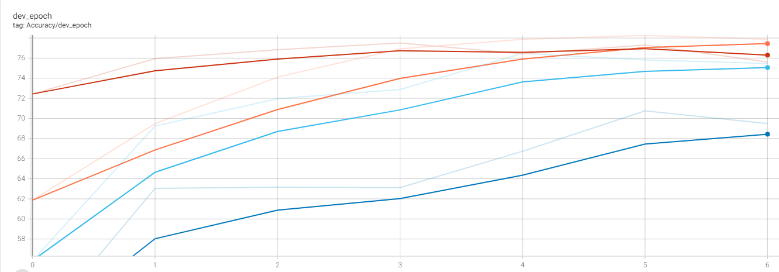
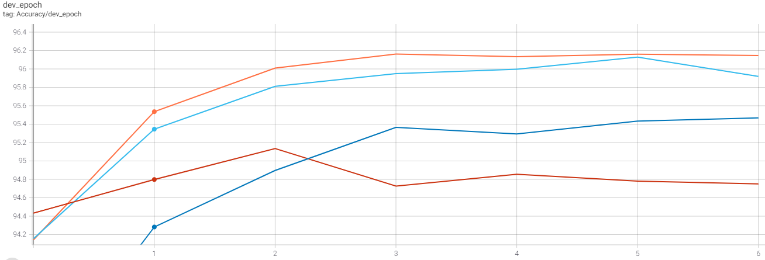
Part 5

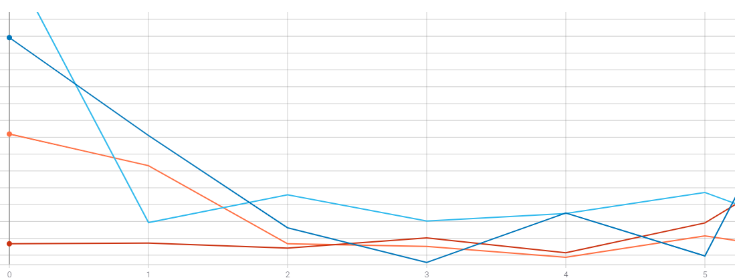
First, in order to compare all the runs, we freeze the hyper parameters and the model seed in sake of comparability and only changed the model. Both between part 1,3,4 and 5 and both between the different window size and number of filters. See graphs below, when the - orange is the part 5 and the red is part 4 with the same hyper parameters.

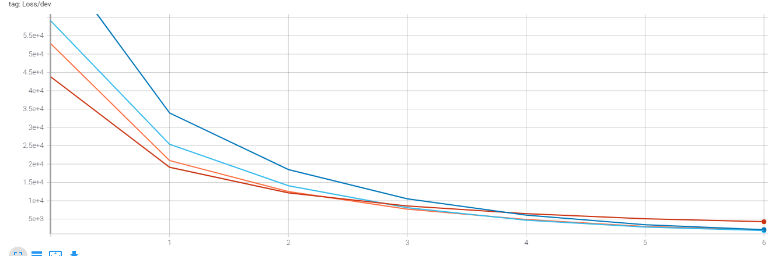
In the technical view, we chose to pad the words to the length of 20 taking, and if there is a word with more than 2 characters we would take only the first 20 and only their vectors would be an input to the Conv. We chose the length of 20 by the distribution of the length in the English language.

As you can see, the model in part 5 out preforms all the other model (at list in this specific setup – disclaimer, we didn’t check the mean of various seeds due lack of computation power and lack of time). Although there is slight improvement in the NER, it is by 2 points, but you can see that if we continued more than 7 epochs we would reach an higher improvement. Regarding the POS, we believe 96.2 is close to the limit of the data. And it is really hard to tell if there is more room for improves although the CNN Char model does yield the best results.

NER Accuracy – Part 5 is the Orange POS Accuracy – Part 5 is the Orange



Loss Dev all models NER and POS – part 5 is the orange



We examined a set of 3 types of filters: 20, 30 (the one used in the paper) and 40 and a set of 3 window sizes 3 (the one used in the paper), 4 and 5.

**Pos:**

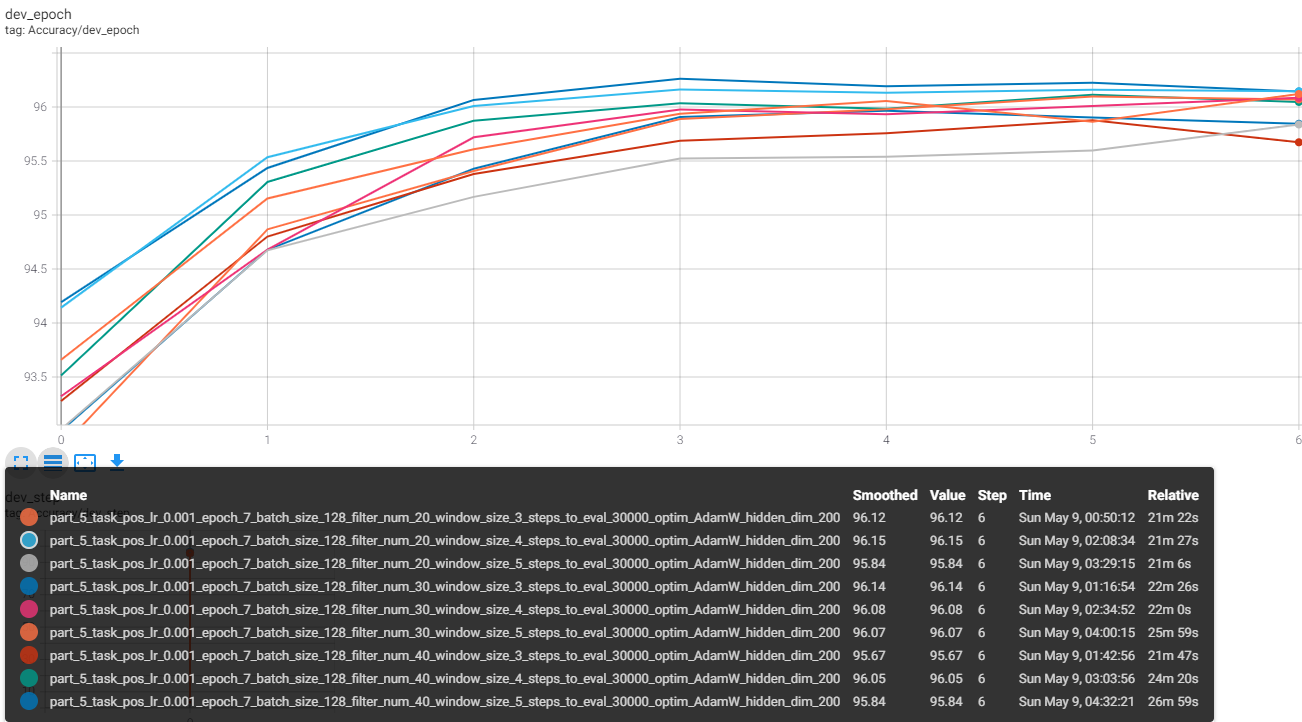
Window Effect:

We can observe in the graph below that there is almost no difference between window size 4 and window size 3 where all the results were very similar to each other. In the other hand, size 5 was a bit more scattered and with a drop of half precent. We believe that it due to the fact that there is more signals is triplet or quadrants of words and a window of five is a bit too much in order to get signal for the POS tag. Also many of the words are shorter than 5 words what would result a vast use of padding and maybe more noise.

Filter Effect:

In the sense of filters 20 was the best with 96.14, but the difference was really minor between the 3 (96.07-96.14)

In the graph below you may see the various accuracy of the runs



.

**NER:**

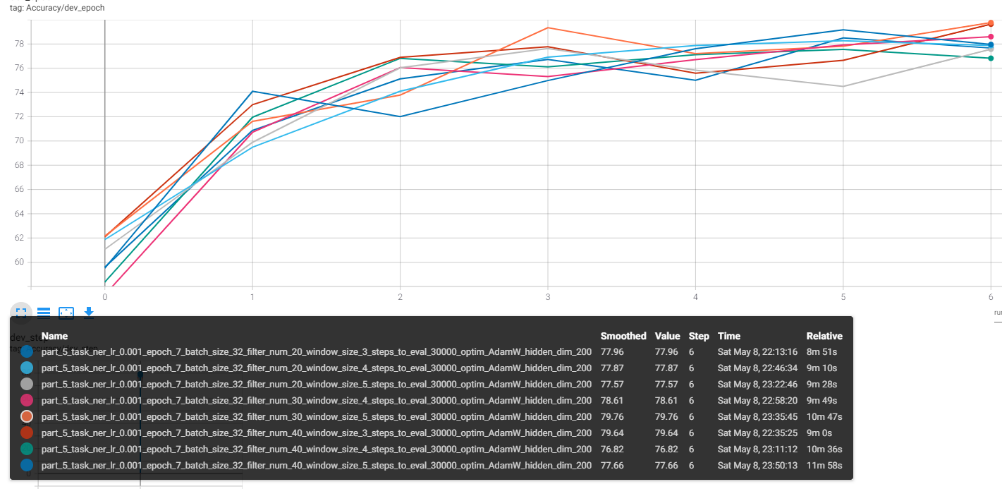
Window Effect:

Differently from the POS in the NER we noticed the opposite trend, when bigger window size yielded better results. 79.76 for window size 5, 78.61 for window size 4 and 76 for windows 3. We thought that maybe in NER all the characters in the word has a lot of meaning not only small patterns, and this may explain it.

Filter Effect:

Same as the window in the NER more == better when the best result was 79.64 for 40 filters but in the other hand 20 was better than 30 (77.9 vs 76). So it was hard for us to reach a conclusion.

In summary bigger windows were better for NER and smaller were better for POS without a clear effect of the filters on both of the tasks.



Can you identify meaningful learned \_lters? Are the \_lters learned

for POS and for NER di\_erent, or are they similar?

Describe your e\_orts and results in the pdf \_le. For the analysis part, in

addition to answering the questions, describe also your method for identifying