Reaction paragraph for Lecture on Transformers.

I noticed 3 themes from the lecture. **First**, the comparison between RNN v/s CNN v/s Transformer models. RNNs are the model of choice for variable length representations but are inherently slow. CNN models can be parallelized are hence faster. Transformer models are the best of both worlds. Transformer models are fast and leverage attention for representation. **Second,** the lecture went on to describe the main components of the Transformer model. there are 3 main components I noticed: Positional Encoding, Multi-Head Attention and use of Residuals. Multi-head attention uses parallel attention layers which capture different linear transformations on the input/output. Residuals are important in the model because it helps carry positional information to higher layers. Unfortunately, I did not much about the use of positional encoding in the model architecture. **Third**, and lastly the speakers went on to give a preview of how Transformer models are used for Image and Music Generation. This gave me some insights on how the model is used in practice, and how it works better than existing models in areas such as music generation.