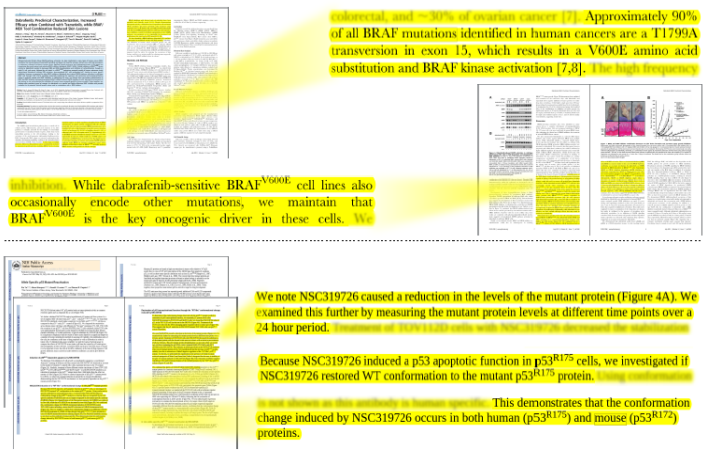
Deep learning for reading and interpreting biomedical papers

L.A. Bugnon, C. Yones, J. Raad, M. Gerard, M.Rubiolo, G.Merino, M. Pividori, L Di Persia, D.H. Milone and G. Stegmayer
Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH/UNL-CONICET, (3000) SF, ARG

Background:
Next-generation sequencing together with novel preclinical reports have led to an increasingly large amount of results published in the scientific literature. However, due to the huge amount of papers available, identifying novel treatments or predicting a drug response in, for example, cancer patients remains a laborious and challenging work. This task requires "reading" a lot of documents for identifying just a small set of papers that have the proper relations between input keywords. There is an urgent need for computational methods that can automatically do this task.

Results:
Our method based on deep learning (DL) is capable of analyzing and interpreting papers in order to automatically extract relevant relations between specific biomedical keywords. It is an end-to-end DL model, trained with full documents and several keyword-pairs with binary labels, indicating whether there is or there is not a relation between them within each full text. Given the input keywords, the model outputs a prediction score for each paper in the corpus, with the probability that the input keywords are related in the text. For training the DL model, a corpus of documents is vectorized using a word embedding. In order to represent a complete text, all the word vectors are concatenated, together with a one-hot-encoding vector for the input entity-type information, which indicates to which of the possible biomedical entities each word belongs (gene, mutation or drug). The embeddings pass through convolutional layers, which compress the word embeddings, and then more convolutional layers grouped in identity blocks (residual and pooling layers) with ELU activations and batch normalization layers. The DL model has been evaluated using a manually curated (labeled) corpus including biomedical entities in oncology, with more than 100 full papers. The results of a 10-fold cross validation showed that our DL model has outperformed state-of-the-art proposals achieving average F1 over 90%. Furthermore, the reliability of the output list of papers was measured, revealing that 100% of the first two documents retrieved for a particular search contain relevant relations. This means that our model can guarantee that the keywords relation can be effectively found in the top-2 papers of the ranked list. Furthermore, our method is capable of highlighting, within each paper, the specific fragments that have the associations of the input keywords (see Figure).

Conclusions:
This proposal could be used for rapidly identifying relationships in full text documents between genes and their mutations, drug responses and treatments in the context of a certain disease. This can certainly be a useful and valuable resource for the advancement of the precision medicine field.



A web-demo is available at: http://sinc.unl.edu.ar/web-demo/dl4papers/