Lab class

Computational Semantics 2021

Adam Ek

The vocabulary

- A torchtext `Field` is
  1. a procedure of preprocessing steps
  2. a mapping of words to indices (numbers)
- Thus, we don't want a context `Field` and a center `Field`, but a shared Field.
- Since in fact, we only have one vocabulary

Projection function

```
TOKENS = bla bla bla

# read the csv files
train, test = TabularDataset.splits(path   = ddir,
                                    train  = 'train1.csv',
                                    test   = 'test1.csv',
                                    format = 'csv',
                                    fields = [('column1', TOKENS),
                                              ('column2', TOKENS)],
                                    skip_header       = True,
                                    csv_reader_params = {'delimiter':'\t',
                                                         'quotechar':'½'})
```

What is an embedding?

- An embedding is:

$$e_{cat} = \left[x_1 x_2 \ldots x_m\right] \tag{1}$$

- the individual numbers have *no meaning*
- embeddings obtain meaning through linear and non-linear transformations
- embeddings obtain meaning from what we can *predict* using them
- embeddings obtain meaning from their position in *space*

Model parameters

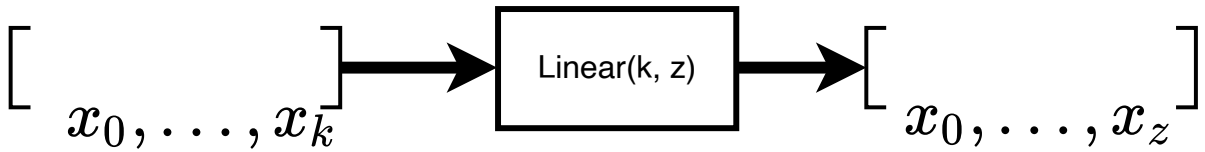
- An embedding layer

$$\begin{bmatrix} x_0^0, \ldots, x_k^0 \\ \vdots \\ \vdots \\ x_0^V, \ldots, x_k^V \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \left[x_0^2, \ldots, x_k^2\right]$$

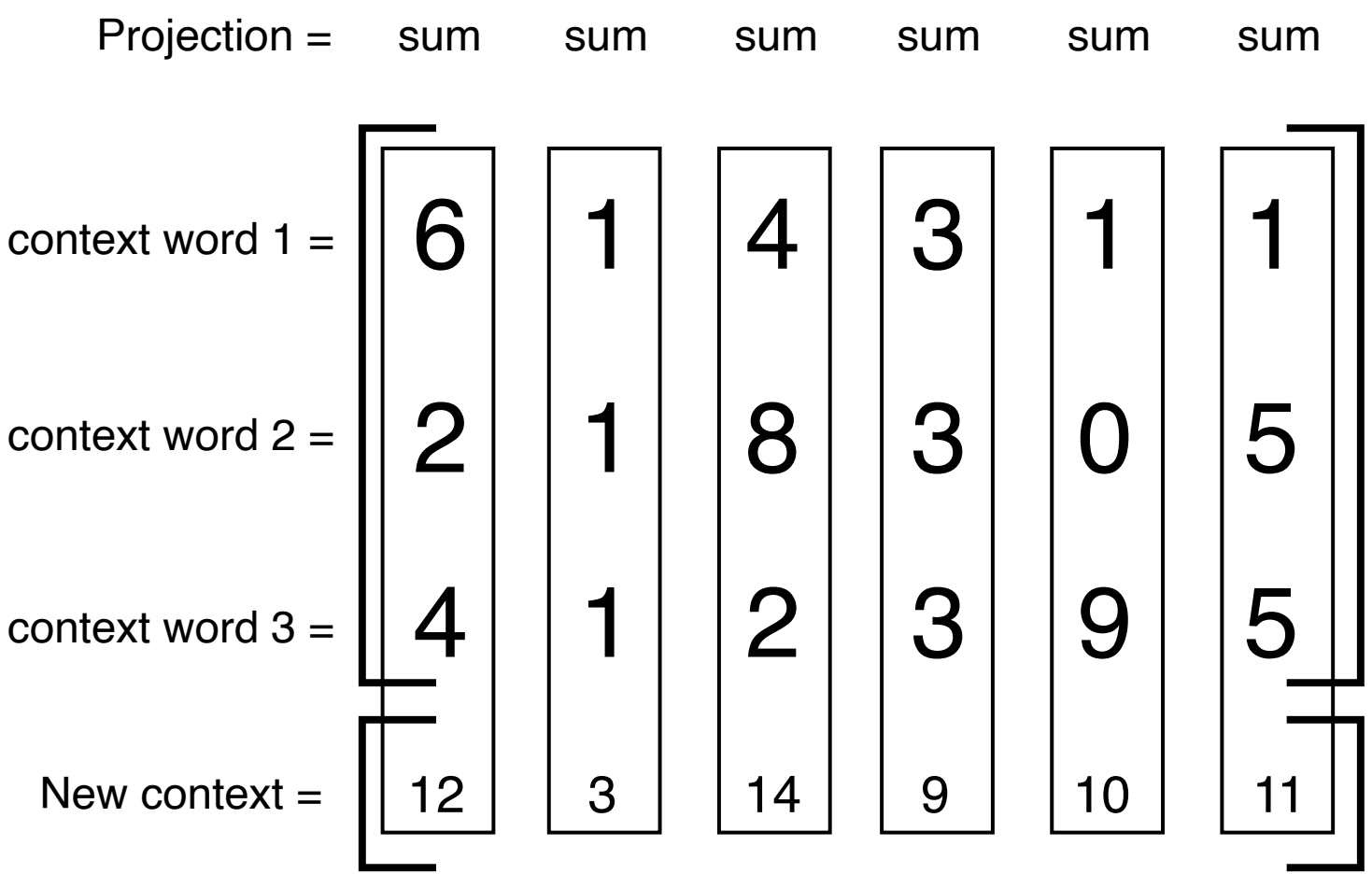
embedding layer one-hot word embedding

- A prediction layer

Where z is the number of words in our vocabulary, and k the size of our vector

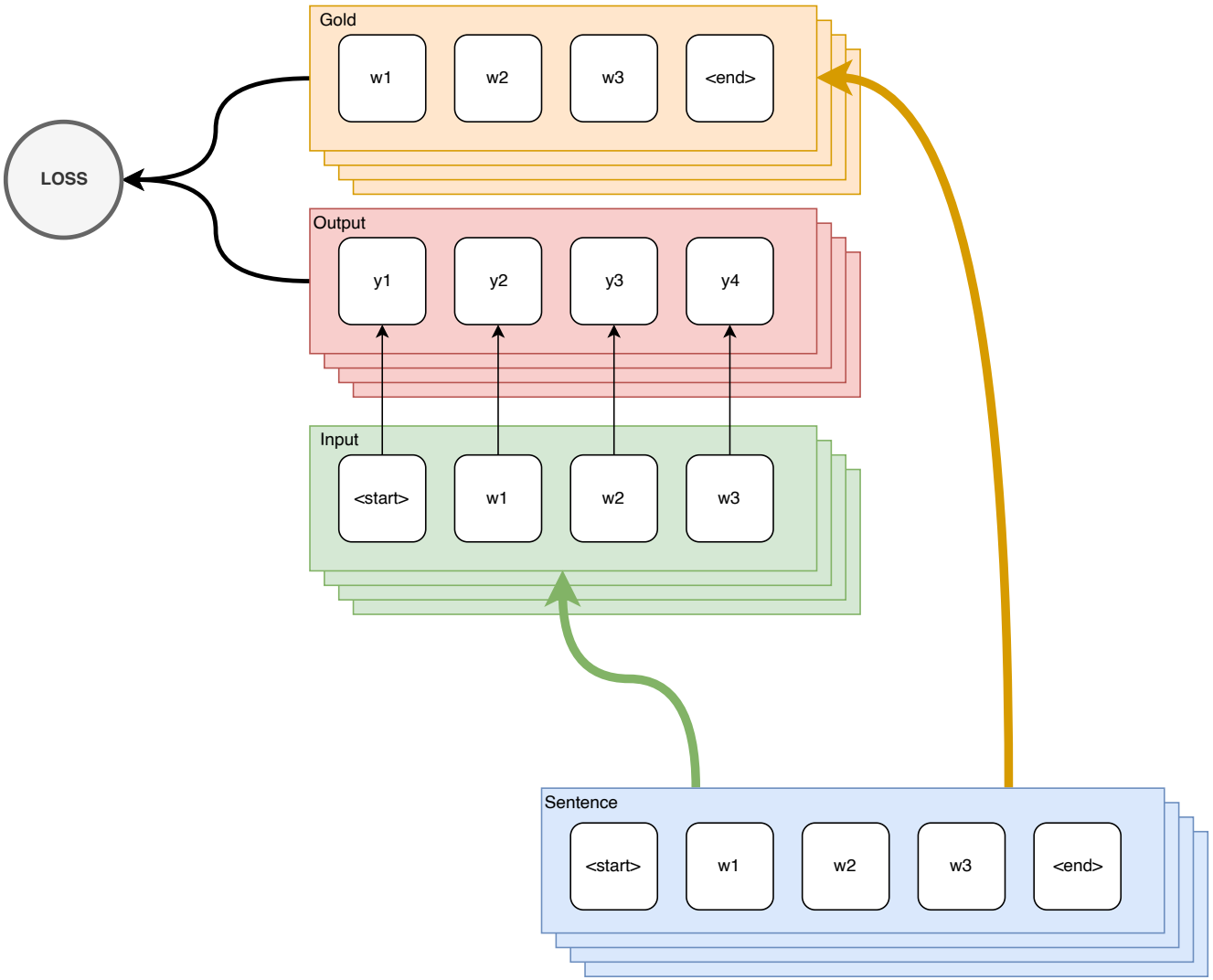


Projection function



Cross Entropy Loss

Language modeling



Open questions