Supplementary Material: Transferability of Adversarial Active Learning for Deep Networks

When faced with a new classification problem, we don't know the hyperparameters that are best suited for the problem. One can argue that a network with high capacity is likely to give high accuracy and is sufficient enough when combined with some human expertise on the problem: several architectures have been handcrafted for specific tasks and are available online (Chollet and others 2015). Still, their efficiency is known for large datasets. (Yanyao Shen 2018) pointed out a flaw in active learning: their active learning heuristics perform well if and only if they use it on a lightweight architecture instead of the architecture of reference for NER classification. Such an issue is inherent to active learning. Combining model selection with active learning has been investigated for shallow models (Sugiyama and Rubens 2008). One of the main issues raised is that multiple hypotheses (i.e. candidate networks) trained in parallel may require labeling different training points.

Furthermore, (Fawzi et al. 2017) empirically demonstrated a strong correlation between the vulnerability of a network to small adversarial perturbations and an asymmetry in the curvature of its decision boundary: if a model is not robust to an adversarial attack, it is likely that the curvature in that direction is negative and vice-versa. Thus, not only that the decision boundaries would lie close one to another but they would likely share some strong topological properties. Based on those arguments, we assume adversarial queries are useful for a diverse set of architectures, not only for the CNN they have been queried for.

First of all, we assert this assumption by evaluating the classification regions overlap between LeNet5 and VGG8; both trained on the QuickDraw dataset. Results are presented in Figure 2. We observe that most of the test samples share the same classification regions (● blue dots) for both networks, LeNet5 and VGG8, while few of them (● red dots) are in different classification regions. Notice that, this does not mean that the networks disagree on their prediction on such samples but put them in different classification regions. Thus, it appears than CNNs may have significant overlaps on their classification regions, at least for LeNet5 and VGG8.

On the contrary, we analyzed the transferability of deterministic and Bayesian uncertainty by plotting the confidency of test samples on VGG8, given their confidency on LeNet5, on the Quick Draw dataset. We denote by *Bayesian uncertainty*, the average uncertainty over a committee of models sampled with dropout. We refer as confidency the highest probability emitted by the network. In case of transferability, either for deterministic or bayesian uncertainty, we would expect the confidency values of the test samples to be aligned: the uncertainty of a sample given one model would then reflect the uncertainty in the second model, which is not the case for both type of uncertainty as depicted in Fig

	DFAL	CORE-SET	RANDOM
LeNet5→ VGG8	**97.80**	96.90	94.46
VGG8→ LeNet5	**97.93**	97.40	95.31

(a) *MNIST*

	DFAL	CORE-SET	RANDOM
LeNet5→ VGG8	**92.87**	91.06	89.94
VGG8→ LeNet5	89.23	89.41	**89.42**

(b) *Quick-Draw*

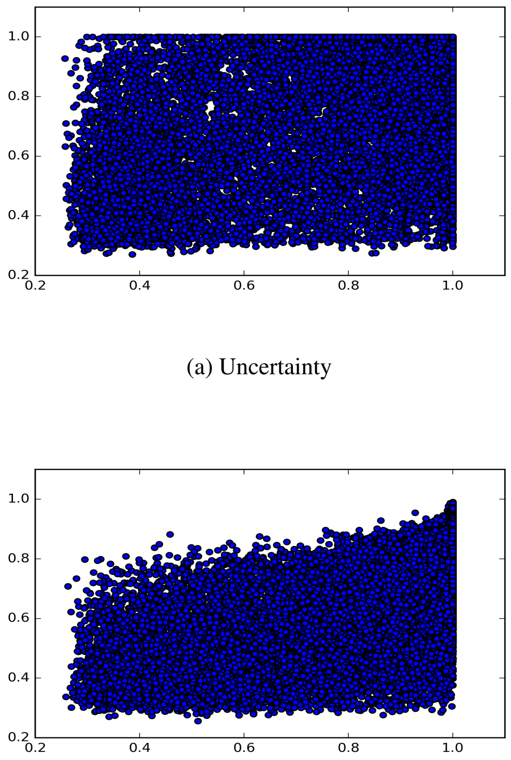
	DFAL	CORE-SET	RANDOM
LeNet5→ VGG8	**99.40**	99.12	97.08
VGG8→ LeNet5	**98.75**	98.50	98.07

(c) *Shoe-Bag*

Table 1: Comparison of the transferability of DFAL and CORE-SET with 1000 annotations

When it comes to the transferability, we empirically demonstrate DFAL's potential for a baby task. We compare the test accuracy of DFAL and CORE-SET transferred dataset on 1000 samples in Table 1. Surprisingly the transferred queries from CORE-SET perform better than random. However, the transferred queries from DFAL outperform CORE-SET and RANDOM.

However, it has been shown that under some constraints of similarities between the architectures, adversarial examples of a network A are very likely to be adversarial for a network B. This turns to be a significant advantage for our adversarial active learning strategy since the training set built with DFAL for the network A will then be very likely to be a relevant training set for the network B.



(a) Uncertainty

(b) Bayesian Uncertainty

Figure 1: **Transferability of deterministic and Bayesian uncertainty**: confidency of test samples computed on VGG8 given their confidency on LeNet5. Both networks are trained on the whole training set

When it comes to the transferability, we empirically demonstrate DFAL potential on a toy task: in Figure 3 we have recorded *Shoe-Bag* adversarial queries for LeNet5 and use them for training VGG8. We did the same protocol on *MNIST* and LeNet5, in Figure 4. While the test accuracy achieved is lower than with the adversarial active strategy directly applied for the training of VGG8, the transfered training set achieves better accuracy than RANDOM. When reaching 1000 annotated samples, it is also better than considering other active criteria designed for VGG8. We go further and compare the accuracy on 1000 test samples of DFAL and CORE-SET trained on the transfered training set in Table 1. Surprisingly the transfered queries from CORE-SET perform better than RANDOM. However, in almost every case, the transfered queries with DFAL outperform CORE-SET and RANDOM. We have therefore shown the relevance of transfering adversarial examples generated within active learning from one architecture to another. This opens up promising perspective for the design of tractable methods to explore network architectures.

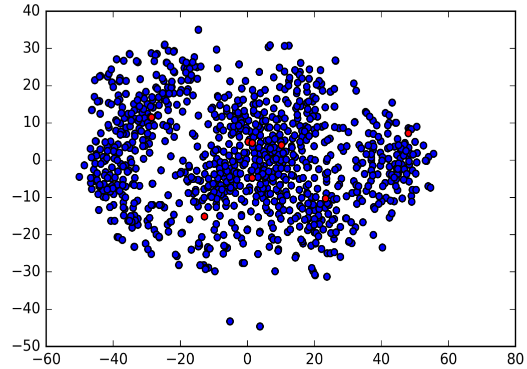


Figure 2: **Overlap of the classification regions of LeNet5 and VGG8 trained on the QuickDraw datasets**. Blue dots ● are test samples that fall into the same classification regions for both networks, while red dots ● do not fall into the same region. We proceed by looking for a convex path so that every point in that path share the same prediction. To do so, we check the validity of the path in the convex combinations of consecutive anchor points, as proposed by Fawzi *et al.* . Then we check, whether paths exist for both networks and project the test samples in a two-dimensional space using T-SNE.

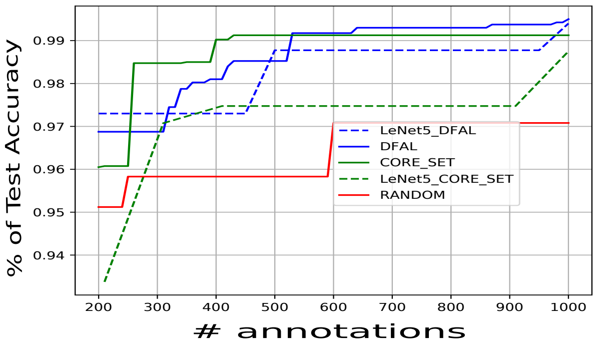


Figure 3: Evolution of the test accuracy for (*Shoe-Bag*, VGG8) trained with different labeled training set: we compare the efficiency of DFAL and CORE-SET built on LeNet5 (LeNet5_DFAL and LeNet5_CORE-SET) and transfered to VGG8.

References

[Chollet and others 2015] Chollet, F., et al. 2015. Keras.

[Fawzi et al. 2017] Fawzi, A.; Moosavi-Dezfooli, S.-M.; Frossard, P.; and Soatto, S. 2017. Classification regions of deep neural networks. *arXiv preprint arXiv:1705.09552*.

[Sugiyama and Rubens 2008] Sugiyama, M., and Rubens, N. 2008. A batch ensemble approach to active learning with model selection. *Neural Networks* 21(9):1278–1286.

[Yanyao Shen 2018] Yanyao Shen, Hyokun Yun, Z. C. L. Y. K. A. A. 2018. Deep active learning for named entity recognition. *International Conference on Learning Representations*.

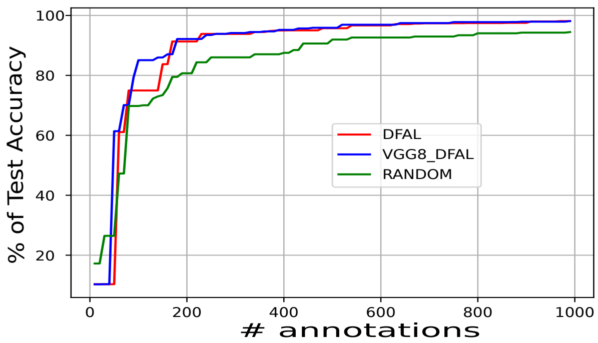


Figure 4: Evolution of the test accuracy for (*MNIST*, LeNet5) trained with different labeled training set: we compare the efficiency of DFAL on LeNet5 (VGG8_DFAL and transfered to LeNet5.