UNIT-LEVEL SURPRISE IN NEURAL NETWORKS

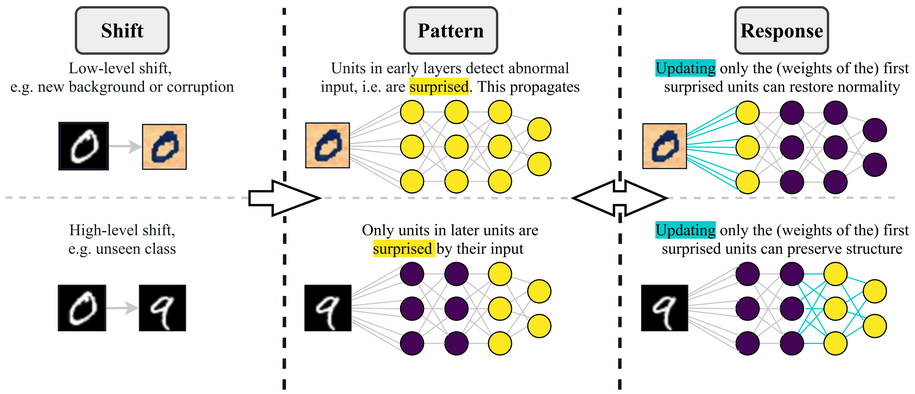
Cian Eastwood*, **Ian Mason***, **Christopher K. I. Williams**

THE UNIVERSITY of EDINBURGH

I (Still) Can't Believe It's Not Better @ NeurIPS 2021—A workshop for "beautiful" ideas that *should* have worked

Motivation

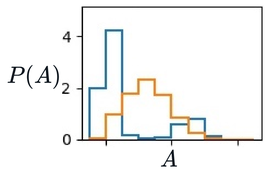
- To adapt quickly to OOD data, we need to adapt some part of our trained network
  - New task/classes → final layer [2]
  - New domain → batchnorm layers [5] or residual adapters [6]
- **But what if we don't know what type of shift will occur?**
  - Unit-level surprise!

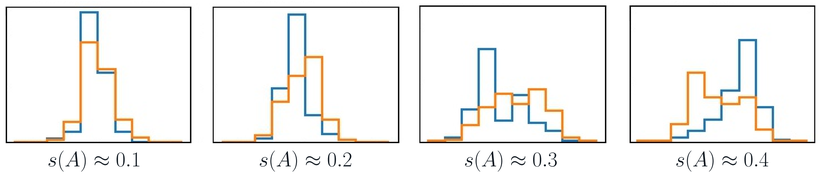


Unit-level surprise can help determine which few parameters should be updated. Purple units are unsurprised, yellow surprised. Blue indicates the weights to be updated.

Calculating unit-level surprise

- Train: $P(A)$
- Test: $Q(A)$
- $s(A) = D_{KL}(Q(A)||P(A))$
- Bayesian surprise [3]



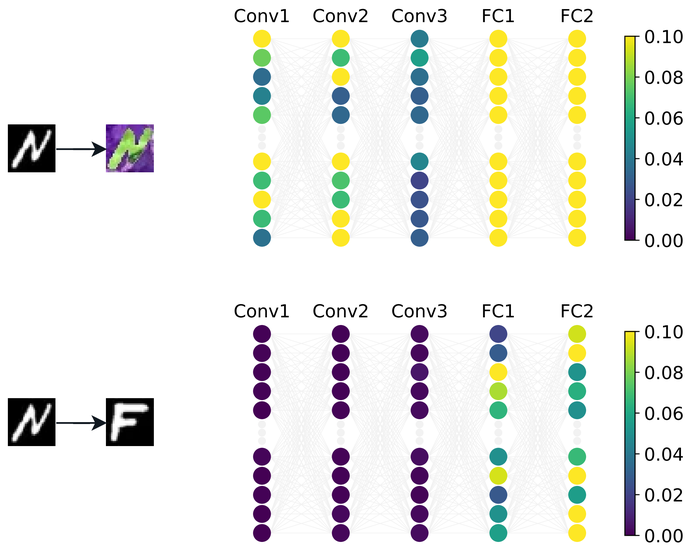


Utilising unit-level surprise

Controlled setting

EMNIST-DA dataset [1].

Surprise patterns



Update rule

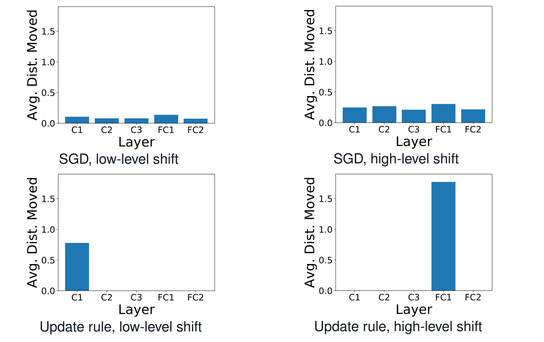
- Update if **you are surprised** but **your parents are not**
- $\boldsymbol{w} \leftarrow \boldsymbol{w} - I[s > \alpha] I[p < \beta] \cdot \eta \nabla \mathcal{L}$

Baselines

- **Conv1/FC2:** update only the first/last layer.
- **SGD:** update all layers (with SGD).
- **FlexTune [7]:** select the single best layer to update using a *test-domain* validation set.

Results

Shift	Conv1	FC2	SGD	Update rule	FlexTune [7]
Low	82.7 ± 0.4	51.5 ± 0.3	71.3 ± 2.1	82.9 ± 0.5	82.7 ± 0.4
High	0.0 ± 0.0	94.5 ± 0.9	74.4 ± 5.2	79.2 ± 4.2	94.5 ± 0.9
Avg.	57.9 ± 0.3	64.4 ± 0.3	72.2 ± 2.9	81.8 ± 1.2	86.3 ± 0.5



Roadblocks to more general settings

"Why isn't it better!?"

1. **Networks are not very modular by default**
   - Meta-learn modular structure over multiple environments?
   - #units surprised ≈ #parameters to be updated
2. **Surprise may not be sufficient**
   - Incorporate other unit-level information?
   - E.g. gradient magnitude or parameter importance [4]
3. **Validation is difficult**
   - No ground-truth update structure

Please reach out – open to collaborations!

References

[1] C. Eastwood, I. Mason, C. K. I. Williams, and B. Schölkopf. Source-Free Adaptation to Measurement Shift via Bottom-Up Feature Restoration. *arXiv:2107.05446*, 2021.

[2] C. Finn, P. Abbeel, and S. Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *34th International Conference on Machine Learning*, pages 1126–1135, 2017.

[3] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[4] J. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526, 2017.

[5] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou. Revisiting batch normalization for practical domain adaptation. In *International Conference on Learning Representations Workshop*, 2017.

[6] S.-A. Rebuffi, H. Bilen, and A. Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, pages 506–516, 2017.

[7] A. Royer and C. Lampert. A flexible selection scheme for minimum-effort transfer learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2191–2200, 2020.