Сжатие без потерь с помощью нейронных сетей

Чернышев Кирилл Владиславович

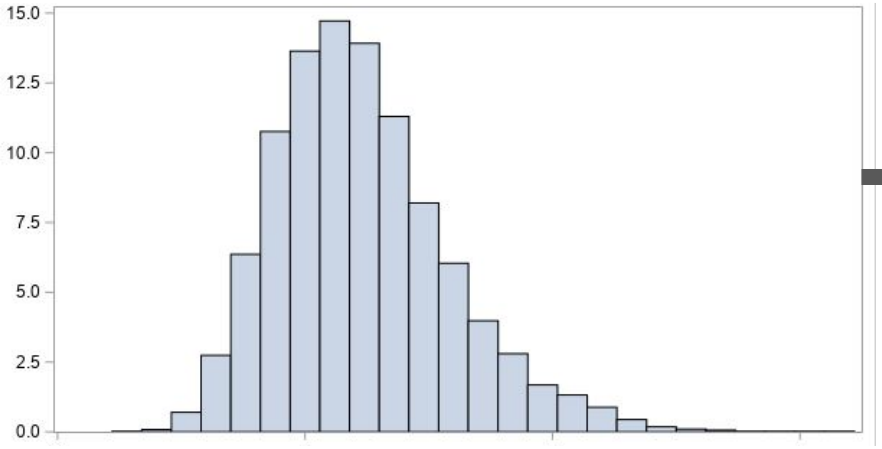
Научный руководитель:
к.ф.-м.н. Кураленок И.Е.

Научный консультант:
Шпильман А. А.

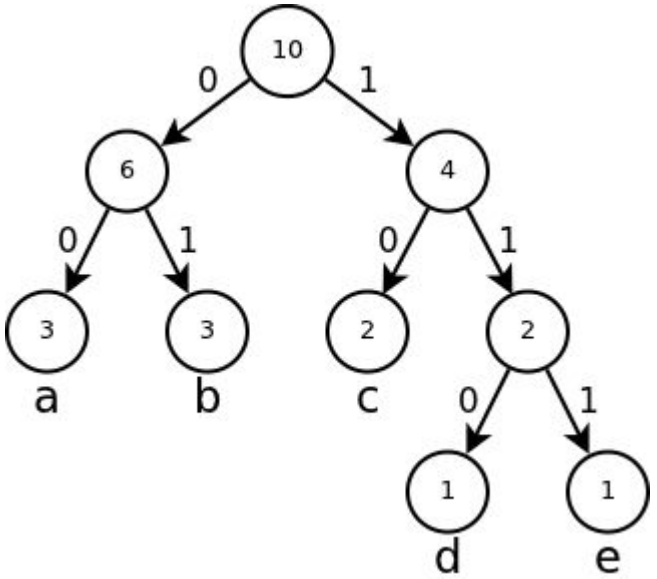
Санкт-Петербургская школа физико-математических и компьютерных наук

Сжатие без потерь

Modeling



Coding



Сжатие без потерь

Генеративные модели для сжатия изображений:

- normalizing flows[1]
- VAE[2]

Сжатие текстов:

- CMIX
- NNCP[3]

[1] E.Hoogeboom et al. -- 2019 *Integer Discrete Flows and Lossless Compresssion*

[2] J.Townsend et al. -- 2019 *Practical Lossless Compression With Latent Variables Using Bits Back Coding*

[3] F.Bellard -- 2021 *NNCP: Lossless Data Compression with Neural Networks*

Цели

Цель работы - использовать нейронные сети для моделирования, выделить особенности такого подхода

Задачи:

- Реализовать алгоритм сжатия
- Произвести сравнение с архиваторами 7zip и zip
- Определить свойства алгоритма

Алгоритм

1. Моделирование условного распределения $p(c|c_{-1}, c_{-2}, \ldots, c_{-n})$ с помощью:
   - MLP
   - LSTM
2. Кодирование -- алгоритм Хаффмана
3. Функция потерь -- усредненная кросс-энтропия $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log \hat{y}_i)$

Сравнение с 7zip

	bib	book1	geo	obj1	gaussian
original	111,3	768,8	102,4	21,5	200
zip	35,4	313,7	68,6	10,5	197,4
7zip	**30,7**	**261,1**	53,4	9,5	144,2
MLP	53,3	421,4	49,9	9,7	143,2
LSTM	44,4	286,5	**48**	**5,4**	**134,4**

Взвешенная кросс-энтропия

веса	LSTM	MLP
обратные частоты	51,9	57,9
относительные частоты	45,9	50,4
отсутствуют	**40**	**45,8**

Уменьшение размера выборки

	MLP	LSTM
Размер обучающей выборки	10,2	2
Число эпох	5	15
Число скрытых нейронов	128	64
Размер сжатого файла	4,68	**4,24**

Выводы

- Реализован алгоритм сжатия, использующий нейронные сети для моделирование распределений символов.
- Обнаружены данные, эффективно сжимаемые представленным алгоритмом.
- Несмотря на аналогии между классификацией и данной задачей, позволяющие сократить число данных обучающей выборки, было показано, что изменение кросс-энтропии на иную функции потерь приводит к серьезной потере качества.