Experiments in Combining Boosting and Deep Stacked Networks

Manuel Montoya-Catalá, Ricardo F. Alvear-Sandoval, Aníbal R. Figueiras-Vidal
Signal Theory and Communications Department of the Univ. Carlos III of Madrid

Universidad Carlos III de Madrid

Objectives

1. To develop new Deep Learning architectures and training algorithms
2. To combine Deep Learning architectures with Boosting, creating systems with high expressivity and resistance to overfitting
3. To propose flexible emphasis functions that can moderate boosting

Introduction

Both boosting and deep stacking sequentially train their units taking into account the outputs of the previously trained learners. This parallelism suggests that it exists the possibility of getting some advantages by combining these widely known techniques, i.e., emphasis and injection, in appropiate manners.
We propose a first mode for such a combination by simultaneously applying a general and flexible enough emphasis function and injecting the aggregated previous outputs to the learner which is being designed. We call this kind of classification mechanism Boosted and Aggregated Deep Stacked Networks (B-ADSNs).

Deep Stacked Networks (DSNs)

DSNs are a family of Deep Learning architectures in which each unit (layer) consists of a MLP whose input is:

– The observed features and
– The outputs of all previously trained learners

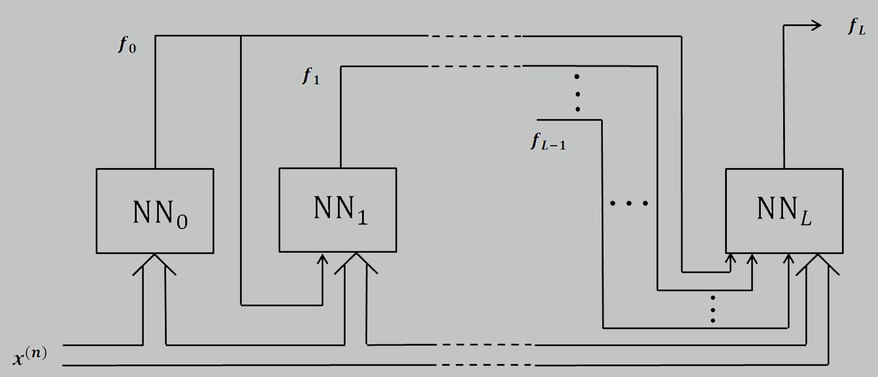

Figure 1: Deep Stacked Network architecture

- The output of the ensemble is the output of the last unit.
- This architecture has a high expressivity but it is prone to overfitting.
- The dimensionality of the input space increases with the number of units.

Boosting

Boosting is an ensemble method in which learners are sequentially trained using information from the aggregation of all previously trained units.

– During the training of every unit, each sample is assigned an emphasis value $\mathbf{e_l(x^{(n)})}$. The set of emphasis value generates a discrete distribution over the samples.

– This value indicates the "importance" of the sample during training. It weights its contribution to the loss function.

– The emphasis function assings this value to every sample according to their current error value. Samples that are easily classified by the system obtain low values of importance.

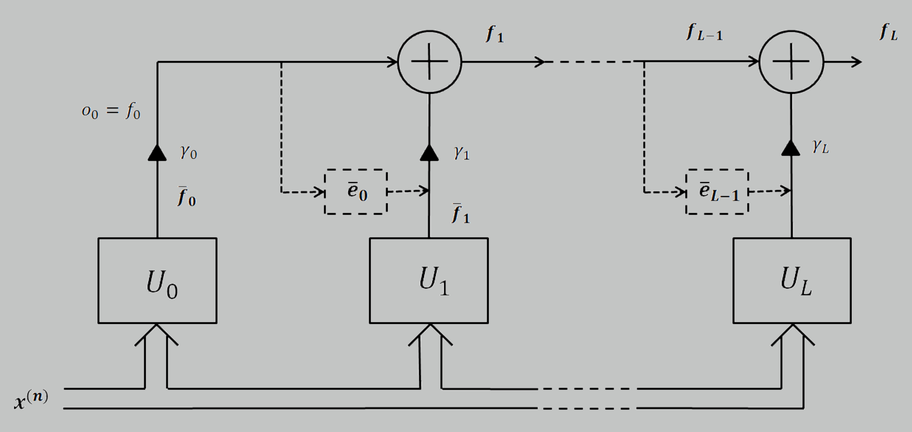

Figure 2: Boosting architecture

- The output of the ensemble is a linear combination of all unit outputs.
- Resistant to overfitting.
- Boosting ensembles usually require weak learners.

Boosted Aggregated Deep Stacked Networks

Combination of DSNs and Boosting by means of an aggregated output injection and a flexible emphasis function. Each unit has two additional sources of information:

– Injection of the aggregated output of all previously trained units.
– Emphasis values of the samples.

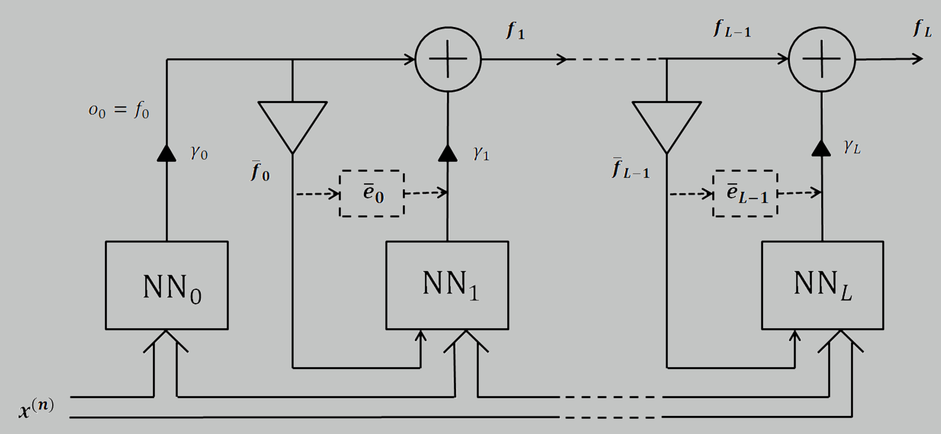

Figure 3: Boosted Aggregated Deep Stacked Networks architecture

- The emphasis function $\mathbf{e_l^*(x^{(n)})}$ constains the hyperparameters α and β which control the intensity of the boosting so that the optimal combination of boosting and outputs' injection can be found.

$$\mathbf{e_l^*(x^{(n)}) = \frac{\alpha}{N} + \frac{1-\alpha}{Z_l}\left[e_l^*(x^{(n)})\cdot \exp\left(\beta\left(t^{(n)} - \bar{f}_{l-1}^2(x^{(n)})\right)^2 - (1-\beta)\bar{f}_{l-1}^2(x^{(n)})\right)\right]} \quad (1)$$

- The learners are trained using Online Backpropagation.
- The values of α, β and the hyperparameters of the architecture are obtained by means of crossvalidation (CV).

Results

The following table contains the average error rate $\pm$ standard deviation for the considered architectures – deep learning architectures B-ADSN including the restricted version ADSN, and the boosting ensemble B1. CV values for α, β, $\mathbf{N_{ep}}$ and $\mathbf{H}$ are also included.

	BADSN ($\alpha/\beta/N_{ep}/H$)	ADSN (N_{ep}/H)	B ($\alpha/\beta/N_{ep}/H$)
Aba	18.4 ± 0.2 $(0.9, 0.5, 100, 12)$	18.6 ± 0.2 $(100, 16)$	19.1 ± 0.1 $(0.9, 0.1, 100, 10)$
Kwo	11.5 ± 0.05 $(0.9, 0.8, 150, 10)$	11.6 ± 0.05 $(150, 15)$	11.6 ± 0.05 $(0.8, 0.9, 100, 5)$
Wav	10.4 ± 0.14 $(0.2, 0.9, 100, 4)$	11.8 ± 0.20 $(100, 7)$	10.1 ± 0.1 $(0.7, 0.7, 100, 5)$

Figure 4: Results Table

- The accuracy of the system varies smoothly with respect to the parameters α and β.
- The performance of the system converges with the number of layers.

Conclusions

- The combination of the expressivity of DSNs and the resistance to overfitting of boosting can be succesfull.
- A flexible emphasis function is required to moderate the boosting contribution.
- There are many other possible combinations of boosting and deep learning.

References

[1] L.Deng and D.Yu, "Deep convex net: A scalable architecture for speech classification", in *Proc. Interspeech 2011*, pp. 2285-2288. Florence (Italy), 2011.
[2] R. E. Schapire and Y. Freund. Boosting: Foundations and Algorithms. Cambridge, MA: MIT Press, 2012.
[3] V. Gómez-Verdejo, M. Ortega-Moral, J. Arenas-García and A. R. Figueiras-Vidal, "Boosting by weighting critical and erroneous samples", *Neurocomputing*
[4] J. T.-Y. Kwok. "Moderating the outputs of support vector machine classifiers"

REF- 96149218249375386387392398419420435502