

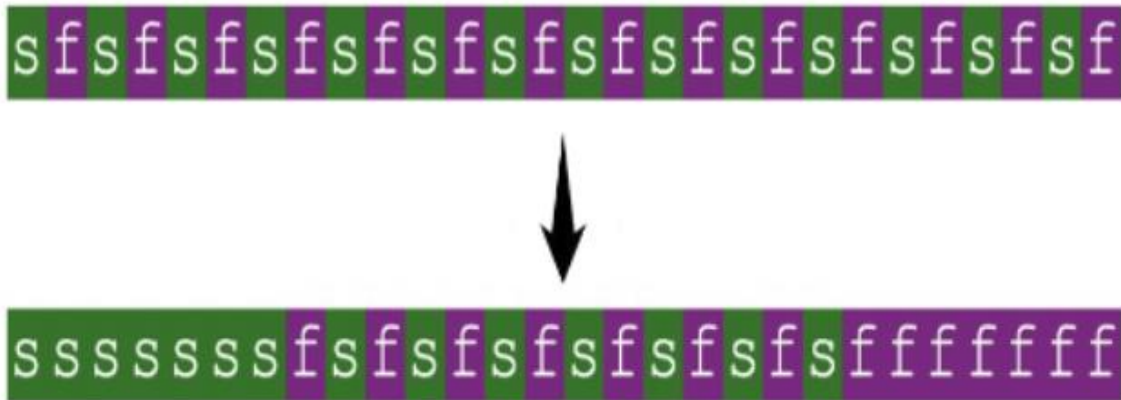
HW4 Gradescope

B08901051 電機四 王維紳

1. Make brief introduction about one of the variant of Transformer, and use an image of the structure of the model to help explain.

Ans:

One of the variant of Transformer is **Sandwich Transformer**. The original transformer has one self-attention layer corresponding to one feedforward layer in each level where as **Sandwich Transformer** has more self-attention layer distributed toward the bottom of the whole model and more feedforward layer toward the top.



2. Briefly explain what's the advantages of this variant under certain situations.

Ans:

The **Sandwich Transformer** doesn't increase any extra parameters and achieve better performance on word-level language modeling benchmark and character-level language modeling.

Reference [1]: <https://www.youtube.com/watch?v=lluMBz5AoOg>

Reference [2]: <https://aclanthology.org/2020.acl-main.270.pdf>