbol{\Phi}} \parallel p_{\mathbf{z}}^m)\right] + \sum_{n=1}^{N} \mathbb{E}_{p_{\mathbf{x}}^n}\left[\mathbb{E}_{q_{\boldsymbol{\Phi}}}\left[\ln p_{\boldsymbol{\Psi}_s^n}(\mathbf{x}|\mathbf{z})\right] - \mathrm{D}(q_{\boldsymbol{\Phi}} \parallel p_{\mathbf{z}}^n)\right] \tag{2}$$

Proposed Method

- What prior distribution $p_{\mathbf{z}}$ to use?
  - Commonly a Gaussian is used for VAE
  - Each class is assigned one prior $p_{\mathbf{z}}^m$
  - Means $\boldsymbol{\mu}_m$ are chosen such that $\|\boldsymbol{\mu}_m - \boldsymbol{\mu}_n\|_2 = c \; \forall \; m \neq n$

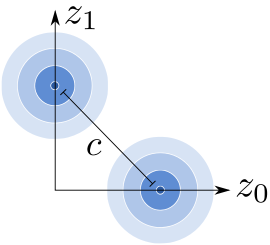

Figure: Prior placement for individual classes

Table of Contents

Data sets

- SplitMNIST
  - Based on MNIST: $\mathbf{x} \in \mathbb{R}^{28\times 28}$, $y \in \{0, \ldots, 9\}$
  - 60000 training and 10000 test examples
  - Split into ten tasks containing only one class

Figure: MNIST examples

- SplitFashionMNIST
  - Based on Fashion MNIST: $\mathbf{x} \in \mathbb{R}^{28\times 28}$, $y \in \{0, \ldots, 9\}$
  - 60000 training and 10000 test examples
  - Split into ten tasks containing only one class

Figure: Fashion MNIST examples

Architecture & Training

- Shared encoder
  - Densely connected
  - $400, 300, 200, 100, 10/10$ Neurons
  - ReLU, linear and softplus activation
- Class specific decoders
  - One densely connected layer
  - 49 ReLU activated neurons
  - Convolutional layers
  - $16, 32, 1$ filters with 3×3 kernel
  - ReLU and Sigmoid activation
  - Bilinear upsampling after conv. layer
- Optimizer
  - RMSprop
- Learning rate
  - 0.001
- Batch size
  - 128
- Reporting
  - Avg. and Std. over ten runs
- Threshold estimation
  - On training data only
  - $\gamma^n = \mu^n_{ELBO} - 6\sigma^n_{ELBO}$

Results

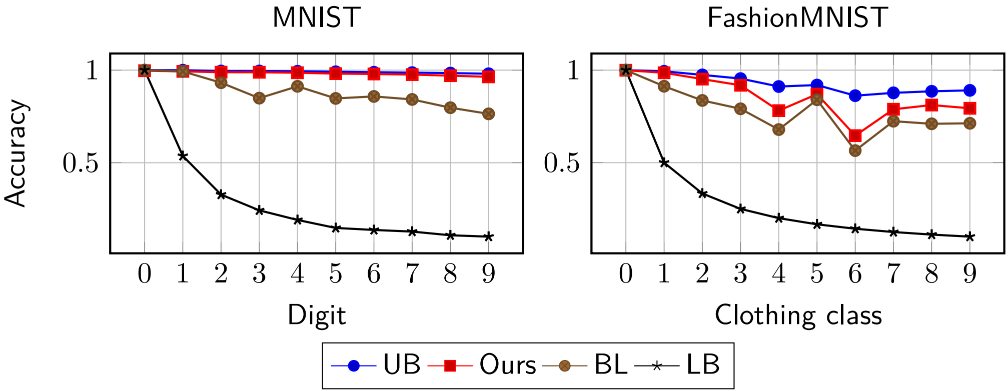

Figure: Comparison of proposed method for incrementally learning all classes in MNIST and FashionMNIST data sets with upper bound (UB), lower bound (LB) and base line (BL).

Results

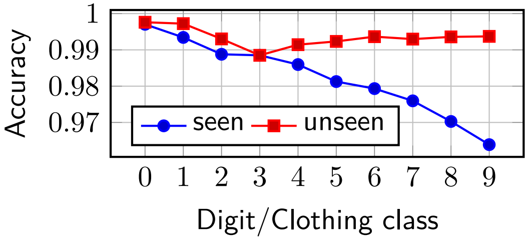

Figure: Classification accuracy on seen (splitMNIST) and detection of unseen (FashionMNIST) classes during ICL.

Table: Accuracy on data sets after learning all classes.

Method	splitMNIST [%]	splitFashionMNIST [%]
SI	19.67 ± 0.29[HLK18]	-
EWC	19.80 ± 0.05[HLK18]	15.96 ± 4.86
DGR	91.24 ± 0.33[HLK18]	72.84 ± 3.03
RtF	92.56 ± 0.21[HLK18]	75.21 ± 2.42
Ours	$\mathbf{96.39 \pm 0.23}$	$\mathbf{79.38 \pm 0.69}$

Table of Contents

Conclusion

- New method for CL based on OCC
- Multi-class classification as series of OCC problems
- CL is enabled through generative replay
- Results competitive on common benchmarks
- Our method detects unknown classes

Thank you for your attention!

References

Jinwon An and Sungzoon Cho.
Variational autoencoder based anomaly detection using reconstruction probability.
Special Lecture on IE, 2:1–18, 2015.

Robert M French.
Catastrophic forgetting in connectionist networks.
Trends in cognitive sciences, 3(4):128–135, 1999.

Yen-Chang Hsu, Yen-Cheng Liu, and Zsolt Kira.
Re-evaluating continual learning scenarios: A categorization and case for strong baselines.
In *Continual Learning Workshop NeurIPS*, 2018.

Yuta Kawachi, Yuma Koizumi, and Noboru Harada.
Complementary set variational autoencoder for supervised anomaly detection.
In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2366–2370. IEEE, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al.
Overcoming catastrophic forgetting in neural networks.
Proceedings of the national academy of sciences, page 201611835, 2017.

David Lopez-Paz et al.
Gradient episodic memory for continual learning.
In *Advances in Neural Information Processing Systems*, pages 6467–6476, 2017.

References

Mary M Moya and Don R Hush.
Network constraints and multi-objective optimization for one-class classification.
Neural Networks, 9(3):463–474, 1996.

Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner.
Variational continual learning.
In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Suwon Suh, Daniel H Chae, Hyon-Goo Kang, and Seungjin Choi.
Echo-state conditional variational autoencoder for anomaly detection.
In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 1015–1022. IEEE, 2016.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim.
Continual learning with deep generative replay.
In *Advances in Neural Information Processing Systems*, pages 2990–2999, 2017.

Michiel van der Ven and Andreas S. Tolias.
Generative replay with feedback connections as a general strategy for continual learning.
CoRR, abs/1809.10635, 2018.

Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, et al.
Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications.
In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 187–196. International World Wide Web Conferences Steering Committee, 2018.

References

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang.
Lifelong learning with dynamically expandable networks.
In *International Conference on Learning Representations*, 2018.

Friedemann Zenke, Ben Poole, and Surya Ganguli.
Continual learning through synaptic intelligence.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3987–3995. JMLR. org, 2017.