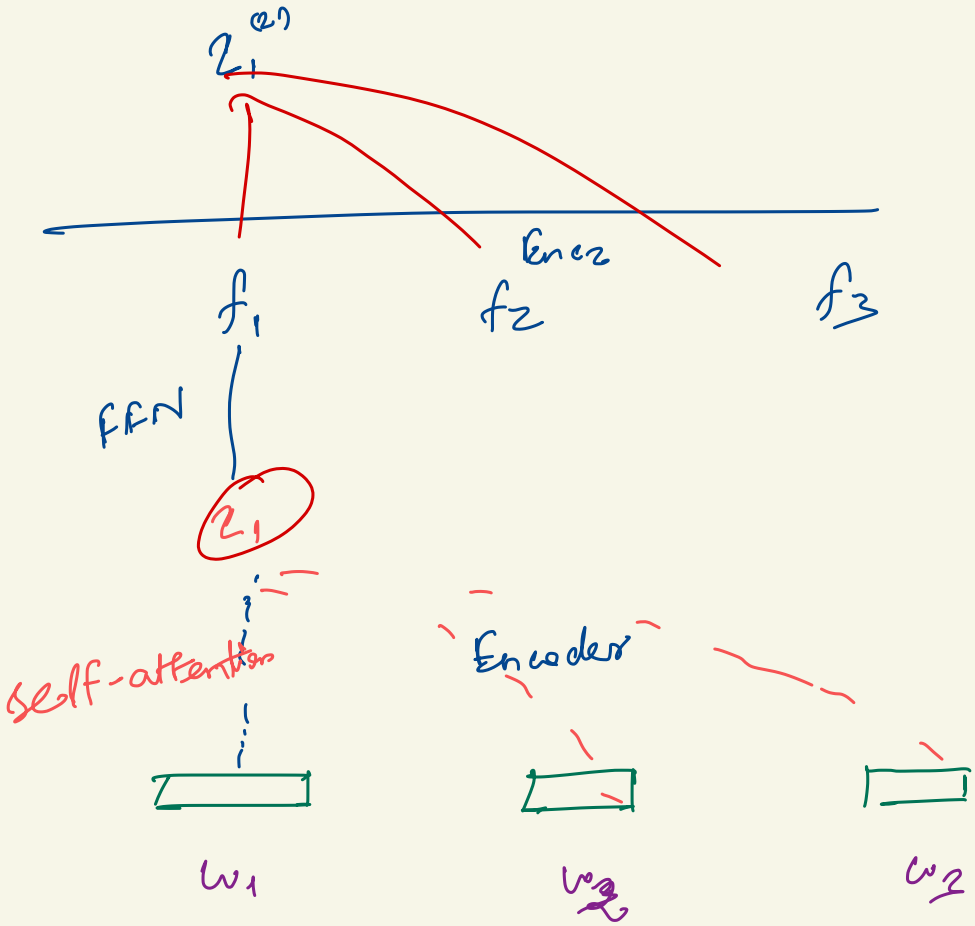
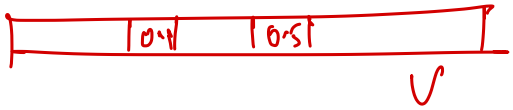

z_1
Encz
f_1
f_2
f_3
FFN
z_1
self-attention
Encoder
w_1
w_2
w_2

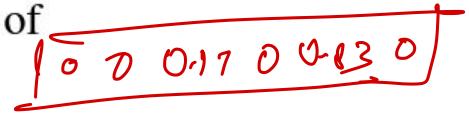
Top-k Sampling

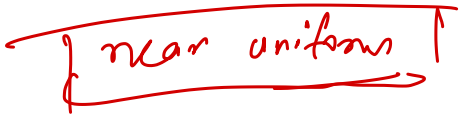


In Top-k Sampling, only Top k tokens (as per prob.) are considered for generation, so the less probable words would not have any chance

1. Choose in advance a number of words k
2. For each word in the vocabulary V, use the language model to compute the likelihood of this word given the context $p(w_t|\mathbf{w}_{<t})$
3. Sort the words by their likelihood, and throw away any word that is not one of the top k most probable words.
4. Renormalize the scores of the k words to be a legitimate probability distribution.
5. Randomly sample a word from within these remaining k most-probable words according to its probability.

renormalize





Nucleus Sampling or Top-p sampling

Issues with Top-k Sampling

Shape of the probability distribution differs in different contexts. Top-k may include most of the probability mass in some cases, and very small mass in other cases.

Nucleus Sampling or top-p sampling

Keep not the top k words but top p percent of the probability mass

Given a distribution $P(w_t|w_{<t})$, top-p vocabulary $V^{(p)}$ is the smallest set of words such that

$$\sum_{w \in V^{(p)}} P(w|w_{<t}) \geq p$$

Try this problem

Suppose you have a vocabulary of size 5 and during decoding, the output vector is [3, -1, 2, 1, -2]. Write down the effective probability distribution when you use the following sampling strategies.

- Random sampling with temperature 0.5
- Top-2 sampling
- Nucleus sampling with p = 0.5