Sentence-Level Extractive Text Summarization
Nathan Zhao, Yi Liu, Yijun Jiang

Introduction
Extractive summarization is the identification of the most relevant sentences in a document which encapsulates its main points. Graph-based, Bayesian, and machine learning have all been applied to this difficult task. Recently, deep learning has also found success in this domain. Here, we investigate a recent end-to-end deep learning framework called NeuSum.

Model
The model consists of two parts:
1. Sentence encoder: two BiGRUs that encode sentences on sentence level and then document level.
2. Joint sentence scoring and selection: scores encoded sentences and selects one at each time step. The sentence scores dynamically change with selection.

Objective function requires evaluation of two distributions at each time step $t$:
Model prediction distribution
$$P_t(S) = \frac{\exp(\alpha_t(S))}{\sum_{k=1}^{L} \exp(\alpha_t(S_k))}$$
Reference distribution
$$Q_t(S) = \frac{\exp(\gamma_t(S))}{\sum_{k=1}^{L} \exp(\gamma_t(S_k))}$$

Loss is the KL-divergence between $P_t$ and $Q_t$, summed over $t$.
$$D_{KL}(P_t||Q_t) = -\sum_S P_t(S) \log \left( \frac{Q_t(S)}{P_t(S)} \right) \Rightarrow L = \sum_t D_{KL}(P_t||Q_t)$$

Data
The Cornell Newsroom dataset is a corpus of 1.3 million documents from 38 different news outlets with abstractive summaries to pair. We siphon off a subsample of 100,000 documents for this project (Fig. 3).

To design an extractive baseline, we test a few algorithms which run significantly faster than a brute-force combinatorial search (Fig. 4).

Multi length summaries appear to better optimize Rouge-1 F1 score with the abstractive baseline (Fig. 5).

Results and Discussion
We trained the fixed-length NeuSum ($n$-summary) model on $n$-sentence extraction data:
1. For Rouge-1 F1 score (Table 1), NeuSum does not beat LEAD-3.
2. Distribution of predicted sentence indices (Fig. 6) with a long tail matches training data well, which is also observed in the original paper.

In adaptive-length NeuSum, we allow the model to choose padding sentences. The model learns to pad once the summary reaches optimal length, as <pad> does not count into Rouge.

We trained NeuSum on 2-sentence data while forcing 3-sentence predictions. The ratio of the 3rd sentence being <pad> increases over the time of training (Fig. 7).

Future Work
- Fully vectorize loss evaluation and train on larger datasets
- Investigate into adaptive NeuSum model

References