**Challenge**

Can the two languages be distinguished using a bag-of-words approach?

Explain why

No. A bag of words approach will simply count the frequency of each letter and digit in the sequence without considering their ordering. The difference between the 2 language lies in the order of the characters – in the negative examples, the c’s comes before b’s. The bag-of-words approach will not be able to model this difference since in requires knowing the order of elements.

Can the two languages be distinguished using a bigram or trigram-based approach? Explain why

Bigrams or trigrams-based approaches only consider the local order in the sequence. They will only model local adjacency, (e.g. ‘a1’, ‘7b’ or ‘24c’). The task of distinguishing between the 2 languages requires modeling long-range order. To tell that one sequence is positive, the model needs to learn that the sequence of b’s comes before the sequence of c’s, and this will not be possible to do with bigram or trigram models because the b’s and c’s are separated by a block of digits.

Can the two languages be distinguished using a convolutional neural network? Explain why.

Convolution networks, like Ngram-based approaches, consider mostly local order. Assuming small filters are used, and max/average pooling is applied in the network, the model will learn the local motifs in the sequence, but not their absolute positions. Modeling absolute positional information is necessary to learn that the b’s block comes before the c’s blocks, since they are separated by a long buffer