Homework 4
Abhijit Prakash Bhatnagar

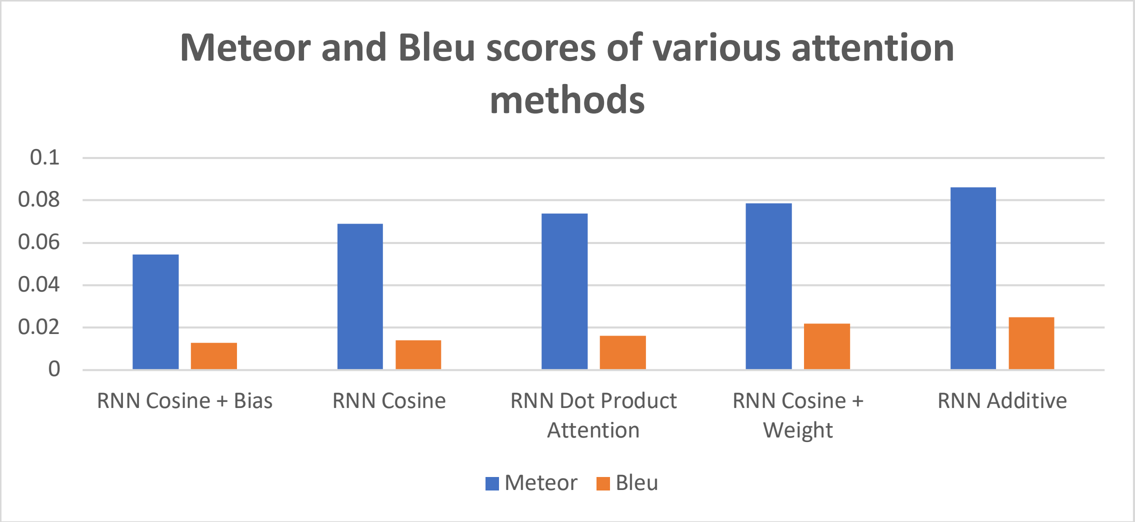
Question 1

1. (1 pt) Incorporate BLEU score into the evaluation script
   See attached code
2. (1 pt) Incorporate METEOR score into the evaluation script
   See attached code
3. (1 pt) Write the code implementing the lead-N baseline
   See attached code
4. (1 pt) report metric values for lead-5, lead-10, and cosine

	Meteor	Bleu
Cosine	0.0666923267195759	0.017648784672166164
Lead-5	0.029603970224661547	0.0006155197606327774
Lead-10	0.054913643306992346	0.005387044545178414

Question 2

1. Implement each kind of attention above. Include this code in your submission.
   See attached code
2. Run the evaluation script from Q1 and report your BLEU and METEOR scores for each kind of attention. Comment on which is the best, and whether you agree (looking at the summaries written in hw4/generations)



	Meteor	Bleu
RNN Cosine + Bias	0.05434909698096362	0.01275059669267597
RNN Cosine	0.0689835096559482	0.014197315458878123
RNN Dot Product Attention	0.07373643350652785	0.016022136762152413
RNN Cosine + Weight	0.07867349433933502	0.02201635047266139
RNN Additive	0.08632443257129131	0.025037633875743086

The sample summaries below show that overall, RNNs are unable to generate himan like statements. Out of all the summaries, the additive attention has the best bleu and meteor scores. However, the summaries are not really legible and it is hard to say if any one model performs better than the other.

Cosine and Bias:
League One side have signed striker season-long loan loan until the end of the season of the season . .

Cosine:
A woman has been found in hospital in hospital after being car in a woman in Gwynedd has been .

Dot product:
slow National the National role the League One until the end of the season of the season of the season .

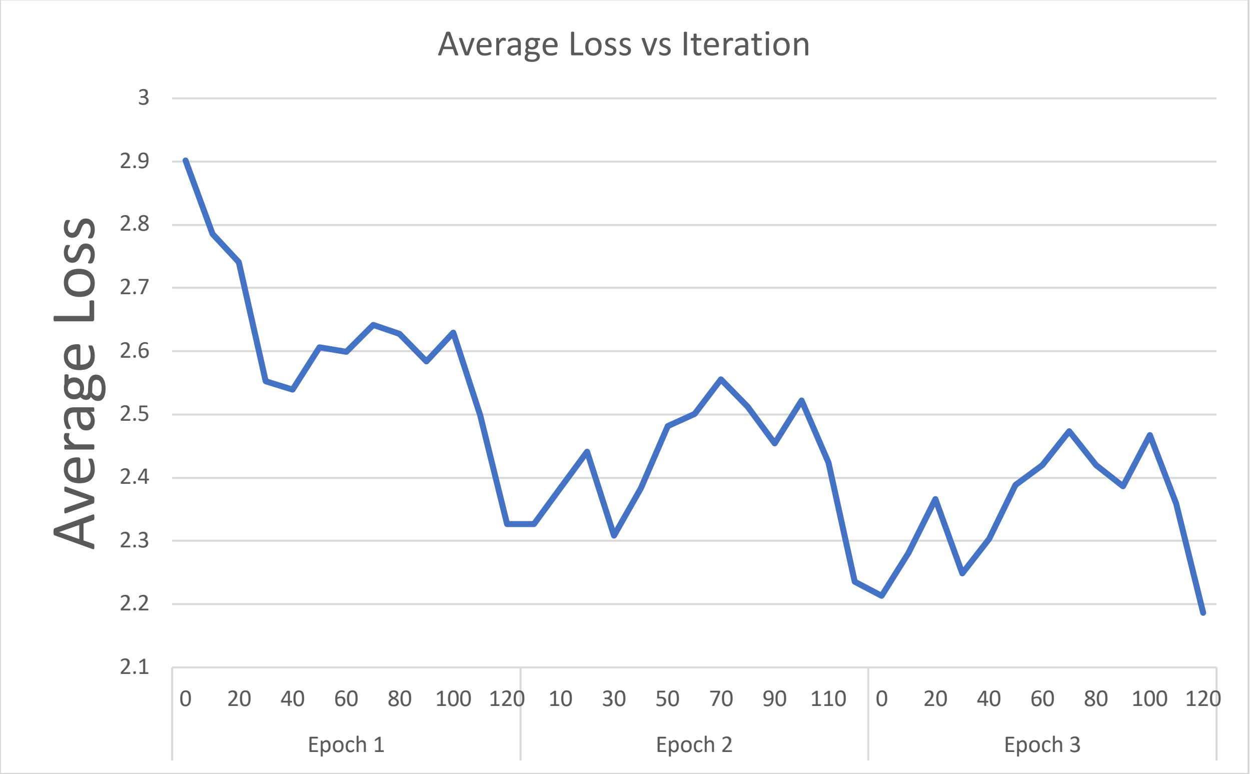
Cosine and weight:
The of of the of the of a Syria of a . in the world , the BBC of the BBC . Saturday of the of the BBC . Saturday . . . ago . Saturday . the of the world . Saturday . the of the . . . . ago . Saturday . a firms . the signing of the of the world . . . Saturday . . . . ago . Saturday . the . . . . ago . Saturday . a firms . the signing of the of the world .

Additive:
More than people have been offered for the first time . months . the first off . . . .

Question 3

1. Code fixing the training loop
   See attached code

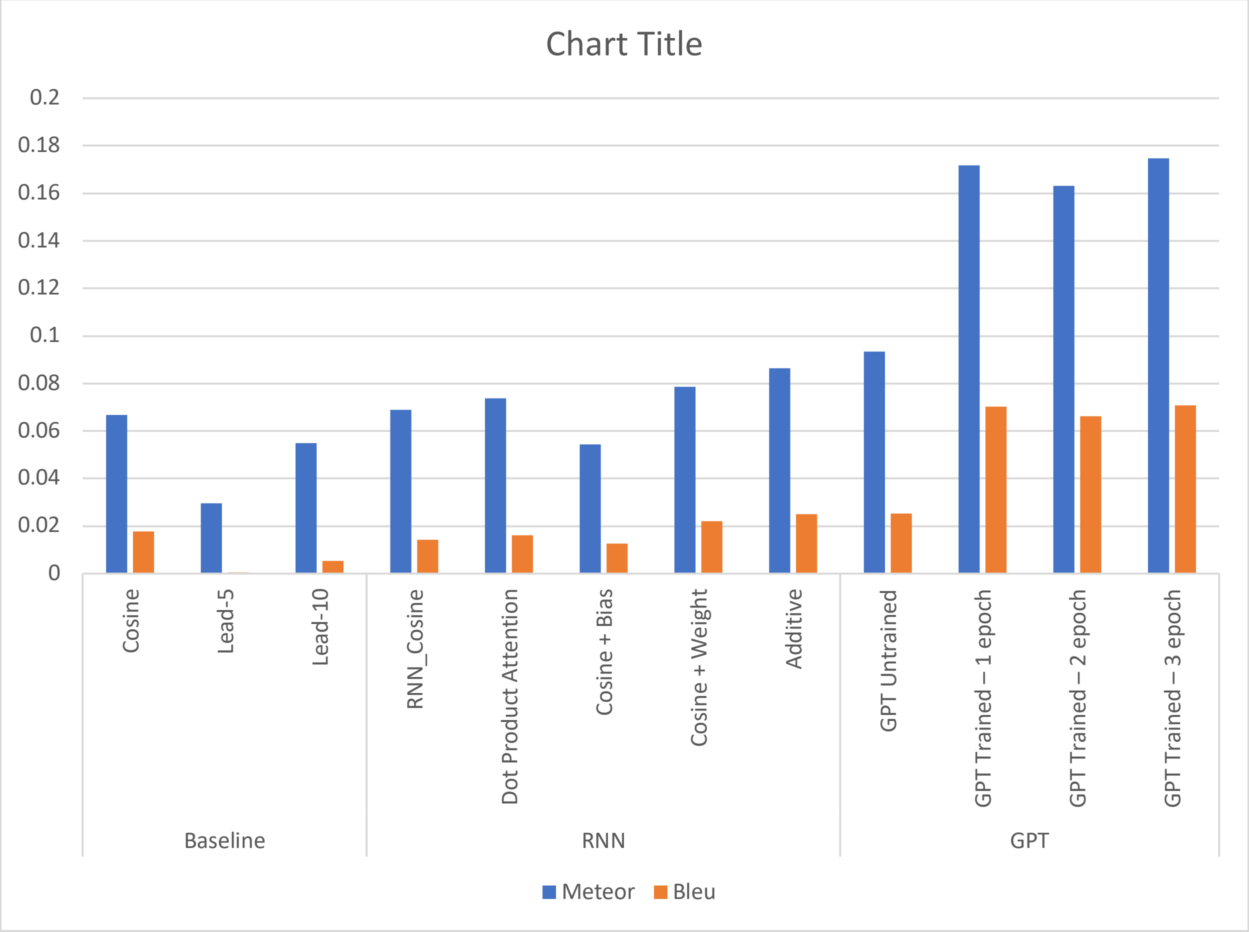
2. Description of what you did to the code for Q3.2
   I modified the code to try total epochs = 0, 1, 2, 3.
   The graph below shows the average loss per iteration across 3 epochs. The average loss decreases as epochs increase.



3. Scores for all models tested

	Meteor	Bleu
GPT Untrained	0.09347619536694315	0.025283065292573268
GPT Trained – 1 epoch	0.17167498701254766	0.07038252142803962
GPT Trained – 2 epoch	0.16321751628798245	0.06619563969947423
GPT Trained – 3 epoch	0.17472824178494872	0.07069547691627812

4. Your overall comparison from Q3.3



Model		Meteor	Bleu
Baseline	Cosine	0.0666923267195759	0.017648784672166164
	Lead-5	0.029603970224661547	0.0006155197606327774
	Lead-10	0.054913643306992346	0.005387044545178414
RNN	RNN_Cosine	0.0689835096559482	0.014197315458878123
	Dot Product Attention	0.07373643350652785	0.016022136762152413
	Cosine + Bias	0.05434909698096362	0.01275059669267597
	Cosine + Weight	0.07867349433933502	0.02201635047266139
	Additive	0.08632443257129131	0.025037633875743086
GPT	GPT Untrained	0.09347619536694315	0.025283065292573268
	GPT Trained – 1 epoch	0.17167498701254766	0.07038252142803962
	GPT Trained – 2 epoch	0.16321751628798245	0.06619563969947423
	GPT Trained – 3 epoch	0.17472824178494872	0.07069547691627812

Comparison of summaries:

Overall, I agree with the reported Meteor and Bleu scores because the GPT model perform best and produce most legible summaries out of all the models. Samples below are clearly more legible than RNN or Lead-N based models. The only confusion is between the scores reported for GPT models trained for 1 and 3 epochs and that trained for 2 epochs. The one trained for 2 epochs has a slightly lower score but the summary produced seems more legible than 1 epoch and 3 epoch models to my eyes.

Untrained GPT:

US banks will continue to face pressure from investors to increase profits. For the first time since 2008, banks are expected to report profits at a lower level

GPT Trained 1 epoch:

In the US, shares of US-listed equities rose at the fastest pace since the start of the year. The Dow Jones industrial average rose 0

GPT Trained 2 epochs:

The latest US data on China's economy is not yet complete. The Dow Jones Industrial Average (DJIA) was down 0.5% at midday

GPT Trained 3 epochs:

Australia's share price fell on news that a quake in southern Japan had left five people dead. The shares of Australian insurers were down 0.1%