CS-583: Deep Learning
Beam Search

Abdul Rafae Khan
Department of Computer Science
Stevens Institute of Technology
akhan4@stevens.edu

October 20, 2023

Beam Search

Lets say we have a machine translation model already trained
Translation output is one word at a time
We want to use it to output the best translation for any given input sentence

Beam Search

Given the input sentence (french) and the previous output word (English), the model outputs the next word

Beam Search

If we have V words in the vocabulary and maximum sentence length is L
Exhaustive search will be V^L
If $vocab = 1000$ and $max\_length = 10$, then 10^{30} possible choices to select from

Beam Search

A simpler option is to select the best at each position

Beam Search

Typically greedy selection for the word at each position

Position	1	2	3	4
A	0.5	0.1	0.2	0.1
B	0.2	**0.4**	0.2	0.2
C	0.2	0.3	0.4	0.1
<END>	0.1	0.2	0.1	0.6

Output probability $= 0.5 \times 0.4 \times 0.4 \times 0.6 = 0.048$

Beam Search

It can happen that this is not the optimal
Selecting 2*nd* best at position 2 gives better total probability

Position	1	2	3	4
A	0.5	0.1	0.1	0.1
B	0.2	0.4	0.6	0.2
C	0.2	**0.3**	0.2	0.1
<END>	0.1	0.2	0.1	0.6

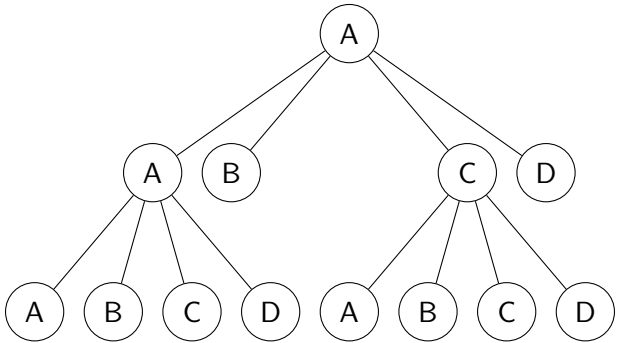
Output probability $= 0.5 \times 0.3 \times 0.6 \times 0.6 = 0.054$

Beam Search

1*st* step:
Select the K maximum probability words

For each subsequent step:
Fix the previous selections and generate K possibilities and select the overall K maximum

Beam Search



Whiteboard