

Alleviating Sequence Information Loss with Data Overlapping and Prime Batch Sizes

Noémien Kocher, Christian Sciuto, Lorenzo Tarantino, Alexandros Lazaridis, Andreas Fischer, Claudiu Musat

 github.com/nkcr/overlap-ml

Hes-so

 swisscom

EPFL

iCoSys
Institute of Complex Systems

1 Introduction

Problem

RNNs and self attention models take as input data points. When modeling a sequence of tokens, those data points are created by extracting sub-sequences of tokens with a predefined length.

While this discretization process is necessary for the machine understanding of the sequence, its side effect is some loss of the token order information. (See figure 1)

Contributions



A first contribution in this work is a mechanism to ensure that all token sequences are taken into account: the *Alleviated Token order Imbalance*.



A second contribution is a strategy for batch creation when using our proposed mechanism, which ends up using a prime batch size.

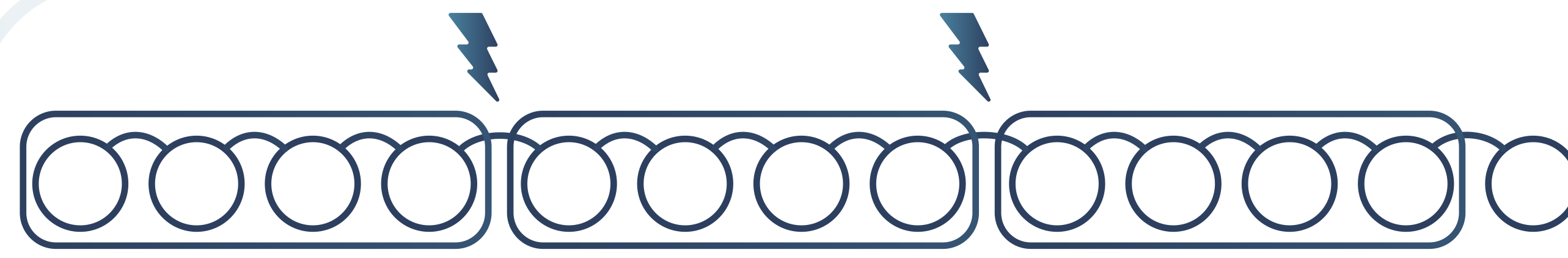





Figure 1

Discretization process on a sequence of 13 contiguous tokens and a predefined length of 4. This process keeps the order of the tokens inside the data points, but loses the order information from token pairs that happen to fall between adjacent data points.

 Contiguous tokens
 Data point
 Order knowledge lost

2 Alleviated Token order Imbalance



To solve an imbalance of token pairs in the data points, we hypothesize that using multiple overlapped sequences of tokens alleviates this effect.

As shown in figure 2, instead of splitting the sequence of token only once, we repeat this process P times using different offsets.

We use the acronym ATOI- P to describe an Alleviated Token Order Imbalance with P overlapped sequences.

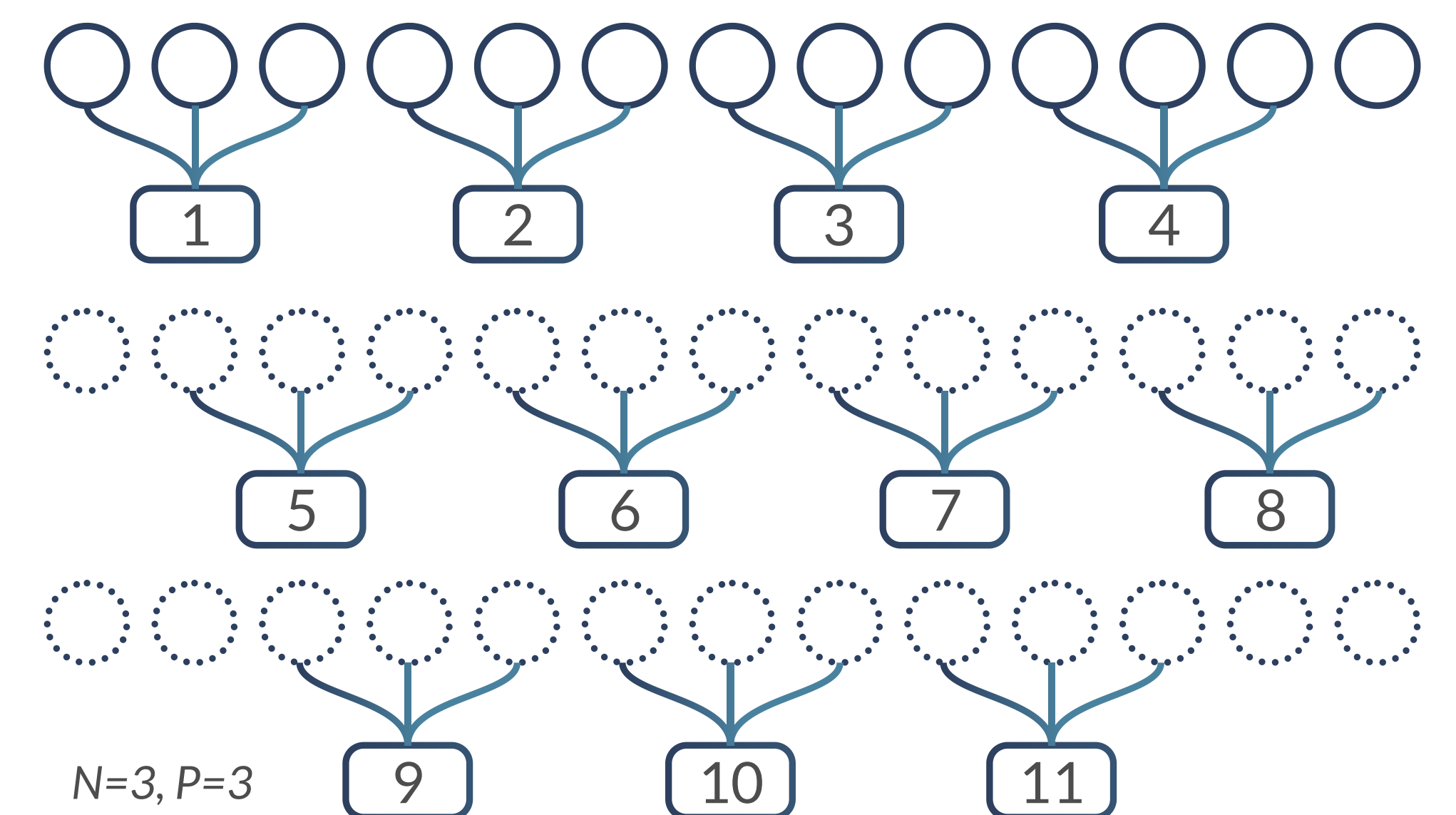


Figure 2

Illustration of our Alleviated Token Order Imbalance mechanism made from a single contiguous list of 13 tokens. With a conventional discretization process that uses $N=3$ tokens per data point, only the data points 1 to 4 would be used, which is illustrated by the first sub-sequence. In this example, we use $P=3$ sub-sequences, which yields a total of 11 data points.

3 Batch creation



We have observed an unintended effect when using our mechanism to build batches for mini-batch training. In this type of training, each batch should contain data points that form a representative distribution of the dataset. In figure 3, we illustrate how the batch size can affect the distribution of batches. Our proposition is simple: use a prime batch size.

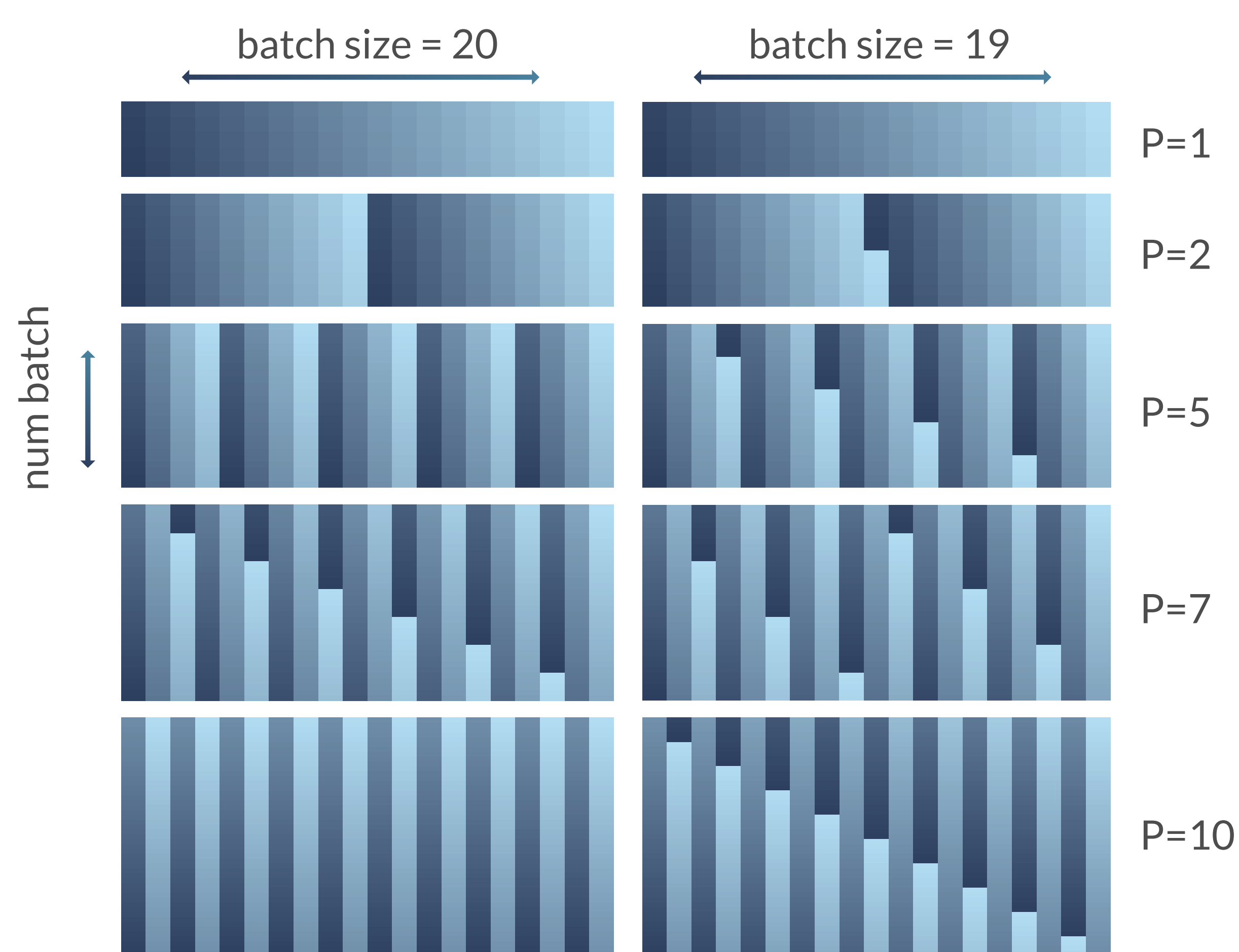


Figure 3

Illustration of the matrix of batches with different P -values and batch sizes. Each data point is a pixel and each row is a batch. The color models the proximity of the data points with respect to the dataset. Two pixels with a similar color represents two data points that are close in the dataset.

4 Results

Our proposed method has been validated in sequence modelling tasks both in text and speech domain outperforming state of the art techniques.

We set up experiments with language modeling for text and emotion recognition for speech, using the same hyperparameters as the baseline models. Here we present some of our results on the text domain.

Table 1 demonstrates how models can be improved by applying our proposed method and Table 2 demonstrates how using a prime batch size with our proposed method actually impacts the scores.

LSTM for language modelling

Model	test ppl
AWD-LSTM (Merity et al., 2017)	58.8
AWD-LSTM + ATOI	56.46
AWD-LSTM-MoS (Yang et al., 2017)	55.97
AWD-LSTM-MoS + ATOI	54.58
Simple-LSTM	75.36
Simple-LSTM + ATOI	74.44

Table 1: Comparison on the PTB dataset between state-of-the-art models and a Simple LSTM, and the same models with Alleviated TOI. The comparison highlights how the addition of Alleviated TOI is able to improve state-of-the-art models, as well as a simple model that does not benefit from extensive hyper-parameter optimization.

Batch size: original vs prime

Experiment	K=20	K=19
ATOI 2	59.37	57.97
ATOI 5	60.50	57.14
ATOI 7	56.70	57.16
ATOI 10	65.88	56.46

Table 2: Perplexity score comparison on the PTB dataset and the AWD model. We use two different values for the batch size K – the original one with $K = 20$, and a prime one with $K = 19$. The results directly corroborate the observation portrayed in Figure 3, where the obtained score is related to the diversity of colors in each row.