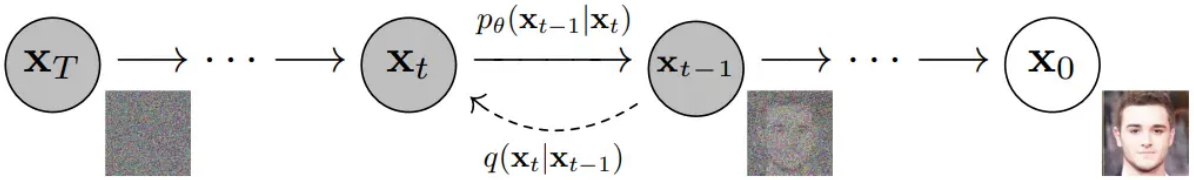
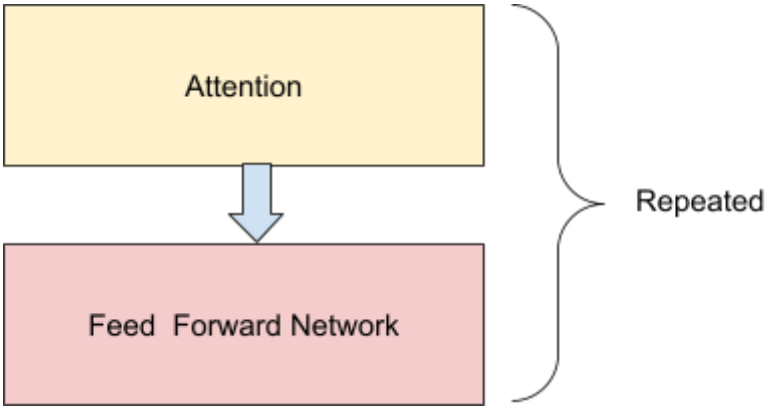
PROJECT WORK IN ARCHITECTURES AND PLATFORMS FOR ARTIFICIAL INTELLIGENCE

This project would like to examine the architecture of current transformers to improve their efficiency for possible deployment on a microcontroller unit. Currently, some research are done on Diffusion model that tries to reduce the number of iterations applied on input noise to obtain a final image



I would like to apply a similar technique to transformers. Transformers are nothing more than a stack of attention blocks layered on top of each other.



What we would like to discover is if it is possible to compress a whole transformer in a single attention layer: In your project work you will :

- Familiarize with the hugging face transformer library (https://huggingface.co/docs/transformers/index)
- Extrapolate the attention maps and the features of a transformer architecture, possible idea the LLama (https://arxiv.org/pdf/2302.13971) model, analyze their changes over the depth of the model
- Verify what happens if:
    - we try avoid the computation of some attention blocks
    - we selected a single attention block and we apply in succession.
    - (optional) testing some possible extension with kV cache compression