UniNE at CLEF 2018: Character-based Convolutional Neural Network and ResNet18 for Twitter Author Profiling

Nils Schaetti

University of Neuchatel

nils.schaetti@unine.ch

10th of September, 2018

PAN@CLEF 2018 : Author Profiling

- Common ground to allow the large study in author profiling on social network data.
- Collections of tweets and images
  1. 100 tweets per authors.
  2. 10 images per authors.
- Three collections : English, Spanish and Arabic.
- Predict gender : two-classes problem (*male*, *female*).
- Evaluation measures :
  - Accuracy on tweets
  - Accuracy on images
  - Joint classification accuracy
- All models evaluated with the same methodology on the *TIRA* platform.

PAN17 & PAN18 dataset

Corpus		Training			
		Authors	Tweets	Images	Gender
English	PAN17	3'600	360k	0	1'800
	PAN18	3'000	300k	30k	1'500
	Total	**6'000**	**660k**	**30k**	**3'300**
Spanish	PAN17	4'200	420k	0	2'100
	PAN18	3'000	300k	30k	1'500
	Total	**7'200**	**720k**	**30k**	**3'600**
Arabic	PAN17	2'400	240k	0	1'200
	PAN18	1'500	150k	15k	750
	Total	**3'900**	**390k**	**15k**	**1'950**

Table: Corpora statistics

Proposed models

Proposed models for 2018 PAN@CLEF Author Profiling

- A **character-based Convolutional Neural Network** (CNN) with 3 different pattern sites (2, 3 and 4) and temporal max-pooling layers for tweet classification.
- The **Residual Neural Network** with 18 layer (ResNet18) for image classification.

Character-based Convolutional Neural Network

- Convolutional model based on a trained embedding layer of character bi-grams.
- The embedding layer reduces the dimensionality of the inputs (300).
- Language-independent model.
- How to find the best patterns of character bi-grams separating the two classes.
- Model very effective on authorship attribution of short texts with more that 8'000 authors.
- Additional space is felt with a special index (zero) if the tweet is shorter than 160 (fixed-size).
- All URLs are removed and tweets are transformed to lower case.

Character-based Convolutional Neural Network

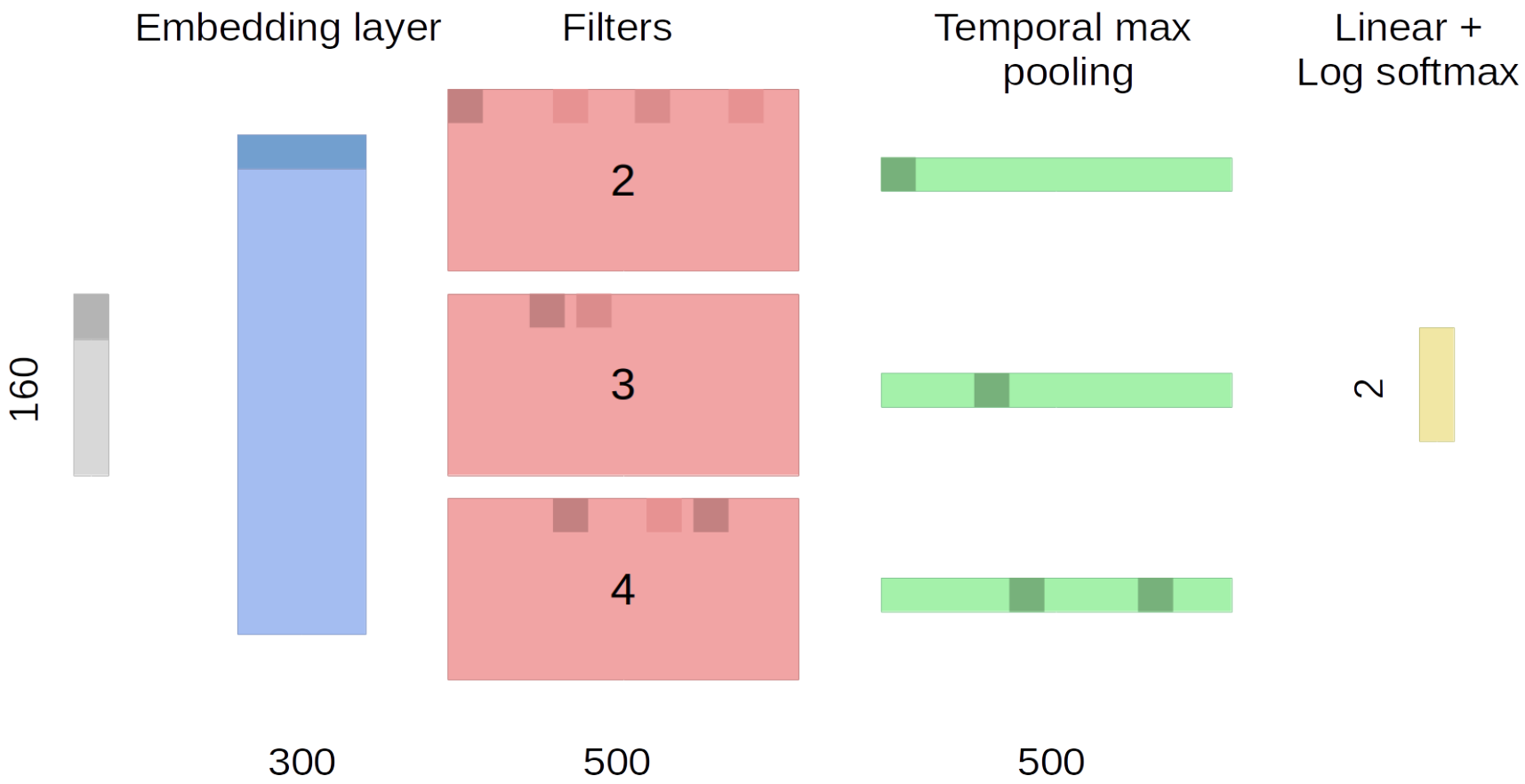

Figure: Structure of the character 2-grams based *Convolutional Neural Network* with the following layers : embedding layer (dim=300), three convolutional layers (kernel size 2, 3 and 4), three temporal max pooling layers, a final linear layer of size 2 with log softmax outputs.

Character-based Convolutional Neural Network

- Dataset split into training and validation sets.
  - The whole PAN17 dataset for training.
  - 90% of the PAN18 dataset for training.
  - 10% of the PAN18 dataset for validation.
- We used *TorchLanguage*, a package based on *pyTorch* designed for Natural Language Processing.
- *Stochastic gradient descent* algorithm to train our model with,
  - *Learning rate* of 0.001.
  - *Momentum* of 0.9.
  - *Cross-entropy* as loss function.
  - Trained for 150 iterations.
- Average probabilities over the 100 tweets.

ResNet 18

- Residual Neural Network.
- Deep neural network using *skip connections* or *short-cuts*.
- Introduced in 2015 and won several competition in computer vision.
  - 1st place in the ILSVRC 2015 classification competition with top-5 error rate of 3.57%.
  - 1st place in ILSVRC and COCO 2015 competition.
- Structures similar to the brain's cerebral cortex.
- Avoid the well known problem of vanishing gradients using activation from a previous layer until the next one has learned its weights.
- Average probabilities over the 10 images.

ResNet 18

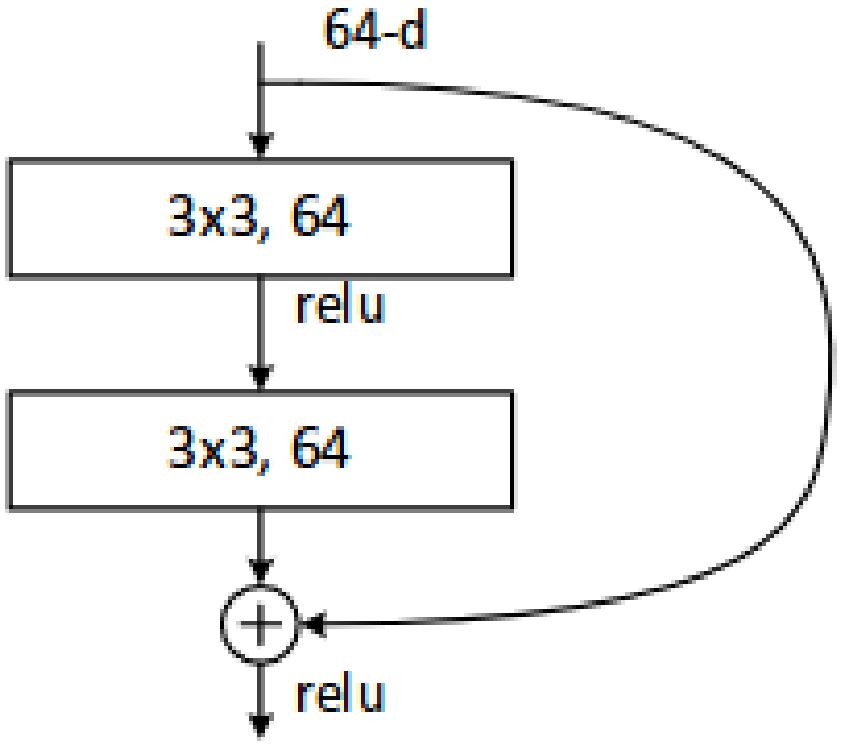

Figure: ResNet18 building block.

ResNet 18

- Model trained on the whole data set independently of the language collection.
- Models with 18, 34, 50, 101 and 152 layers available in *TorchVision*.
- We choose the model with 18 layers to avoid over-fitting.
- We used 90% of the PAN18 image training set for training and 10% to evaluate the performances at each iteration.
- We used the *stochastic gradient descent algorithm* to train our model with a *learning rate* of 0.001, a *momentum* of 0.9 and *cross-entropy* as loss function.
- Model with 32 layers, and *AlexNet* tested on validation set but with lower performances.

Results

Results on the test dataset (TIRA)

Corpus	Gender	Images	**Both**	*Random*
English	0.7711	0.5782	**0.7711**	0.5000
Spanish	0.7359	0.5763	**0.7359**	0.5000
Arabic	0.7390	0.5430	**0.7390**	0.5000
Overall	0.7487	0.5658	**0.7487**	0.5000

Table: Evaluation for the three *test* collections

Results

Overall positioning

	Team	Arabic	English	Spanish	Avg.
1	takahashi18	0.7850	0.8584	0.8159	0.8198
	...				
15	vondaeniken18	0.7320	0.7742	0.7464	0.7509
16	Schaetti	0.7390	0.7711	0.7359	0.7487
17	aragonsaenzpardo18	0.6670	0.8016	0.7723	0.7470
18	bayot18	0.6760	0.7716	0.6873	0.7116

Table: Positioning in the PAN18 challenge

Future research

- Use completely independent training and validation set to avoid under/over-fitting.
- Add a *Drop-Out* layer after max-pooling to avoid over-fitting.
- Try different number and sizes of pattern filters.
- Try with character, character 3-grams, etc.
- Adagrad, Adadelta, RMSProp, with various parameters.

Questions