Multi-Task Learning in Deep Neural Networks at EVALITA 2018

Andrea Cimino, Lorenzo De Mattei and Felice Dell'Orletta

Italia**NLP** Lab - *www.italianlp.it*

Instituto di Linguistica Computazionale "Antonio Zampolli" - CNR - Pisa

Dipartimento di Informatica - Università di Pisa

EVALITA, Turin, November 12th, 2018

Our participation at EVALITA 2018

- In EVALITA edition the majority of task were classfication (binary) tasks, among them:
  - **ABSITA** (Aspect Based Sentiment Analysis)
  - **HaSpeeDe** (Hate Speech Detection)
  - **GxG** (Gender X-Genre, author profiling in terms of gender)
  - **IronITA** (Irony and Sarcasm Detection)
- We designed and developed a general purpose system based on deep neural networks and evaluated the performance on these 4 shared tasks

In this presentation...

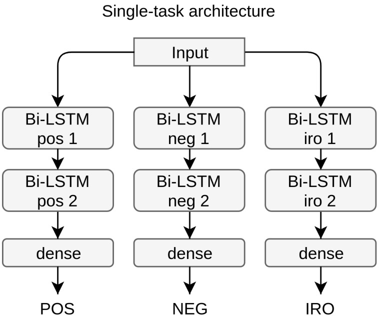
- I will describe our multi task learning our architecture
- I will report the resources and experiments we performed on the HaSpeeDe 2018 shared task

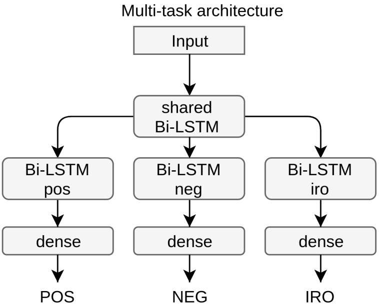
Single-task Learning vs Multi-task Learning

- In classical **single task learning**, given a labeled dataset, we create a statistical model we employ on unseen examples
- But many times we deal with a dataset with multiple annotation (Sentipolc 2016), or with multiple datasets with related annotation such as hate speech, sentiment polarity and irony (HaspeeDe, ABSITA, IronIta)
- Wouldn't be nice if a single statistical model would exploit all these datasets to improve the final model?
- Here **multi-task learning** comes to the rescue

ITALIAN NATURAL LANGUAGE PROCESSING LAB@ILC
www.italianlp.it

Single-task & Multi-task architectures





- The Sentipolc 2016 case: very few positive IRO labels compared to POS and NEG
- We conducted a study in *Multi-Task Learning in Deep Neural Network for Sentiment Polarity and Irony classification* (De Mattei et al., 2018): MTL improvements w.r.t. STL

www.italianlp.it

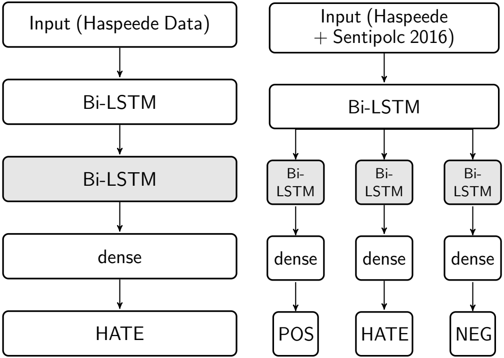
Multi Task Learning

Actually MTL is not a very novel idea:

- *A unified architecture for natural language processing: deep neural networks with multitask learning*(Collobert et al., 2008)
  - Part-Of-Speech Tagging, Chunking, Semantic Role Labeling, Language Models

But GPUs and DL frameworks opened the access to this technique:

- Machine Translation: *Multi-Task Learning for Multiple Language Translation* (Dong et al. 2015)
- Sentence Compression: *Improving sentence compression by learning to predict gaze*, (Klerke et al., 2016)



(a) STL Models (b) MTL Models

STL and MTL architectures. The 1-layer Bi-LSTM models do not include the components marked with a gray background.

Train STL and MTL

- STL: we performed three different training steps, one for each task.
- MTL: we run a shared training by iteratively optimizing at each step a loss function for each task. For the MTL the global loss function is given by the sum of the three individual loss functions.

In STL and MTL architectures, we stopped the training after 100 epochs without improvements of the loss function on the validation set, choosing the parameters with the best performances.

5-fold vote approach

Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

- Train 5 different models, using 5 different validation sets
- Use a vote approach in classification phase
- Each validation fold follows the label distribution of the training set

Resources and DNN Features

- We downloaded 50.000.000 Tweets: built 3 lexicons for neutral, positive and negative tweets using seed words or emojis
  - Each TW_{POS}, TW_{NEG}, TW_{NEU} entry contains the relative frequency of the word in each lexicon
  - What we expect: $freq_{POS}(bello) > freq_{NEU}(bello) > freq_{NEG}(bello)$
  - We used this as feature in the DNN classifier
- ItWac corpus word embeddings
- Twitter corpus word embeddings
- Automatically postagged the dataset → POStags

Results (Test Set)

Configuration	**TW**	**FB**	**C_TW**	**C_FB**
baseline	0.403	0.404	0.404	0.403
best official system	0.799	0.829	0.699	0.654
linear SVM	0.798*	0.761	0.658	0.451
1L STL	0.793	0.811*	0.669*	0.607*
2L STL	0.791	0.812	0.644	0.561
1L MTL	0.788	0.818	**0.707**	**0.635**
2L MTL	0.799*	**0.829***	0.699*	0.585*
1L MTL NO SNT	**0.801**	0.808	0.709	0.620
1L STL NO FOLD	0.785	0.806	0.652	0.583

- Our system ranked 1st in the *TW*, *FB*, and $Cross_{TW}$ tasks

ITALIAN NATURAL LANGUAGE PROCESSING LAB@ILC
www.italianlp.it

Results (Test Set)

Configuration	TW	FB	C_TW	C_FB
baseline	0.403	0.404	0.404	0.403
best official system	0.799	0.829	0.699	0.654
linear SVM	0.798*	0.761	0.658	0.451
1L STL	0.793	0.811*	0.669*	0.607*
2L STL	0.791	0.812	0.644	0.561
1L MTL	0.788	0.818	**0.707**	**0.635**
2L MTL	0.799*	**0.829***	0.699*	0.585*
1L MTL NO SNT	**0.801**	0.808	0.709	0.620
1L STL NO FOLD	0.785	0.806	0.652	0.583

- SVMs almost near DL in the TW task
  - Domain shift problems, -0.20 f-score in the C_{FB} task

ITALIAN NATURAL LANGUAGE PROCESSING LAB@ILC

www.italianlp.it

Results (Test Set)

Configuration	TW	FB	C_TW	C_FB
baseline	0.403	0.404	0.404	0.403
best official system	0.799	0.829	0.699	0.654
linear SVM	0.798*	0.761	0.658	0.451
1L STL	0.793	0.811*	0.669*	0.607*
2L STL	0.791	0.812	0.644	0.561
1L MTL	0.788	0.818	**0.707**	**0.635**
2L MTL	0.799*	**0.829***	0.699*	0.585*
1L MTL NO SNT	**0.801**	0.808	0.709	0.620
1L STL NO FOLD	0.785	0.806	0.652	0.583

- MTL **helps!**
  - Particularly true in the out domain tasks, +0.04 f-score

www.italianlp.it

Results (Test Set)

Configuration	TW	FB	C_TW	C_FB
baseline	0.403	0.404	0.404	0.403
best official system	0.799	0.829	0.699	0.654
linear SVM	0.798*	0.761	0.658	0.451
1L STL	0.793	0.811*	0.669*	0.607*
2L STL	0.791	0.812	0.644	0.561
1L MTL	0.788	0.818	**0.707**	**0.635**
2L MTL	0.799*	**0.829***	0.699*	0.585*
1L MTL NO SNT	**0.801**	0.808	0.709	0.620
1L STL NO FOLD	0.785	0.806	0.652	0.583

- Vote approach improves performances
  - In each task significant gains

Results (Test Set)

Configuration	TW	FB	C_TW	C_FB
baseline	0.403	0.404	0.404	0.403
best official system	0.799	0.829	0.699	0.654
linear SVM	0.798*	0.761	0.658	0.451
1L STL	0.793	0.811*	0.669*	0.607*
2L STL	0.791	0.812	0.644	0.561
1L MTL	0.788	0.818	**0.707**	**0.635**
2L MTL	0.799*	**0.829***	0.699*	0.585*
1L MTL NO SNT	**0.801**	0.808	0.709	0.620
1L STL NO FOLD	0.785	0.806	0.652	0.583

- Distant supervision lexicons behaviour: unstable
  - Overfitting?

www.italianlp.it

Conclusions

- First deep multi-task learning system at EVALITA
- Our system ranked 1st in the *TW*, *FB*, and $Cross_{TW}$ tasks
- The MTL architecture showed performance improvements w.r.t. the STL counterpart
  - Particularly true in the C_{FB} outdomain task (+3 f-score points)

Thanks for your attention!
Questions?