not saturated by ResANN 26, the delay appears significant. Hence, EANNs may not be the best when the performance is not saturated or when the constant fraction of extra cost is critical.

5.4 Data-set Difficulty versus Adaptive Weights

In Fig. 5c, we plot the final AdaLoss weights of the same ResANN model (25,32) on CIFAR10, CIFAR100, and SVHN, in order to study the effects of the data-sets on the weights. We observe that from the easiest data-set, SVHN, to the hardest, CIFAR100, the weights are more concentrated on the final layers. This suggests that AdaLoss can automatically decide that harder data-sets need more concentrated final weights to have near-optimal final performance, whereas on easy data-sets, more efforts are directed to early predictions. Hence, AdaLoss weights may provide information for practitioners to design and choose models based on data-sets.

6 Conclusion and Discussion

This work devises simple adaptive weights, AdaLoss, for training anytime predictions in DNNs. We provide multiple theoretical motivations for such weights, and show experimentally that adaptive weights enable small ANNs to outperform large ANNs with the commonly used non-adaptive constant weights. Future works on adaptive weights includes examining AdaLoss for multi-task problems and investigating its "first-order" variants that normalize the losses by individual gradient norms to address unknown offsets of losses as well as the unknown scales. We also note that this work can be combined with orthogonal works in early-exit budgeted predictions (Guan et al., 2017; Bolukbasi et al., 2017) for saving average test computation.

Acknowledgements

This work was conducted in part through collaborative participation in the Robotics Consortium sponsored by the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Ba, L. J. and Caruana, R. Do deep nets really need to be deep? In *Proceedings of NIPS*, 2014.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 2009.

Boddy, Mark and Dean, Thomas. Solving time-dependent planning problems. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'89, pp. 979–984, 1989.

Bolukbasi, Tolga, Wang, Joseph, Dekel, Ofer, and Saligrama, Venkatesh. Adaptive neural networks for fast test-time prediction. In *ICML*, 2017.

Cai, Zhaowei, Saberian, Mohammad J., and Vasconcelos, Nuno. Learning Complexity-Aware Cascades for Deep Pedestrian Detection. In *International Conference on Computer Vision (ICCV)*, 2015.

Chen, Minmin, Weinberger, Kilian Q., Chapelle, Olivier, Kedem, Dor, and Xu, Zhixiang. Classifier Cascade for Minimizing Feature Evaluation Cost. In *AISTATS*, 2012.

Chen, Qifeng and Koltun, Vladlen. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017.

Grubb, Alexander and Bagnell, J. Andrew. SpeedBoost: Anytime Prediction with Uniform Near-Optimality. In *AISTATS*, 2012.

Guan, Jiaqi, Liu, Yang, Liu, Qiang, and Peng, Jian. Energy-efficient amortized inference with cascaded deep classifiers. In *arxiv preprint, arxiv.org/abs/1710.03368*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.