Training models in Deep Learning: A Formal basis

Samir El Karrat Moreno

November 28, 2025

Contents

1 Introduction **2**

2 Input data **2**

3 Tokenization **2**
3.1 Byte Pair Encoding (BPE) 2

4 Embeddings **3**

5 Datasets & Data Loading using pytorch **3**

6 Deep neural networks: mathematical formulation **3**
6.1 Bias & Affine transformation 3
6.2 Activation functions 3
6.3 Logits & Softmax 3

7 Types of neural networks **3**

8 Training **3**
8.1 Mathematical motivation (via Statistics) 3
8.1.1 Maximum Likelihood Estimation 3
8.2 Loss functions 4
8.2.1 Cross-Entropy Loss 4
8.3 Regularization 5
8.4 Optimization methods 5

9 Extra material **6**
9.1 BPE 6

1 Introduction

2 Input data

3 Tokenization

In DL we want to train our models by finding the minima of a chosen loss function. Let us denote $x^{(i)} \in \mathcal{D}_{\text{input}}$ be our input data $i = 1, \dots, n$ in the input space $\mathcal{D}_{\text{input}}$ for, n being the amount of data that we are training our model with. Often times our input data $x^{(i)}$ can be highly dimensional, possibly reducing the context window of our models (see **??**). For example, CNNs work directly on the pixel tensor representing an image. However, for when having text as our input data several methods of compression have been developed, as we are going to describe in the following discussion.

In the case that we want to represent our data more condensedly, it is popular to take the following steps:

(1) **Tokenization**: It is a 'statistical' model for encoding input data following certain fixed rules and frequency optimizations to create condensed representations called **token sequences**. The latter consist of strings of **tokens**, statistically optimized encoded subdivisions of our original data. This process is bi-directional, and given any token sequence we can recover its input space representation.

During the training of a tokenizer, we look for efficient mappings between tokens in order to compress token sequences further. We denote as the **vocabulary** this collection of all mappings. It makes up for the entirety of the learnt parameters. However the **vocabulary dimension** is a hyper-parameter.

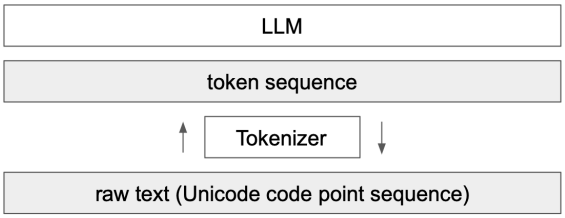

Figure 1: How tokenizer works

3.1 Byte Pair Encoding (BPE)

Byte Pair Encoding, or shortly BPE, is a basic example of text tokenization method. During its training stage, we first find the most frequent bigram (tuple of subsequent tokens) in the corresponding token sequence of our training text. Then we add it to our vocabulary, creating in this way a new **merge**. After this we replace all its occurrences with the correspnding mapping according to the vocabulary, i.e, its index. We perform the whole process iteratively, adding new merges until the desired vocabulary size is achieved. Note that merges containing tokens previously added to the vocabulary are a possiblity.

During BPE's inference stage, we encode/decode any token sequence from an inference text (the text we want to tokenize). For encoding, we simply just loop through each merge in our vocabulary, hierarchically from first merges to last merges, replacing any ocurrence of it with its assigned token. For decoding, we perform the same loop but in the opposite order, from last performed merges to the first ones, replacing them for their corresponding bigram according to the vocabulary.

Usually in BPE we do not work with a raw string of text. We usually format them in a general way that allows us to uniformly treat basically any string of text we can ever encounter. This can be done in various forms, the most common one being formatting to bytes using an UTF-8 encoder. We would then encode the most common byte pairs, and initialize our vocabulary size to be 256. See 1 for a pseudocode and subsection 9.1 for a Python implementation.

However, there exists other approaches as

The process described before is used to *train* the model. We can use any form of text as input data. However, as we tipically want a versatile tokenizer for LLMs, it is beneficial that the training data correctly represents expected inputs during evaluation. For this reason we must make sure that training

Algorithm 1 Byte Pair Encoding using UTF-8 (BPE) – Training

Require: Input text sequence T, vocabulary size V_{target}
Ensure: Learned merge rules M, Final token sequence S

Convert T into a sequence of UTF-8 bytes: $S \leftarrow [b_1, b_2, \ldots, b_n]$ where $b_i \in [0, 255]$
Initialize vocabulary size $V \leftarrow 256$
Initialize merge rules $M \leftarrow \emptyset$
while $V < V_{target}$ **do**
    $C \leftarrow \text{CountPairs}(S)$ ▷ Compute frequency of adjacent pairs
    if C is empty **then**
        break
    end if
    $(p_1, p_2) \leftarrow \arg\max_{(x,y)\in C} C[x, y]$ ▷ Find most frequent pair
    $idx \leftarrow V$
    $M[(p_1, p_2)] \leftarrow idx$ ▷ Record the new merge rule
    $S \leftarrow \text{Merge}(S, (p_1, p_2), idx)$ ▷ Replace all occurrences in sequence
    $V \leftarrow V + 1$
end while
return M, S

input texts contain a combination of code and different languages, making a 'uniform vocabulary density' without over-specializing in any domain.

Let us now look at a python implementation of this algorithm:

BPE can run on code points, so change defintion from utf-8 to So far only seen training definition and we still are left with code/decode. After this talk about regex motivated by token issues. Continue with special tokens (useful for meta data, used in finetuning like adapting to browser)

Difference between token and index (after tokenization) Nn embeddings, relationship with a transformer, what are channels in a transformer, forward/ backward pass transformer, training by distillation, finetuning

4 Embeddings

5 Datasets & Data Loading using pytorch

6 Deep neural networks: mathematical formulation

6.1 Bias & Affine transformation

6.2 Activation functions

6.3 Logits & Softmax

7 Types of neural networks

8 Training

8.1 Mathematical motivation (via Statistics)

8.1.1 Maximum Likelihood Estimation

The **Maximum Likelihood Estimation** is an *statistical approach* to the following problem: given some IID (independent and identically distributed) from which we know what distribution they follow but not its parameters, how can we *estimate* them.

MLE solves this approach by maximizing the so called **parameter likelihood**. In mathematical terms, let $p(x^{(1)}, \ldots, x^{(n)}|\theta)$ be the joint density function for our IID data samples $x^{(1)}, \ldots, x^{(n)}$ and some parameters θ. Then this translates to finding $\arg\max_\theta p(x^{(1)}, \ldots, x^{(n)}|\theta) = \prod_i^n p(x^{(i)}|\theta)$. Note how we are considering our function fixed on the data points as we are trying to find the "best fit" for them.

Computationally, to avoid numerical instability and to be able to conduct optimiziation, we typically aim to minimize quantities. We then can transform our problem: we can take $\text{NLL}(\theta) = -\log p(x^1, \ldots, x^n|\theta)$. Finding the maximum parameter likelihood is then equivallent (due to logarithmic injectivity) to finding $\arg\min_\theta \text{NLL}(\theta)$.

Observation: This statistical concept is closely tied to typical the optimization process in DL, for which we use a loss function to update the weights of our models. The NLL is a bridge between statistics and DL.

8.2 Loss functions

8.2.1 Cross-Entropy Loss

Before we get into detail, let us borrow a very important concept from statistics that help us understand better what cross-entropy is:

Kullback-Leibler Divergence

The **Kullback-Leibler Divergence** $D_{KL}(P||Q)$ is a measure of the information lost when Q is used to approximate P. In more of a mathematical approach we say it is the "distance" between both probability distributions. The formula for discrete distributions is:

$$D_{KL}(P||Q) = \sum_{i \in \mathcal{X}} P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

Let us motivate it through a concrete example. Given the two discrete probability distributions on the sample space $\mathcal{X} = \{0, 1\}$:

- **True Distribution P (Fair Coin X):** $P(X = 0) = P(0) = p_1$ $P(X = 1) = P(1) = q_1$
- **Approximating Distribution Q (Biased Coin Y):** $P(Y = 0) = Q(0) = p_2$ $P(Y = 1) = Q(1) = q_2$

We want to compare for a same event the probability of it happening. This in general cannot be done as they need not share the same sample space, but none the less it is a good way to visualize it.

Consider that we flip the coin N times. Then our sample space is completely determined by strings of this same length of the type $101\ldots1$. For any given event $e \in \Omega$, we can define the ratio:

$$\frac{P(e)}{Q(e)} = \frac{p_1^n q_1^m}{p_2^n q_2^m}$$

Now we can take the log and divide by N:

$$\begin{aligned}
\frac{1}{N}\log\left(\frac{P(e)}{Q(e)}\right) &= \frac{1}{N}\log\left(\frac{p_1^n q_1^m}{p_2^n q_2^m}\right) \\
&= \frac{1}{N}\left[\log(p_1^n) + \log(q_1^m) - \log(p_2^n) - \log(q_2^m)\right] \\
&= \frac{1}{N}\left[n\log(p_1) + m\log(q_1) - n\log(p_2) - m\log(q_2)\right] \\
&= \frac{n}{N}\left[\log(p_1) - \log(p_2)\right] + \frac{m}{N}\left[\log(q_1) - \log(q_2)\right] \\
&= \frac{n}{N}\log\left(\frac{p_1}{p_2}\right) + \frac{m}{N}\log\left(\frac{q_1}{q_2}\right)
\end{aligned}$$

In the limit as $N \to \infty$, by the Law of Large Numbers, the frequency $\frac{n}{N}$ converges in probability to the true probability $P(1) = p_1$, and $\frac{m}{N}$ converges to $P(0) = q_1$.

Thus, taking the expectation over the sequence of N events, $\mathbb{E}_P\left[\frac{1}{N}\log\left(\frac{P(e)}{Q(e)}\right)\right]$, we find that the limit of this average log-ratio is the Kullback-Leibler Divergence:

$$\lim_{N\to\infty} \mathbb{E}_P\left[\frac{1}{N}\log\left(\frac{P(e)}{Q(e)}\right)\right] = p_1\log\left(\frac{p_1}{p_2}\right) + q_1\log\left(\frac{q_1}{q_2}\right)$$

This is the general formula for the KL-Divergence between two Bernoulli distributions:

$$D_{KL}(P||Q) = \sum_{i \in \{0,1\}} P(i)\log\left(\frac{P(i)}{Q(i)}\right)$$

where $P(1) = p_1, P(0) = q_1$, $Q(1) = p_2, Q(0) = q_2$.

8.3 Regularization

8.4 Optimization methods

9 Extra material

9.1 BPE

Here I include pseudocode for some of the functions in Algorithm 1, together with a python implementation of it.

Algorithm 2 Further functions for BPE Training

procedure CountPairs(S)
  $Counts \leftarrow$ empty map
  for $i \leftarrow 0$ **to** $|S| - 2$ **do**
    $pair \leftarrow (S[i], S[i+1])$
    $Counts[pair] \leftarrow Counts[pair] + 1$
  end for
  return $Counts$
end procedure
procedure Merge($S, pair, new\_token$)
  $S_{new} \leftarrow []$
  $i \leftarrow 0$
  while $i < |S|$ **do**
    if $i < |S| - 1$ **and** $S[i] == pair[0]$ **and** $S[i+1] == pair[1]$ **then**
      Append new_token to S_{new}
      $i \leftarrow i + 2$
    else
      Append $S[i]$ to S_{new}
      $i \leftarrow i + 1$
    end if
  end while
  return S_{new}
end procedure

```python
tokens = text.encode("utf-8") # raw bytes
tokens = list(map(int, tokens)) # convert to a list of integers in range 0..255 for convenience

def get_stats(ids): # Get the most frequent pair in the byte-encoded text
    counts = {}
    for pair in zip(ids, ids[1:]):
        counts[pair] = counts.get(pair, 0) + 1
    return counts # Dictionary of the form "pair -> times_appeared"

def merge(ids, pair, idx): # Update the token sequence according to a new merge
    newids = []
    i = 0
    while i < len(ids):
        if i < len(ids) - 1 and ids[i] == pair[0] and ids[i+1] == pair[1]:
            newids.append(idx)
            i += 2
        else:
            newids.append(ids[i])
            i += 1
    return newids

# ---
vocab_size = 276 # the desired final vocabulary size
num_merges = vocab_size - 256
ids = list(tokens) # copy so we don't destroy the original list

merges = {} # (int, int) -> int
for i in range(num_merges):
    stats = get_stats(ids)
    pair = max(stats, key=stats.get)
    idx = 256 + i
    print(f"merging {pair} into a new token {idx}")
    ids = merge(ids, pair, idx)
```

```
38      merges[pair] = idx
```

Listing 1: An implementation of BPE in Python (adapted code from Andrej Karpathy's github)