Gated Attention

Twisha Naik (tn268)
Prakruti Joshi (phj15)
Keya Desai (kd706)

Guide: Prof. Karl Stratos

Problem Statement (Recap)

Implementation of paper:

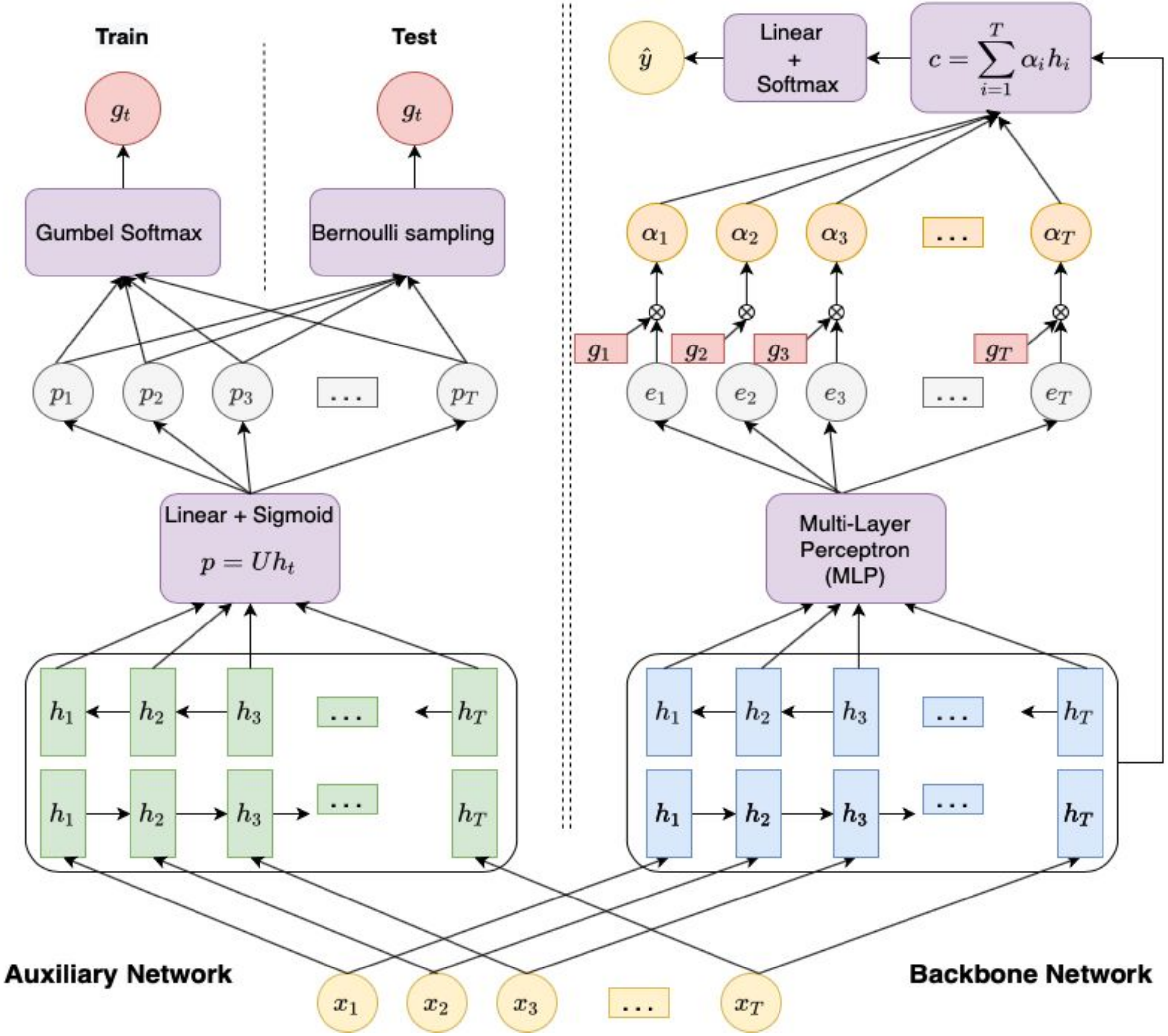
“Not all attention is Needed: Gated Attention Network for Sequence Data” [1]

Gated Attention Concept and Motivation:

- Dynamically select the units to attend to.
- Avoids unnecessary computation and allows model to focus on important elements of the sequence.
- Sparser attention network

Application: Text Classification (TREC: 6 question classes, IMDb: Positive or negative)

Computational Graph



Gumbel Softmax and Loss Function

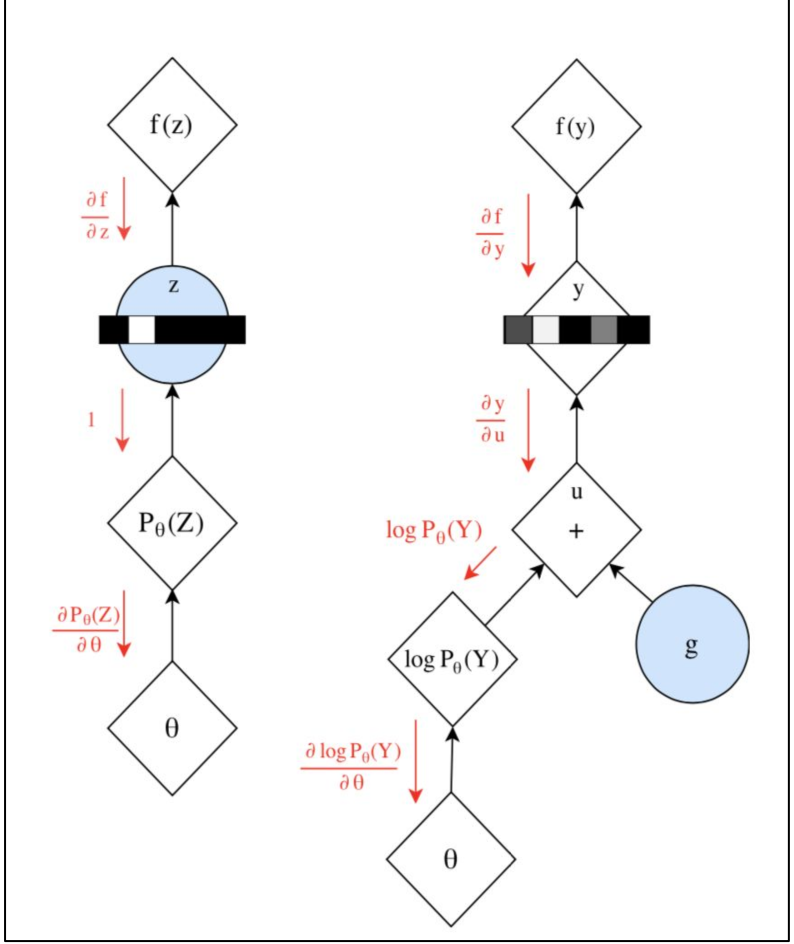
Why Gumbel Softmax?

- Reparameterization trick (similar to the one in the Assignment)
- To make Stochastic node differentiable

Why the new loss function?

- Cross Entropy + L1 norm of Gates

$$\mathcal{L} = -\sum_{k} y_k \log \hat{y}_k + \frac{\lambda \|\mathcal{G}\|_1}{T}$$



f(z)
$\frac{\partial f}{\partial z}$
z
1
$P_\theta(Z)$
$\frac{\partial P_\theta(Z)}{\partial \theta}$
θ
f(y)
$\frac{\partial f}{\partial y}$
y
$\frac{\partial y}{\partial u}$
u
+
$\log P_\theta(Y)$
$\log P_\theta(Y)$
g
$\frac{\partial \log P_\theta(Y)}{\partial \theta}$
θ

Image Courtesy [2]

Issues

- Probability distribution for picking gates getting too small
- Due to this, all the gates sampled were 0 during testing (Predict without input sentence!)

Approaches:

- Test time: Made all gates 1 (This gives baseline accuracy of Soft Attention)
- Train time: Changed loss function to keep at least k gates open - If all gates are 0 then also it is penalised

Hyperparameter Tuning

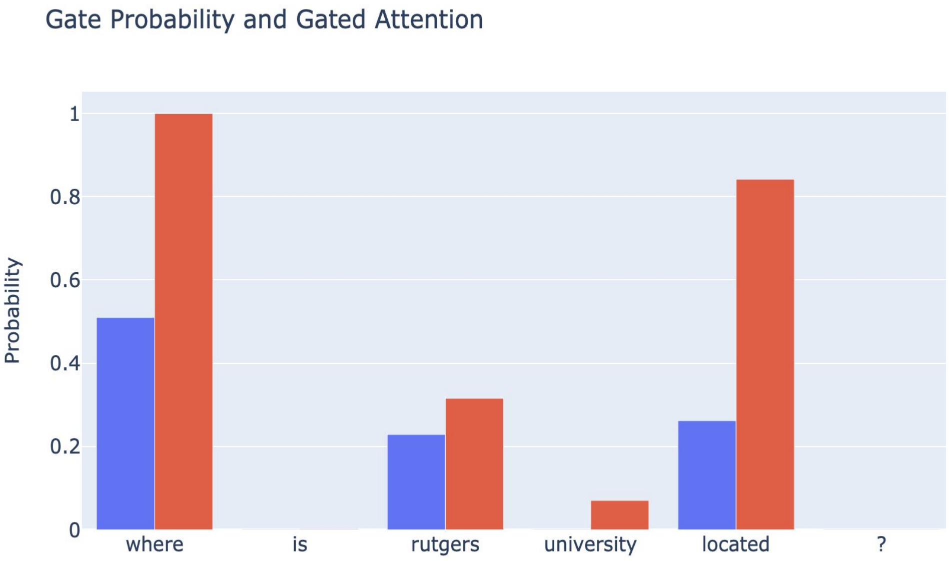
Hyperparameter	Values	Optimum Value
Learning Rate (α)	[2e-5, 1e-4, 2e-4, 5e-4, 1e-3, 2e-3, 5e-3]	2e-4
Temperature ($\boldsymbol{\tau}$)	[0.5, 1, 1.5, 2.0]	1
Regularization parameter (ƛ)	[4e-6, 5e-6, 1e-5, 1e-4, 4e-4, 5e-4]	4e-4
Batch Size	[8, 16, 32, 64, 128]	16 (Trec) 32 (IMDb)

Number of epochs:
Convergence criteria : Validation loss does not decrease for 10 epochs

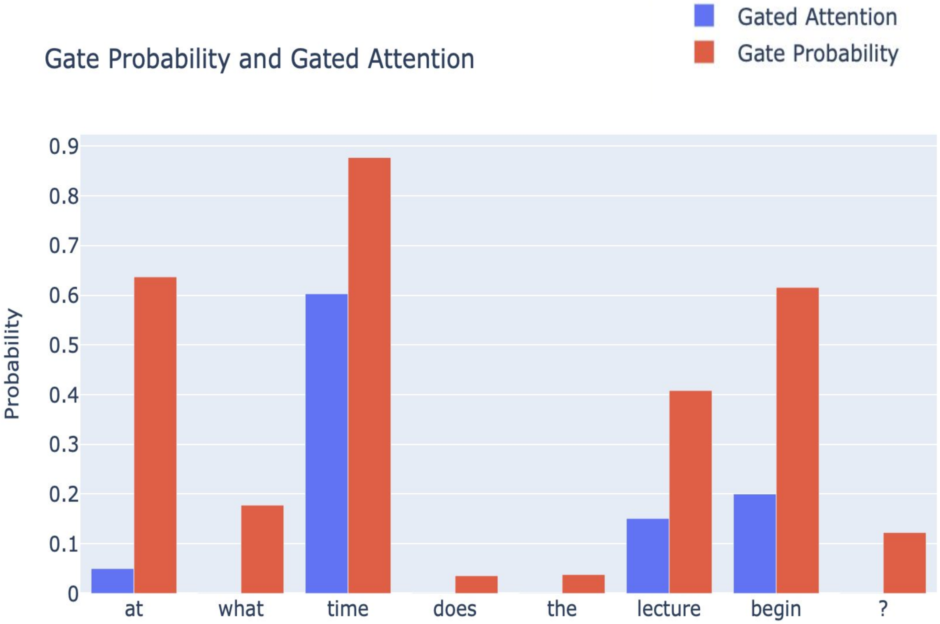
Results

Network	**TREC** (Sequence length = 10)			**IMDB** (Sequence Length = 100)		
	Train Acc.	**Test Acc.**	**Density**	**Train Acc.**	**Test Acc.**	**Density**
LSTM	0.96	0.815	-	0.972	0.755	-
BiLSTM	0.97	0.812	-	0.973	0.765	-
BiLSTM Soft Attention	0.95	0.821	-	0.979	0.772	-
GA-Net (LSTM+LSTM)	0.927	0.830	0.616	0.985	0.780	0.332
GA-Net (FF+BiLSTM)	0.923	0.832	0.552	0.985	0.781	0.327
GA-Net (LSTM+BiLSTM)	0.954	0.836	0.431	0.986	0.785	0.329
GA-Net (BiLSTM+BiLSTM)	**0.964**	**0.842**	**0.371**	**0.989**	**0.787**	**0.324**

Attention results - TREC



Label: Location



Label: Numeric

Attention results - IMDB

Example 1a: "This is one of the **best creation** of Nolan. I can say, it's his **magnum opus**. **Loved** the **soundtrack** and especially those **creative dialogues**."

Prediction: Positive (**Density**: 10/24 = 0.416)

Example 1b: "This is one of the **worst creation** of Nolan. I can say, it's his **magnum opus**. **Hated** the **soundtrack** and especially those **creative dialogues**."

Prediction: Negative (**Density**: 10/24 = 0.416)

Example 2:

Now this is **more like** it! **One of the best movies** I have ever seen! **Despite** it made **very well on all aspects**, this movie was **put down** solely for **not being** too **historically accurate**. Loosen up! There are tons of historical movies out there that were forgiven for **not being too historically accurate** and many of them do not even come close to how grand, how **entertaining** and how **captivating** this movie was! Now this is what a movie ticket is all about! ..... If the viewer of this movie is open minded and has the ability to separate **politics** from art,you will find this movie not only **one of the best classics**, but also **one of the best movies** of all time. I rate it the **second best** western **ever**, right behind Wayne's The Cowboys

Prediction: Positive

Density: 0.371

Takeaways

- Theoretical Study
  - Gumbel Softmax and Backpropagation
  - Tuning different Attention models
- Model Implementation
  - Baseline models of LSTM, BiLSTM
  - Comparison of accuracy and density with Soft Attention
  - GA-Net with different combinations of Backbone and Auxiliary Network
- Results
  - GA-Net with BiLSTM as both Auxiliary and Backbone works the best (in terms of accuracy and density of network)
  - Analysis of results on longer and shorter sequences
- Novel Approaches tried
  - Change of loss function
  - Analysis of misclassified examples (Mostly ambiguous sentences) - Trying out Aspect Based Sentiment Analysis

References

1. Lanqing Xue, Xiaopeng Li, and Nevin L Zhang. 2019.

   “Not all attention is needed: Gated attention network for sequence data”

2. Eric Jang, Shixiang Gu, and Ben Poole. 2016.

   “Categorical reparameterization with gumbel-softmax”

Thank you!

Questions?