Notation

The following notation is used throughout this work.

a	a scalar (real, integer, or word/token)
$\tilde{a}$	a vector, nominally a column vector
A	a matrix
$\mathcal{A}$	a sequence, including a dataset or a sequence of words
$\mathbb{V}$	a set, e.g. the vocabulary
$\tilde{x}_{[i]}$	the ith element of the vector $\tilde{x}$
$X_{[i,j]}$	the row i and column j element X
$X_{[:,i]}$	the ith column *vector* of the matrix X
$X_{[i,:]}$	the ith row *vector* of the matrix X
$w^{[t]}$	a scalar tth element of some sequence
W^{f}	a matrix disambiguated by the name f
$[A\ B]$	the horizontal concatenation of A and B
$[A; B]$	the vertical concatenation of A and B
$P(\ldots)$	A probability (estimated or ground truth)
A	a random variable (when not a matrix)
$\hat{y}$	A network output value, corresponding to target value y or $\tilde{y}$ a vector or scalar quantity as appropriate

Words are treated as integers
We consistently notate words, as if they were scalar integer values. writing for example $w^{[1]}$ as to be the first word in a sequence. Which is then used an an index: $C_{[:,w^{[i]}]}$ is it's corresponding word vector, from the embedding matrix C.

Superscripts and Subscripts
Readers may wonder why we are using $x_{[i]}$, and $x^{[i]}$. Would not $x_{[i]}$ suffice? Why differentiate between elements of a sequence, and elements of a vector?
The particularly problematic case, is that we often want to represent taking the ith element of a vector that is the tth element of a sequence of vectors. The vector, we would call $\tilde{x}^{[t]}$, its i element is $\tilde{x}^{[t]}_{[i]}$.
This is also not ambiguous with the matrix indexing notation $X_{[i,t]}$.
Rarely a superscript will be actually an exponent, e.g. $x^{\frac{2}{3}}$. This should be apparent when in this case. For more common is the natural exponent which we write $\exp(x)$.