In DSL class, we talked about XLNet, which uses TransformerXL for transfer learning, replacing BERT’s masking by sampling different generation orders