Debugging Guide

Here are some of the most common observations and the underlying issue:

- **Observation**: when overfitting on a single headline, the loss decreases very quickly, but the model does not generate the correct headline.
  - **Issue**: Off-by-one in the decode function. In particular, the first token that the decoder outputs should not be the token, but rather the following token after . The `decode` function is replicating the input token, rather than predicting the next token. Since it is very easy to replicate the input token, the loss decreases very quickly.
- **Observation**: when overfitting on a single headline, the same token is always generated.
  - **Issue**: The order of the dimension of the input tensor is incorrect. PyTorch is interpreting the input as `N` different sequences, each of length 1, rather than 1 sequence of length `N`. Then, when making predictions for these `N` sequences, since PyTorch is not conditioning the first token on any information, the prediction for that first token is the same for all `N` sequences.
- **Observation**: Google Colab crashes when initializing the model.
  - **Issue**: The sizes provided to `AutoEncoder.__init__` method is incorrect, and asks Colab to initialize a large number of weights, requiring too much memory.
- **Observation**: Google Colab crashes when computing the embeddings of the validation data.
  - **Issue**: The batch size is too large. In fact, to avoid padding, we should be using a batch size of 1 to compute the embeddings of the validation data.