Partial Boosting of Deep Stacked Networks

Manuel Montoya-Catalá,
Ricardo F. Alvear-Sandoval,
Aníbal R. Figueiras-Vidal
UC3M / RAIng

Contents

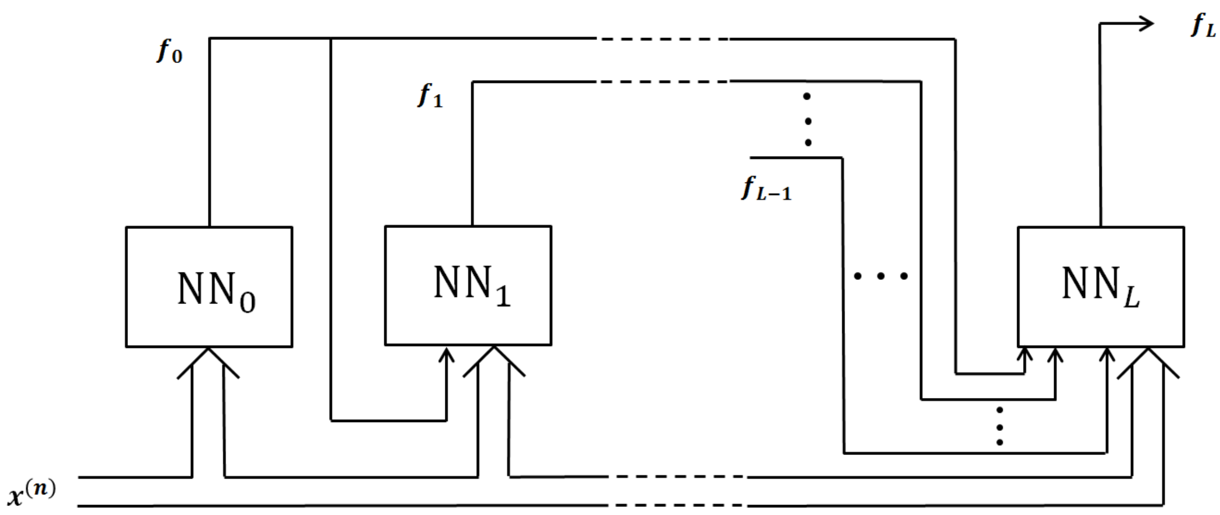
1. Deep Stacked Networks (DSNs)

Deep Learning architecture.

Each unit consists of a MLP whose input is:

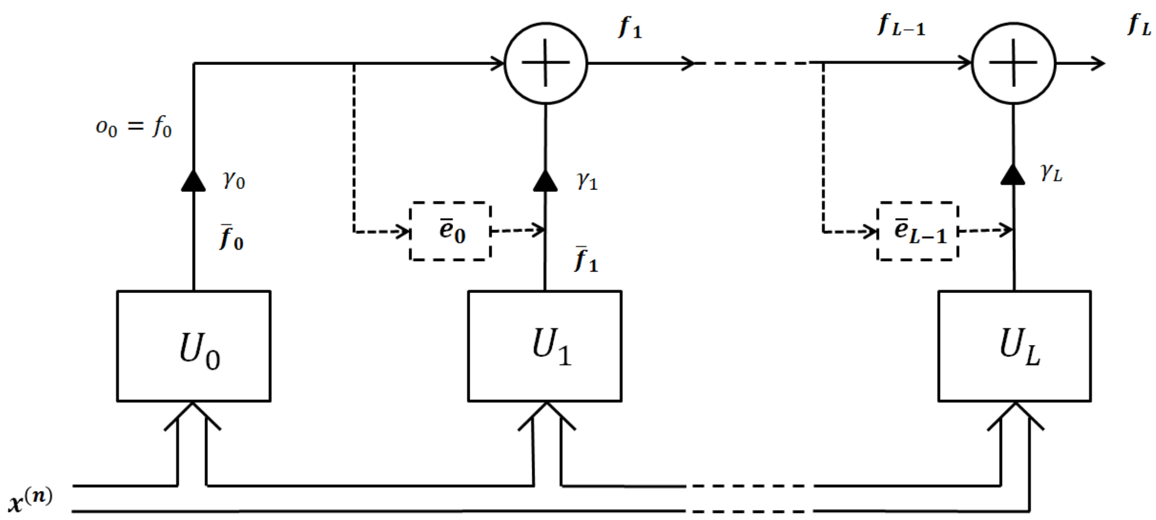
- The observed features and
- the outputs of all previously trained learners.

The output of the DSN is the output of the last unit.



2. Boosting

- Ensemble method in which weak learners are sequentially trained using information from the aggregation of all previously trained units.
  - o Samples are weighted using a emphasis function.
- The output of the ensemble is a linear combination of all unit outputs.
- Resistant to overfitting.



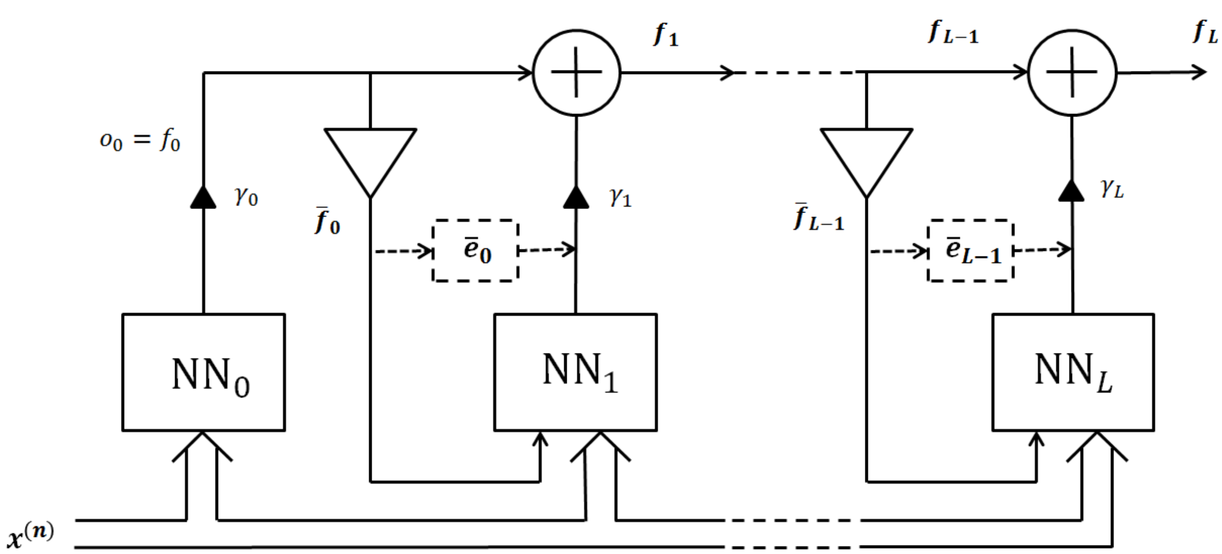
3. Boosted Aggregated Deep Stacked Networks

Combination of DSNs and Boosting by means of an aggregated output injection and a flexible emphasis function. Each unit has 2 additional sources of information:

- Injection of the aggregated output of all previously trained units
- Emphasis function

$\alpha, \beta: CV$

$$p(\boldsymbol{x}^{(n)}) = \frac{\alpha}{N} + \frac{1-\alpha}{Z_l}\left[\beta\left(t^{(n)} - \bar{f}_l(\boldsymbol{x}^{(n)})\right)^2/4 + (1-\beta)\left(1 - \bar{f}_l(\boldsymbol{x}^{(n)})^2\right)\right]$$



4. Experiments

Experiments performed over a set of modetate size binary problems.

Units are MLP sequentially trained using Online Back-Propagation.

Explored values of the non-trainable elements in the CV-search are:

- Number of hidden neurons from 2 to 30.
- Number of epochs from 25 to 200.

	B1-ADSN	B2-ADSN	ADSN	B1	B2
aba	18.4 ± 0.2	18.5 ± 0.2	18.6 ± 0.2	19.1 ± 0.1	19.0 ± 0.1
ima	2.9 ± 0.3	2.9 ± 0.4	3.0 ± 0.3	3.2 ± 0.5	3.2 ± 0.2
hep	6.6 ± 0.0	6.7 ± 0.4	8.0 ± 0.4	6.6 ± 0.5	6.7 ± 0.5

TABLE II

% AVERAGE ERROR RATE $\pm$ STANDARD DEVIATION FOR THE CONSIDERED ARCHITECTURES

5. Properties of the B-ADSNs

- Performance varies smoothly with α, β (some discontinuity for extreme values of alpha).

- Harder problems require extreme values of α.

- Smaller problems (hep) requiere an intermediate α which seems to figh the initial overfitting.

- The value of β is problem dependent.

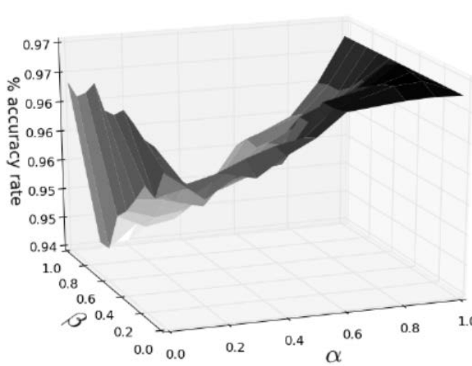


Fig. 2. % average accuracy rate for the Ima dataset with respect to α and β.

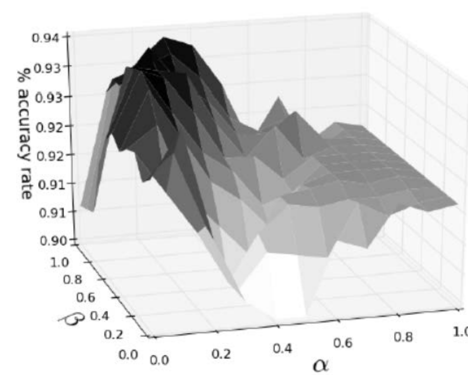


Fig. 3. % average accuracy rate for the Hep dataset with respect to α and β.

6. 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 modetate the boosting contribution.
- There are many other possible combination of boosting and deep learning.