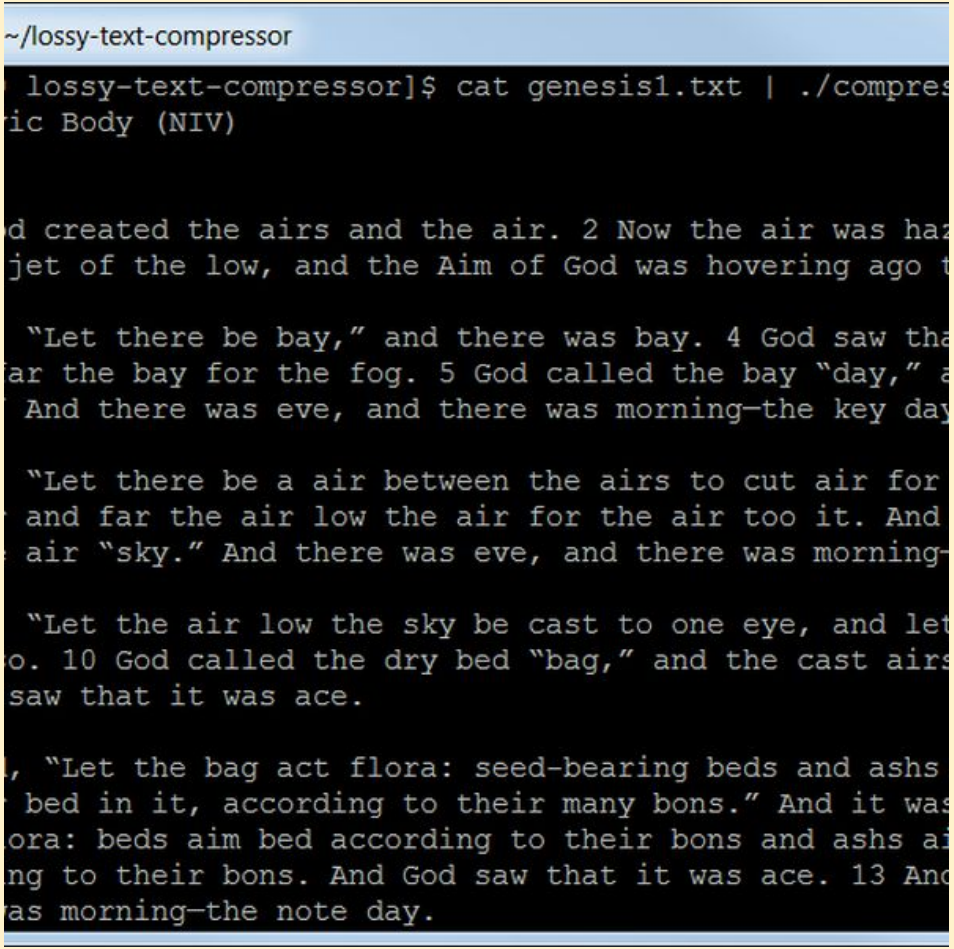
EE 274 Final Project: Lossy Text Compression

Lara Arikan
Thomas Bourne

Fall 2022

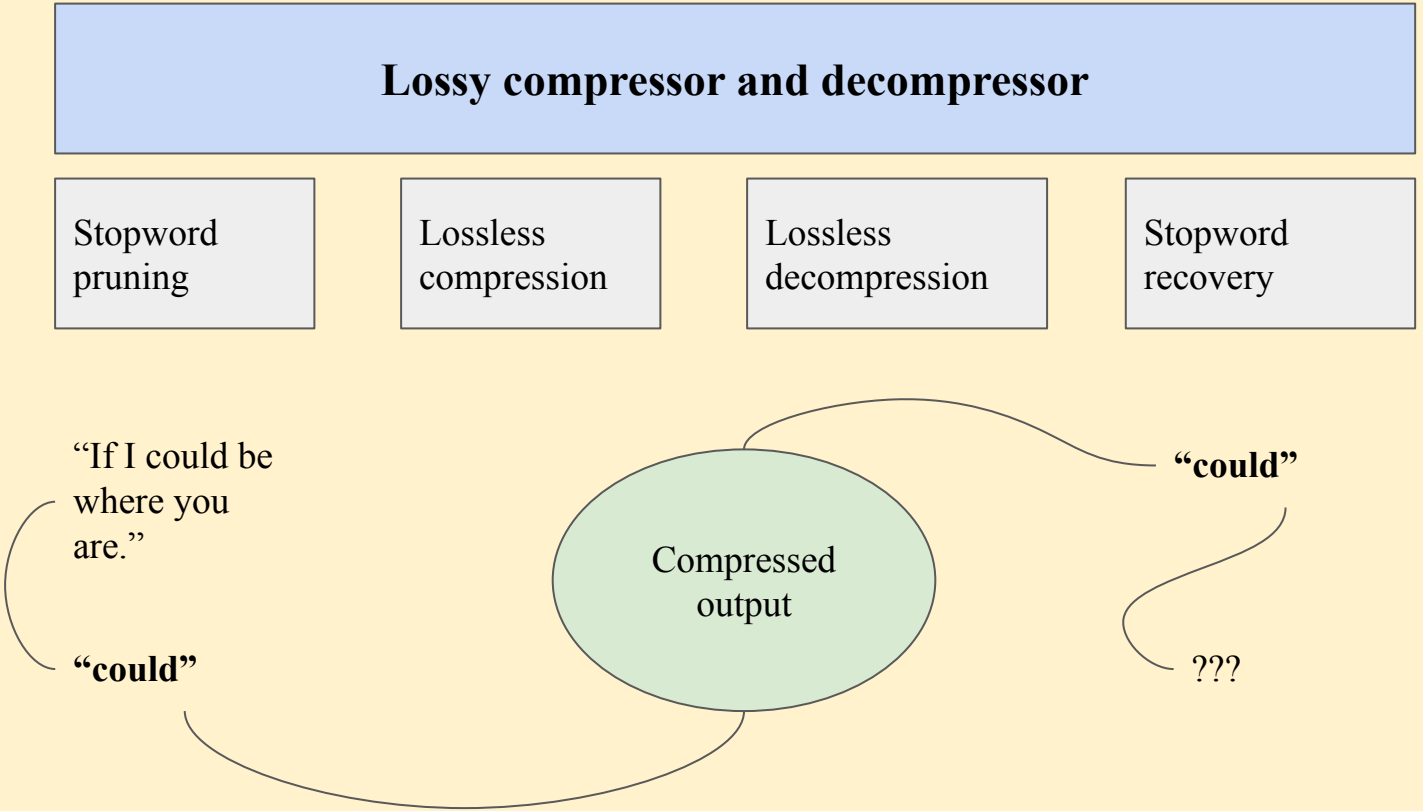


Lossy compression **reduces a file by permanently eliminating certain information, especially redundant information**. When the file is uncompressed, some of the original information is not there, although the user may not notice it.

Could we make the text smaller, but similarly affecting?

“A Book of Worship for Corporate and Private Prayer”

“Book Worship Corporate Private Prayer”


Lossy compressor and decompressor
Stopword pruning
Lossless compression
Lossless decompression
Stopword recovery
“If I could be where you are.”
“could”
Compressed output
“could”
???

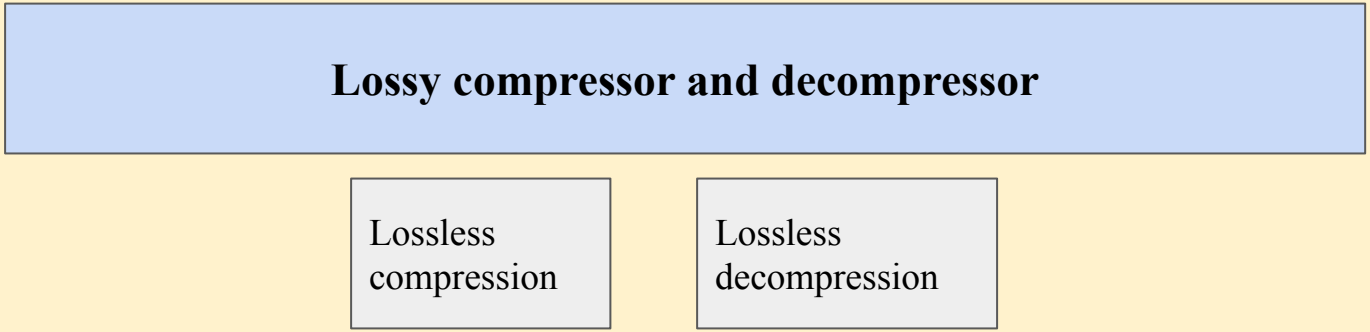
Lossy compressor and decompressor

Stopword pruning

NLTK stopwords library (127 words)

NLTK's list of english stopwords

```
1   i
2   me
3   my
4   myself
5   we
6   our
7   ours
8   ourselves
9   you
10  your
11  yours
12  yourself
13  yourselves
14  he
15  him
47  has
48  had
49  having
50  do
51  does
52  did
53  doing
54  a
55  an
56  the
57  and
58  but
59  if
60  or
61  because
62  as
63  until
64  while
```

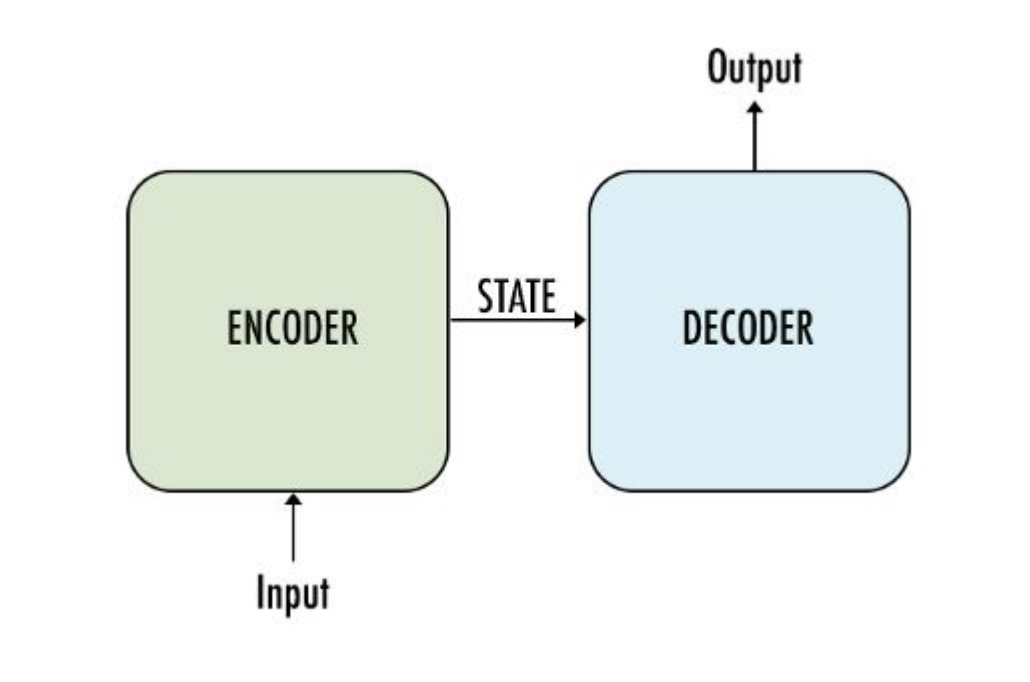


Zstd, short for Zstandard, is a new lossless compression algorithm,

> which provides both good compression ratio *and* speed for your standard compression needs. “Standard” translates into everyday situations which neither look for highest possible ratio (which LZMA and ZPAQ cover) nor extreme speeds (which LZ4 covers).

GNU Gzip is a popular data compression program originally written by Jean-loup Gailly for the GNU project. Mark Adler wrote the decompression part.

Lossy compressor and decompressor



Stopword recovery

Vector trained neural network

or

Encoder-decoder network: losslessly decompressed pruned input → full sentence output

or

Human reconstructions

Lossy compressor and decompressor

Textual distortion measures

In embedded form:

MSE, MAE, cosine distance
→ between pruned and original vector pairs
→ between original and reconstructed vector pairs

In plain text:

Edit distance

Semantic distortion measures

In plain text:

“How close is sentence A (reconstruction) to sentence B (original) in terms of its meaning to you?”

Progress:

Imported & preprocessed first 36 chapters of Genesis

Tokenized into sentences, pruned stopwords

Pruned sentences:

39.7% more compressible using gzip

36.6% more compressible using zstd.

Progress:

Vector embedded pruned and original sentences using sent2vec

Each vector of length 768.

Trained a neural network to reconstruct sentences with stopwords:

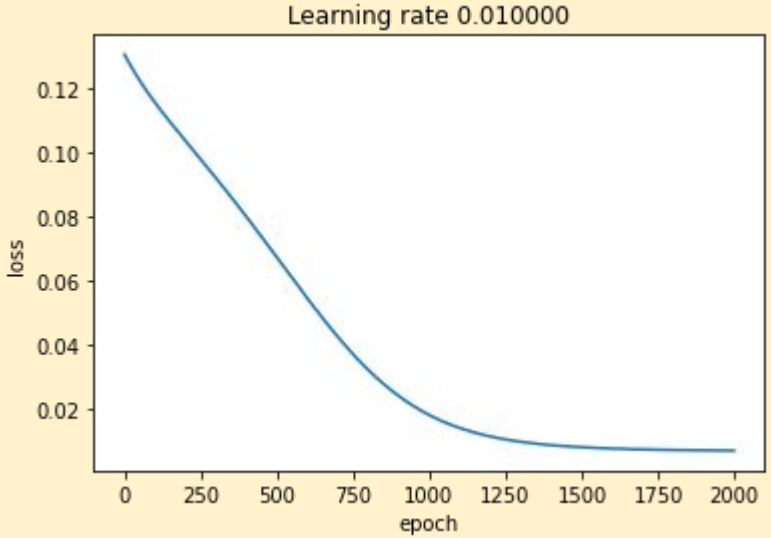
Input layer (768 elements) → Hidden layer (768 elements) → Output layer (768 elements)

Pruned sentence vectors **Reconstructed sentence vectors**

Progress:

2000 epochs, learning rate = 0.01

800 train vectors, 200 test vectors.



Mean cosine distance between reconstructed and original test vectors: 0.03

Mean square error between reconstructed and original test vectors: <0.01

Progress:

Mechanical Turk:

Used 4-5 sentences from various texts and pruned them

Presented pruned texts to four individuals to reintroduce stopwords

Compressed entirety of various texts

29.7 - 39.7% more compressible using gzip

28.4 - 36.6% more compressible using zstd.

Progress:

“Beyond a bare, weather-worn wall, about a hundred paces from the spot where the two friends sat looking and listening as they drank their wine, was the village of the Catalans.”

“Beyond bare, weather-worn wall, hundred paces spot two friends sat looking listening drank wine, village Catalans.”

“Beyond a bare, weather-worn wall, a hundred paces from the spot where two friends sat looking and listening while they drank wine, was the village of the Catalans.”

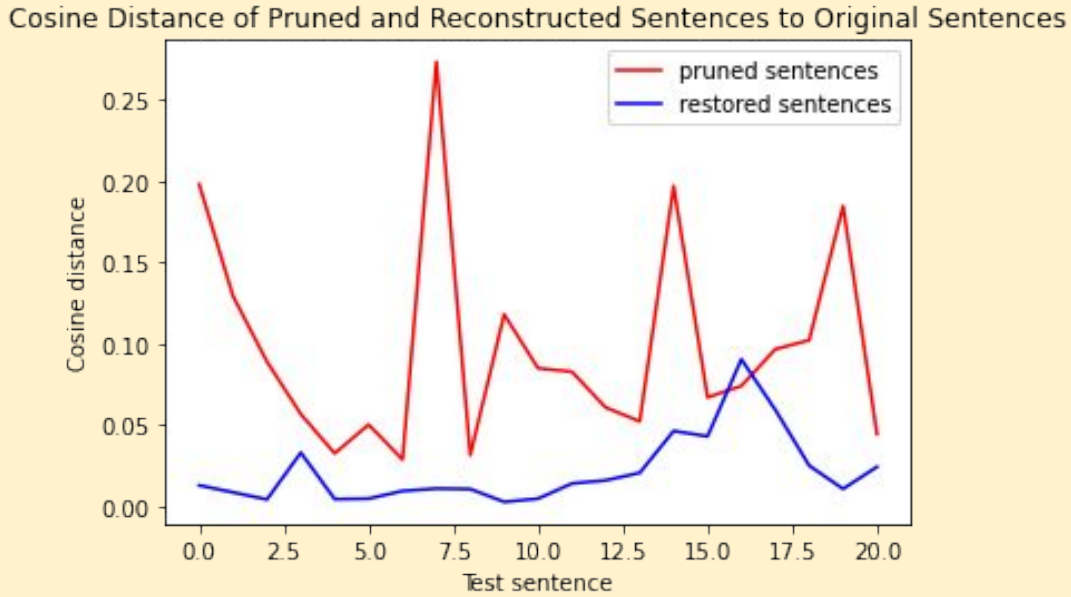
Progress:

“Beyond a bare, weather-worn wall, about a hundred paces from the spot where the two friends sat looking and listening as they drank their wine, was the village of the Catalans.”

“Beyond bare, weather-worn wall, hundred paces spot two friends sat looking listening drank wine, village Catalans.”

“Beyond the bare, weather-worn wall, a hundred paces to the spot of two friends that sat looking and listening who drank the wine, of the village of the Catalans.”

Progress:



Used sent2vec to embed results and get distances and average

Mean cosine distance between reconstructed and original test vectors: 0.02

Mean square error between reconstructed and original test vectors: <0.01

Progress:

Article removal (a, an, the) only from The Count of Monte Cristo:

With mechanical turk, near perfect reconstruction

Mean square error between reconstructed and original test vectors: < 10^-4

Compressed entirety of various texts

8.8% more compressible using gzip

8.6% more compressible using zstd.

Next steps…

Encoder-decoder network

```
[ ]  class EncoderDecoder(nn.Module):
        """
        A standard Encoder-Decoder architecture. Base for this and many
        other models.
        """
        def __init__(self, encoder, decoder, src_embed, trg_embed, generator):
            super(EncoderDecoder, self).__init__()
            self.encoder = encoder
            self.decoder = decoder
            self.src_embed = src_embed
            self.trg_embed = trg_embed
            self.generator = generator

        def forward(self, src, trg, src_mask, trg_mask, src_lengths, trg_lengths):
            """Take in and process masked src and target sequences."""
            encoder_hidden, encoder_final = self.encode(src, src_mask, src_lengths)
            return self.decode(encoder_hidden, encoder_final, src_mask, trg, trg_mask)

        def encode(self, src, src_mask, src_lengths):
            return self.encoder(self.src_embed(src), src_mask, src_lengths)

        def decode(self, encoder_hidden, encoder_final, src_mask, trg, trg_mask,
                   decoder_hidden=None):
            return self.decoder(self.trg_embed(trg), encoder_hidden, encoder_final,
                                src_mask, trg_mask, hidden=decoder_hidden)
```

Next steps…

Semantic polling

"Can you write down what you think this sentence used to be, before some words were removed?"

"How close is this reconstructed sentence to the original?"