STOCHASTIC ANSWER NETWORKS FOR SQUAD 2.0

Paper number : 86

Manan Soni 2017B4A70495P

Vaibhav Ajmera 2017B4A80693P

Vibhu Verma 2017A3PS0189P

STOCHASTIC ANSWER NETWORKS FOR SQUAD 2.0

Xiaodong Liuy, Wei Liy, Yuwei Fangy, Aerin Kimy, Kevin Duhz and Jianfeng Gaoyy

Microsoft Research, Redmond, WA, USA

Johns Hopkins University, Baltimore, MD, USA

Research paper - Aim

1. To build two components a span detector and a binary classifier for judging whether a given question is unanswerable (with respect to the given passage).

2. To Jointly optimize both the components.

3. Getting EM and F1 score for Joint SAN + classifier and comparing it with SAN model and Joint SAN

4. Testing Model's EM and F1-score in SQuAD 2.0 development dataset, SQuAD 2.0 development dataset + ELMo and SQuAD 2.0 test dataset.

Research paper - What it did differently

- The author used a novel stochastic dropout in the answer module. By using it, the models avoids a “step bias problem” and it forces the model to produce good results at every step instead of relying on the prediction of any one particular step

- v2 of the model is jointly optimized by using losses for both the answer module and the unanswerable classifier, which leads to very competitive results.

Research paper - Methodology

1. Used GloVe vectors and CoVe vectors for training.
2. Created Lexicon Encoding Layer, Contextual Encoding Layer, Memory generation Layer to drive the input for getting results.
3. Created Span Detector for getting the length of the answer (and it's endpoints), if it exists.
4. Created an Unanswerable Classifier to detect if the question is unanswerable.
5. Compared EM and F1-score for SAN, Joint SAN and Joint SAN + Classifier.

Research paper - Final Outcome

Single model	EM	F1
SAN [1]	67.89	70.68
Joint SAN	69.27	72.20
Joint SAN + Classifier	69.54	72.66

[1] : http://aclweb.org/ anthology/P18-1157.

Research paper - Final Outcome

SQuAD 2.0 development dataset

	EM	F1
BNA[1]	59.8	62.6
DocQA[1]	61.9	64.8
R.M- Reader[2]	66.9	69.1
R.M- Reader + Verifier Joint[2]	68.5	71.5
Joint SAN[2]	**69.3**	**72.2**

[2] : https://www.arxiv-vanity.com/papers/1808.05759/
[1] : https://www.aclweb.org/anthology/P18-2124/

Research paper - Final Outcome

SQuAD 2.0 development dataset + ELMo

	EM	F1
DocQA	65.1	67.6
R.M -Reader + Verifier	72.3	74.8

SQuAD 2.0 test dataset

BNA	59.2	62.1
DocQA	59.3	62.3
DocQA + ELMo	63.4	66.3
R.M-Reader	71.7	74.2
Joint SAN	68.7	71.4

Introduction - overview of the model

Model tries to understand the passage along with the questions, with context.

Provides an answer using Span Detector

Classifies using unanswerable classifier

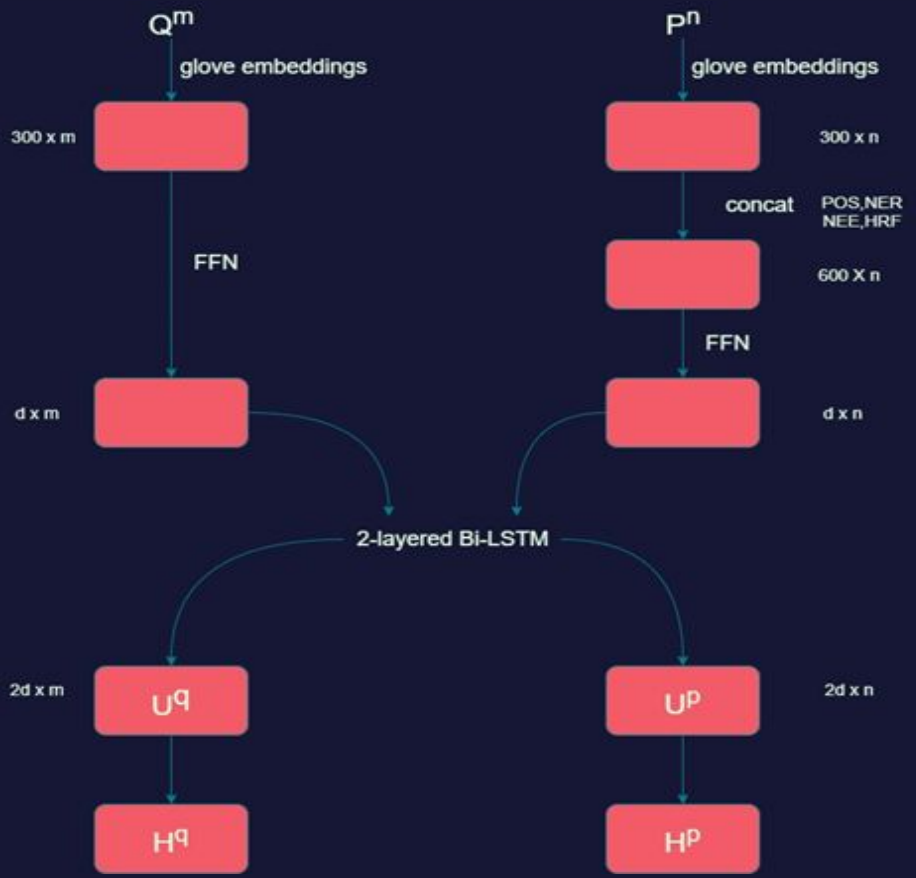
Example

Question: What was one of the Norman's major exports?

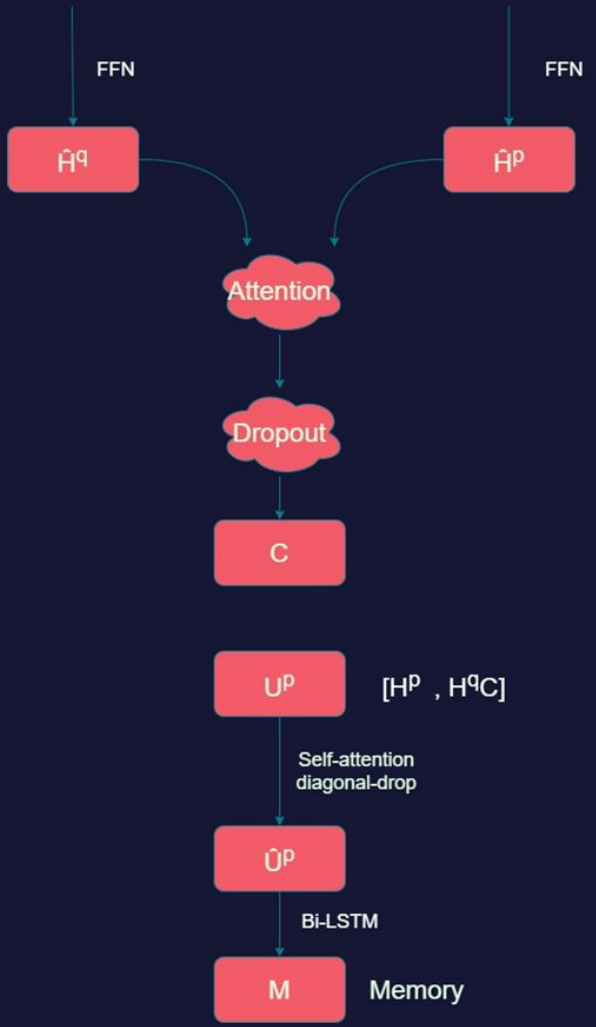
Context: The Normans thereafter adopted the growing feudal doctrines of the rest of France, and worked them into a functional hierarchical system in both Normandy and in England. The new Norman rulers were culturally and ethnically distinct from the old French aristocracy, most of whom traced their lineage to Franks of the Carolingian dynasty. Most Norman knights remained poor and land-hungry, and by 1066 Normandy had been exporting fighting horsemen for more than a generation. Many Normans of Italy, France and England eventually served as avid Crusaders under the Italo-Norman prince Bohemund I and the Anglo-Norman king Richard the Lion-Heart.

Answer: fighting horsemen

Architecture



Architecture



Important Equations

Span Detector

$P_t^{begin} = softmax(s_t W_2 M)$ used to find starting point of answer

$P_t^{end} = softmax(s_t W_3 M)$ used to find ending point of answer

$s_t = GRU(s_{t-1}, x_t)$

$x_t = \sum_j \beta_j M_j$, where $\beta_j = softmax(s_{t-1} W_1 M)$

- M is the model's memory
- s_t is the state of the span detector at the t^{th} time step

Important Equations

Classifier

$$P^u = sigmoid\ ([s_0 ; m_0] W_4)$$

$$m_0 = \sum_j \gamma_j M_j \quad , \quad \gamma_j = \exp(w_5 M_j) / (\sum_{j'} \exp(w_5 M_{j'}))$$

- P^u denotes the probability of the question which is unanswerable.
- Threshold was taken as 0.5
- M is the model's memory
- s_0 is initial state for the GRU
- m_0 represents the summary of the model's entire memory

Important Equations

Loss function

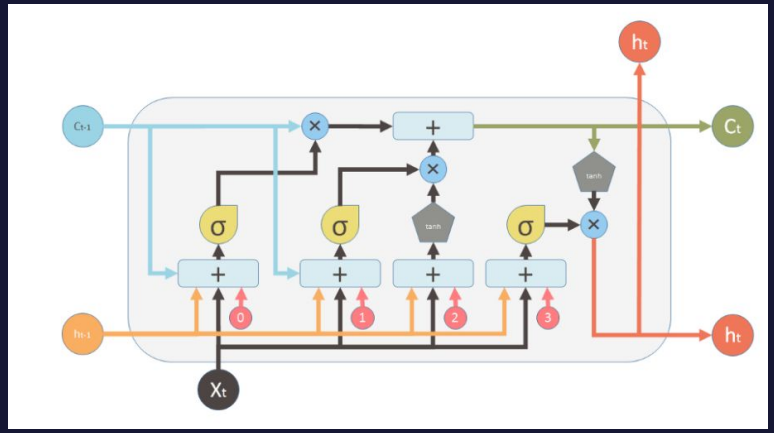
1. $L_{joint} = L_{span} + \lambda L_{classifier}$
2. $L_{span} = -(\log(P^{begin}) + \log(P^{end}))$ (cross-entropy)
3. $L_{classifier} = -y \ln P^{u} - (1-y) \ln (1 - P^{u})$ (Binary cross-entropy)

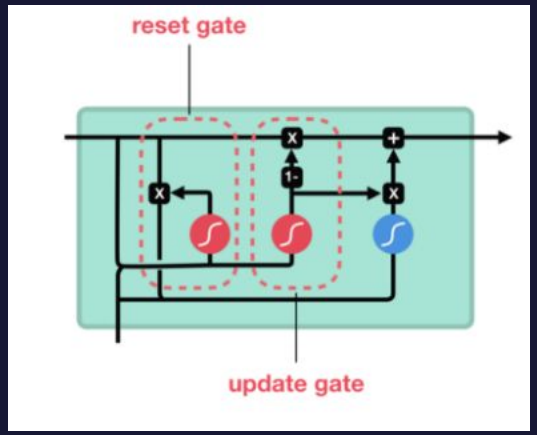
- Author has used $\lambda = 1.5$
- We have experimented on $\lambda = [0.1, 1, 1.5, 10]$

Machine learning concepts

- LSTMs: LSTMs are special kind of RNNs which not only learn about the recent information but also are able to understand the context making them suitable for this application.
- For Implementation Bi-LSTMs are used in the model
- GRUs: Solves the vanishing gradient problem of a standard RNN, can be considered a variation of LSTMs.
- Multi-task learning: Multitask learning involves different tasks which are solved together utilizing the commonalities in them
- Attention: Attention is a way to contextualize the data. When attention is applied it looks among different layers while self attention looks within the same layer.

LSTMs & GRUs - an overview





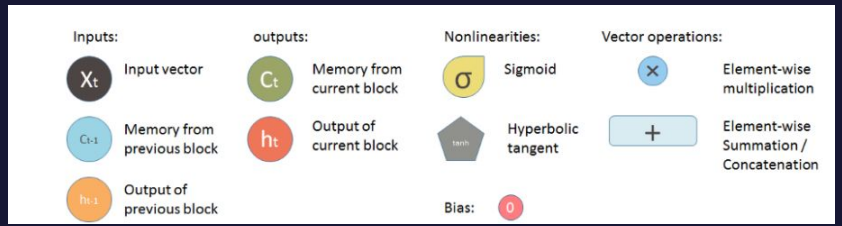
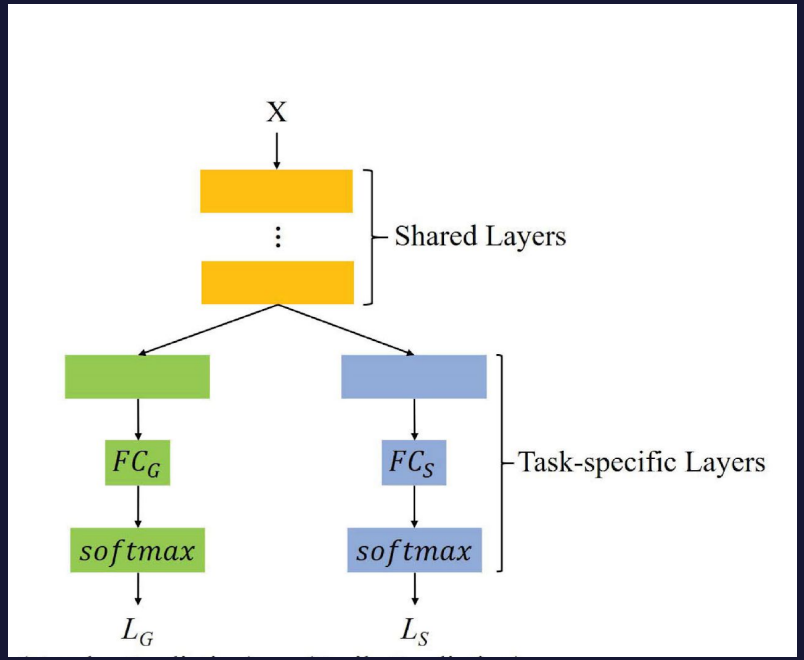
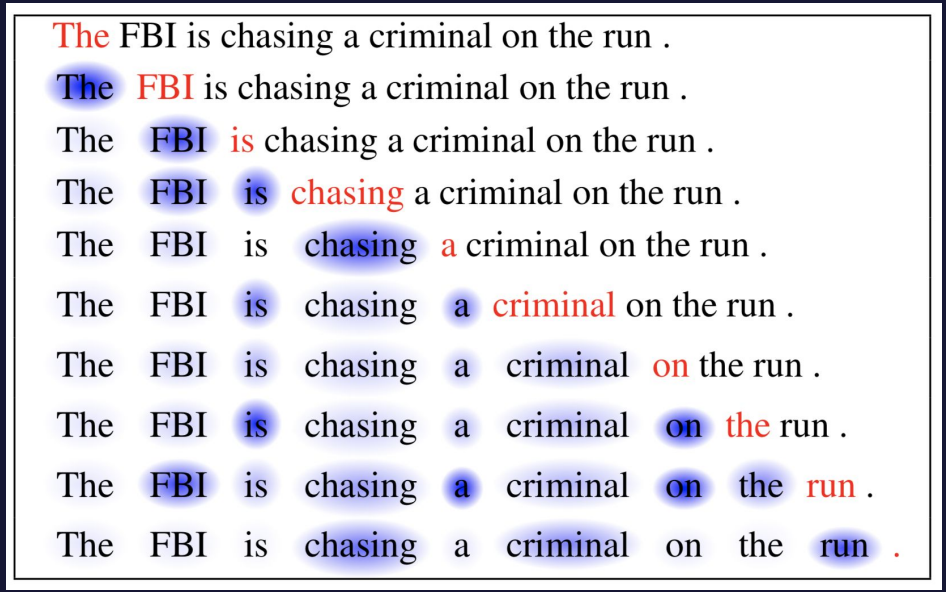


Image Sources : medium.com & towardsdatascience.com

Multi task learning - an overview



Attention



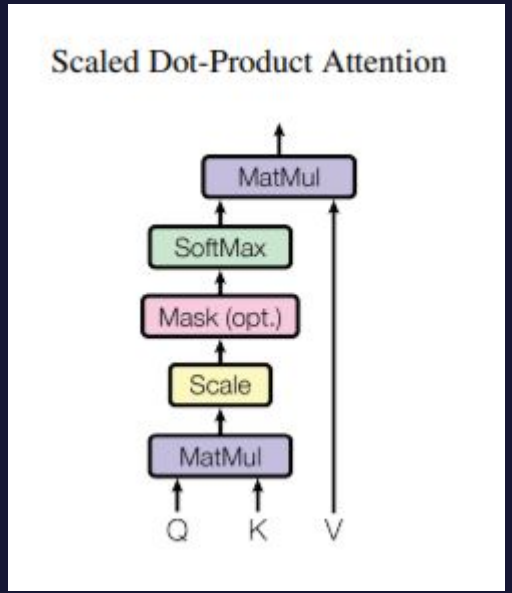


Image Sources : Weng, L. Weng, L. (2018). Attention? Attention!

Dataset Details

For Tokenizing

Dataset	Embeddings
Question	300 - dim GloVe + 600 - dim CoVe vectors
Passage	300-dim Glove + 16-dim POS tagging +8-dim named-entity and 4- dim hard-rule features + 600 -dim CoVe vectors

Dataset Details

For Evaluation

1. SQuAD 2.0 dataset comes under the class of Machine Reading Comprehension(MRC) Dataset.
2. Combination of Stanford Question Answering Dataset (SQuAD 1.0) and unanswerable-question- answers pairs.
3. Contains 23K passages from Wikipedia articles.

Questions	Frequency
Answerable	100K
Unanswerable	53K

Implementation details - Overview

1. Pre-processing: We used the Spacy library to tokenize passages, questions and to generate lemma, part-of-speech and named entity tags.

2. Model: The PyTorch library is used to create the model architecture and for all training and evaluation

3. Plotting: The Matplotlib library is used to generate all the plots that have been shown in this presentation

Implementation details - Installation and setup

- Requires Python 3+ and pip3 to be installed

- To install all dependencies:

  pip3 install -r requirements.txt

  python3 -m spacy download "en"

- For more details, refer the README.md provided in the code

Implementation details - Scripts

- Download the datasets and model weights (trained by us)

  ./get_data.sh
- Pre-process the data

  ./preprocess.sh
- Train the model

  ./train.sh (Recommended to use train.py with correct arguments instead)
- Plot graphs

  ./plot_results.sh (graphs are stored in plot/ directory)
- Evaluate metrics on a model checkpoint

  ./evaluate_metrics.sh path_to_checkpoint
- Predict on custom data using a model checkpoint

  ./predict_sample.sh (data taken from questions.txt and paragraph.txt)

Pseudo code part I - preprocessing

```
train_data, dev_data = load_data(path_to_data)

glove = load_glove_embed(path_to_glove)


# Create vocabulary

vocab = Vocab() # defined in author's code

for word in spacy.tokenize(train_data + dev_data): # tokenize done using spacy directly gives tags as well

    vocab.add(word.text)

    vocab.add_tag(word.tag)

    vocab.add_ner(word.named_entity))
```

Pseudo code part I - preprocessing (continued)

```
# Create Embedding matrix
emb = np.zeros(data_size, 300) # 300 dimensional GLoVE vectors
for word in spacy.tokenize(train_data + dev_data):
    emb.add(glove[word])

# Create pre-processed datasets
questions, passages, raw_ans = process(train_data, vocab, emb)
ans = build_span(passages, raw_ans)
features = get_features(questions) # has PoS, NeR, hard features, etc

# Save everything for later use
save(vocab); save(voab_tags); save(vocab_ner)
save(emb)
save(preproc_train_data); save(preproc_dev_data)
```

Pseudo code part II - training and evaluation

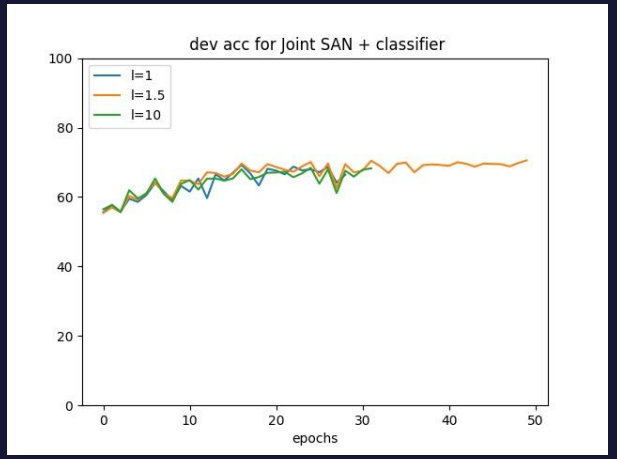
```
model = SAN(classifier='on')
train_data = BatchGen(path_to_preprocessed_train_data) # BatchGen defined by author
train_labels = load_labels(path_to_preprocessed_train_data) # Use for evaluating various metrics
dev_data = BatchGen(path_to_preprocessed_dev_data) # BatchGen defined by author
dev_labels = load_labels(path_to_preprocessed_dev_data) # Use for evaluating various metrics
optimizer = Adam(model.parameters())
```

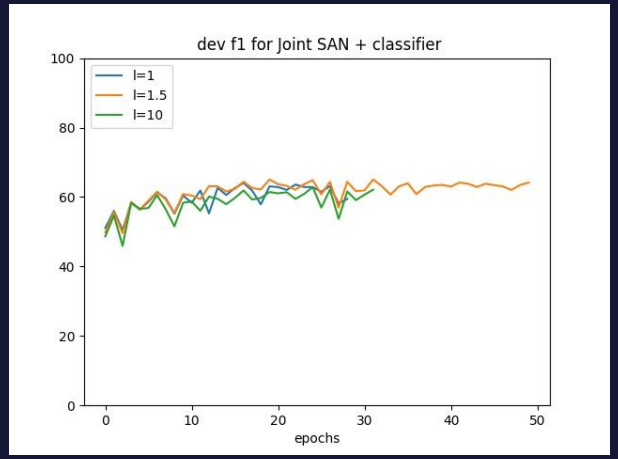
```
for epoch in range(epochs):
    loss = 0.0
    train_em, train_f1, train_acc = 0,0,0
    dev_em, dev_f1, dev_acc = 0,0,0

    for batch in train_data:
        y = model(batch['x'])
        loss =  cross_entropy(y['ans_start'], batch['actual_start']) + cross_entropy(y['ans_end'], batch['actual_end'])
        if classifier is 'on':
            loss += binary_cross_entropy(y['label'], batch['actual_label'])
        loss.backward()
        optimizer.step()

        em, f1 = get_em_f1(train_labels, batch, y)
        train_em += (em / len(batch))
        train_f1 += (f1 / len(batch))
        if classifier is 'on':
            train_acc += (y['label']-batch['acutal_label']).sum().item() / len(batch)
```

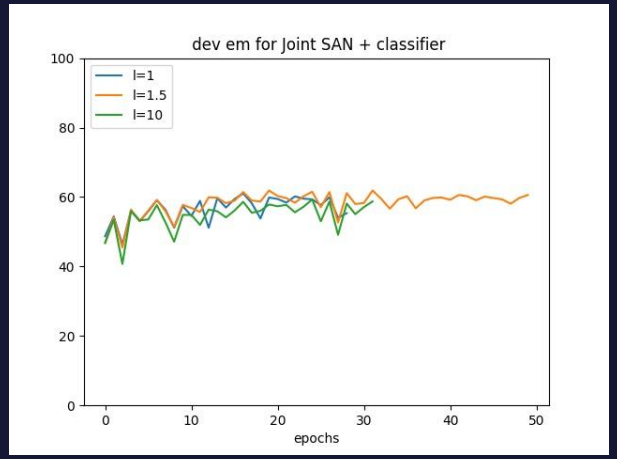
Pseudo code part III - evaluation

```
y_pred = model(dev_data)
em, f1 = get_em_f1(dev_labels, dev_data, y) # dev_data needed for paragraph (context info)
if classifier is 'on':
    acc = (y['label']-dev_labels).sum().item()
    actual_labels = (dev_labels > 0.5)
    predicted_labels = (y_pred['label'] > 0.5)

    cm = confusion_matrix(predicted_labels, actual_labels)) # scipy
    precision, recall, f1, _ = precision_recall_fscore_support(actual_labels, predicted_labels, average='binary') # scipy
else:
    acc = 0
    cm = None
    precision, recall, f1 = 0,0,0
```

Plots - dev metrics for different lambda values (joint SAN + classifier)

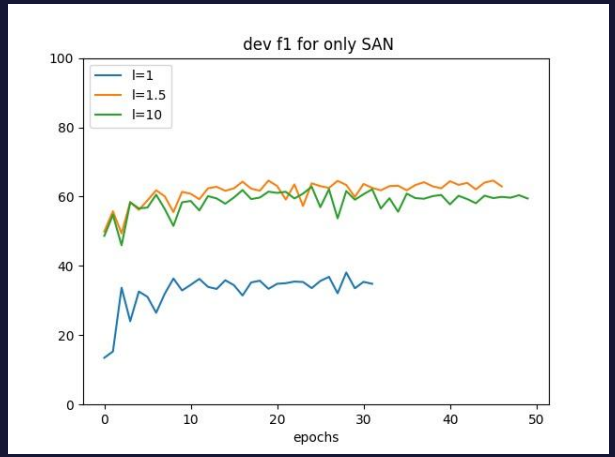


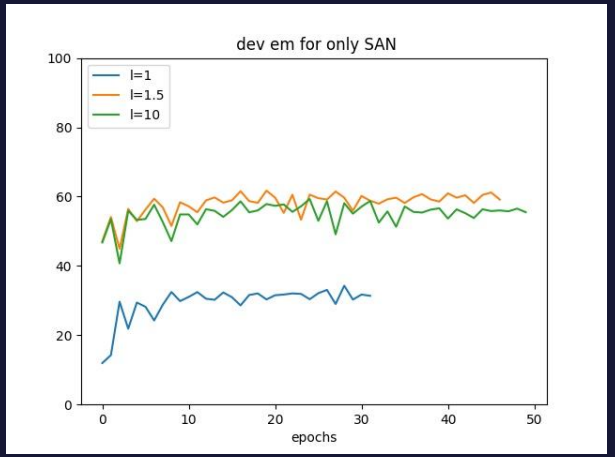




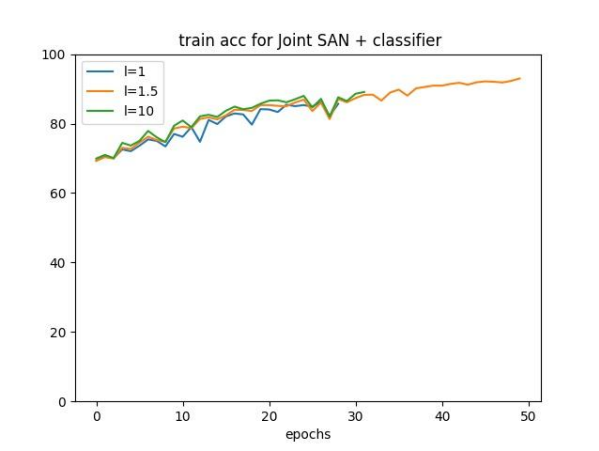
Plots - dev metrics for different lambda values (only SAN)

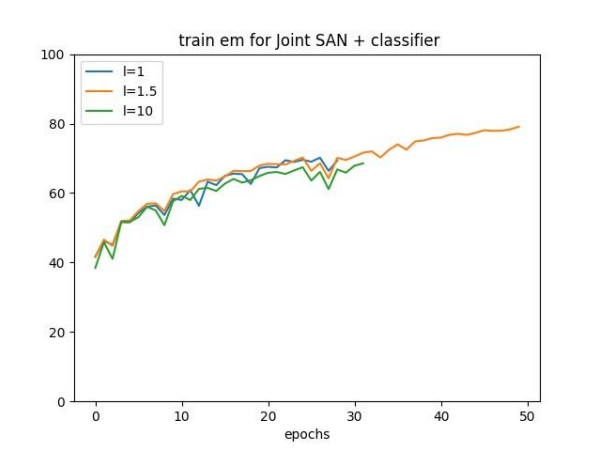
Accuracy can’t be calculated as classifier is disabled

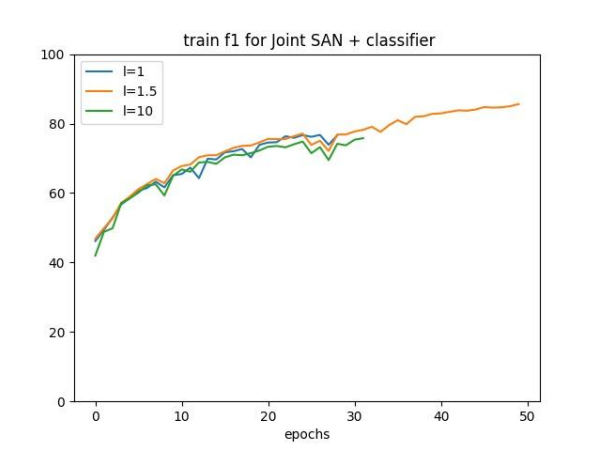




Plots - train metrics for different lambda values (joint SAN + classifier)

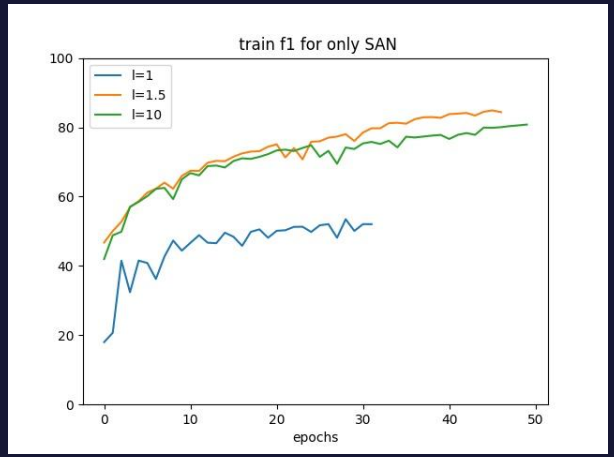


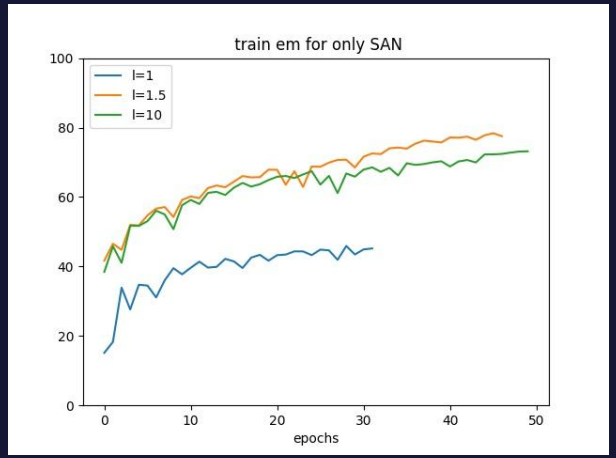




Plots - train metrics for different lambda values (only SAN)

Accuracy can’t be calculated as classifier is disabled

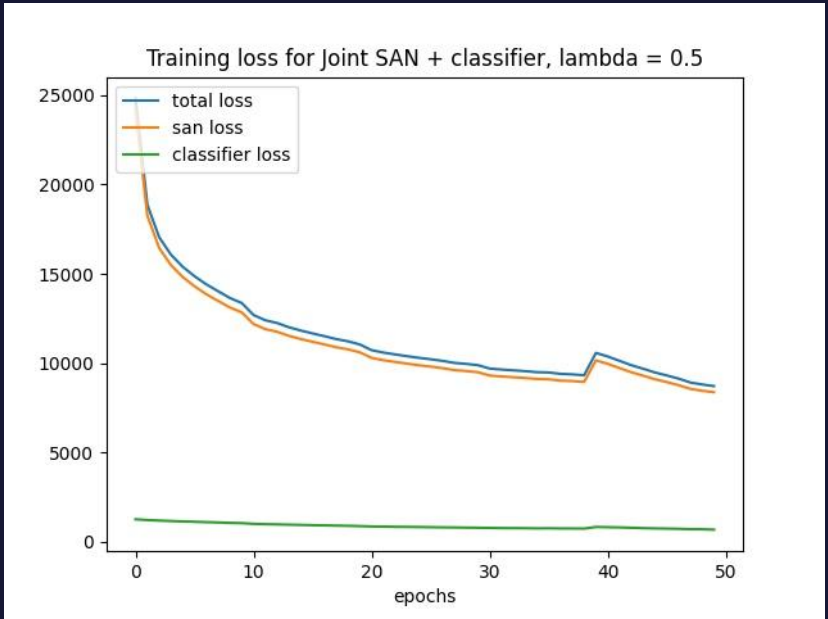




Plots - Train loss for lambda = 0.5

joint SAN + classifier

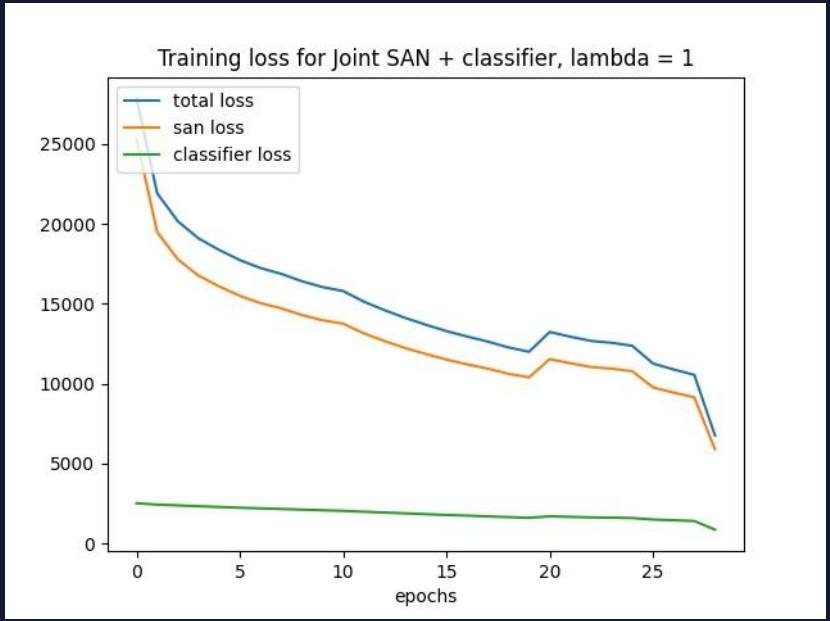
only SAN



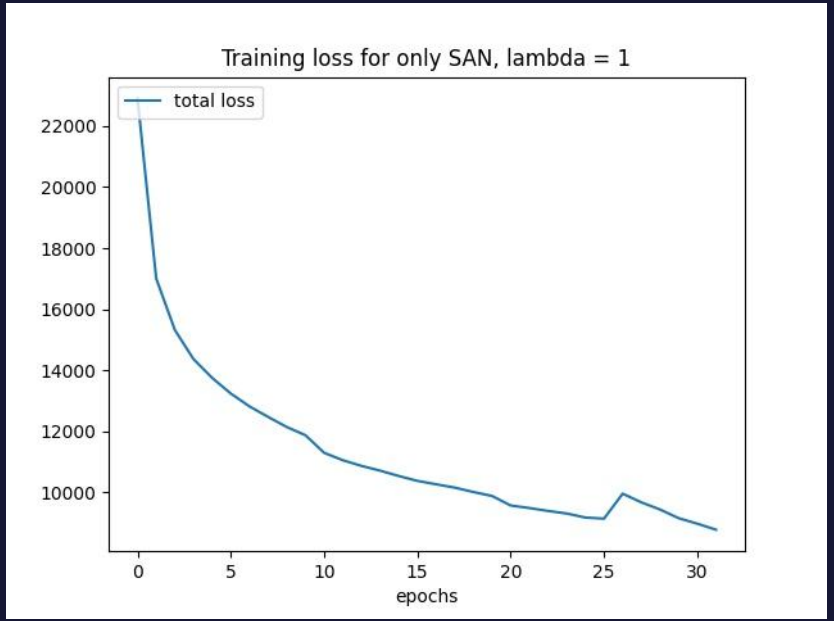
Could not train

Plots - Train loss for lambda = 1.0

joint SAN + classifier

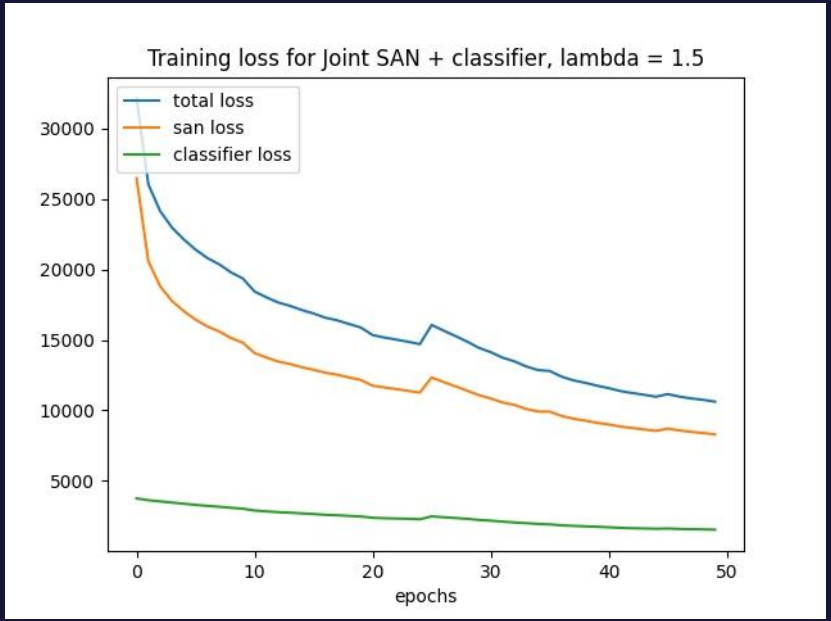


only SAN

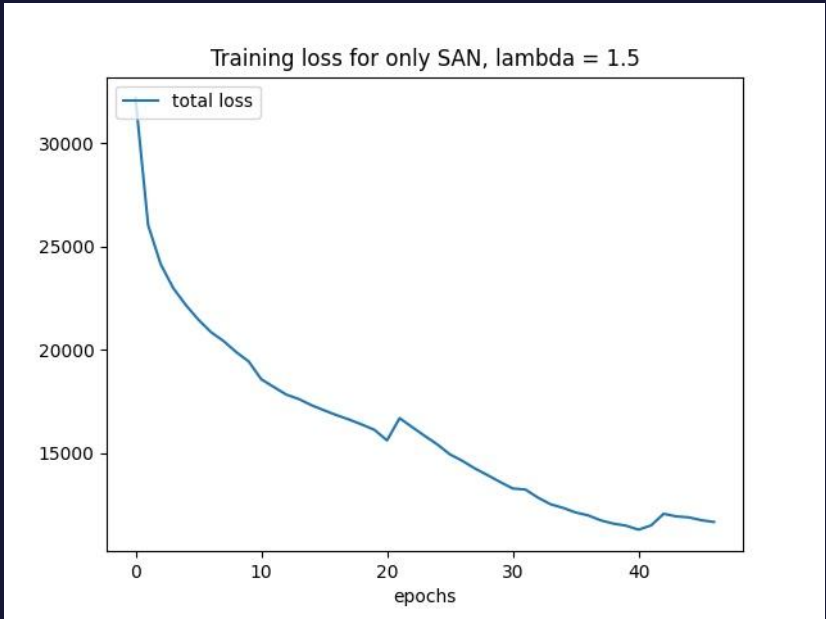


Plots - Train loss for lambda = 1.5

joint SAN + classifier

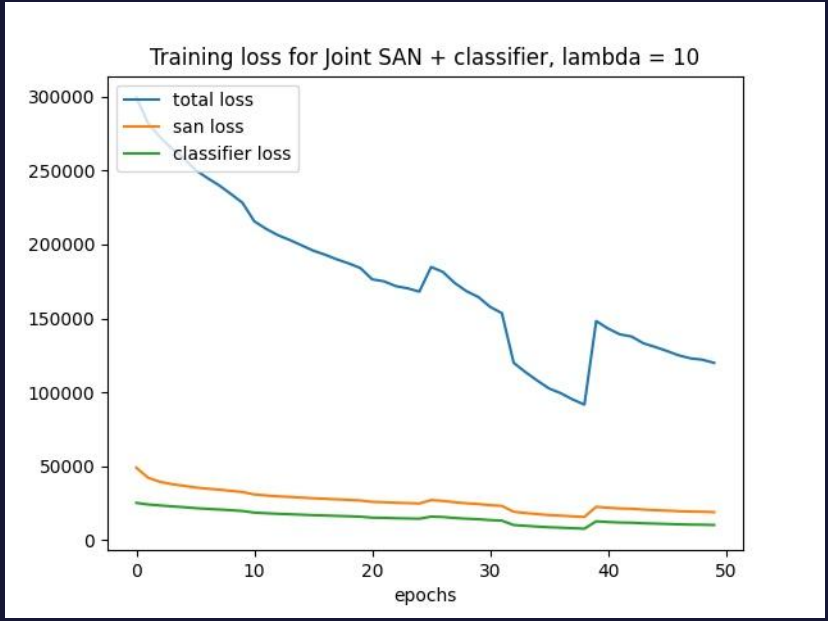


only SAN

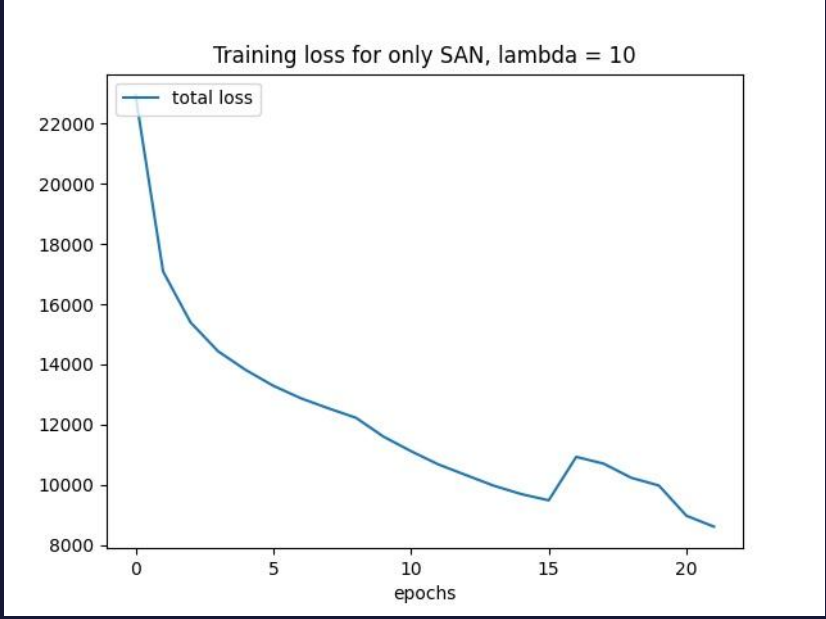


Plots - Train loss for lambda = 10.0

joint SAN + classifier



Only SAN



Plots - clarifications

- Some of the loss function graphs have a sudden spike in the middle, this is most likely because we had to reload our models on Google Colab after ~20-30 epochs and reloading caused the batches to be out of sync and the data also got randomly reshuffled leading to some spikes in the loss function

- For lambda = 0.5, we had trained this model much before the earlier models and could only train on the joint SAN + classifier model

Results - I (we have taken the best values for each column separately)

Lambda value	Model type	Epochs	Train accuracy	Train EM score	Train F1 score	Dev accuracy	Dev EM score	Dev F1 score
0.5	Joint SAN + classifier	50	-	-	-	70.34	62.63	65..75
1.0	Joint SAN + classifier	28	86.27	70.17	76.73	68.69	59.93	63.17
	Joint SAN	31	NA	45.91	53.47	NA	34.28	38.11
10.0	Joint SAN + classifier	50	92.53	73.17	80.79	68.47	59.30	62.84
	Joint SAN	21	NA	44.83	52.38	NA	33.61	36.85
1.5 (best results)	Joint SAN + classifier	50	**92.96**	79.10	85.61	70.55	60.58	63.46
	Joint SAN	50	NA	78.37	84.88	NA	**61.18**	**64.39**

Results - II (evaluation metrics)

1. λ=1

Confusion Matrix :

		Predicted Class	
Actual Class		False	True
	False	-	-
	True	-	-

Precision : -

Recall : -

F1- score: -

Results - II (evaluation metrics)

2. λ=1.5

Confusion Matrix :

		Predicted Class	
Actual Class		False	True
	False	5156	2786
	True	772	3159

Precision : 0.80

Recall : 0.53

F1- score: 0.64

Results - II (evaluation metrics)

3. λ=10

Confusion Matrix :

		Predicted Class	
Actual Class		False	True
	False	5466	3570
	True	462	2375

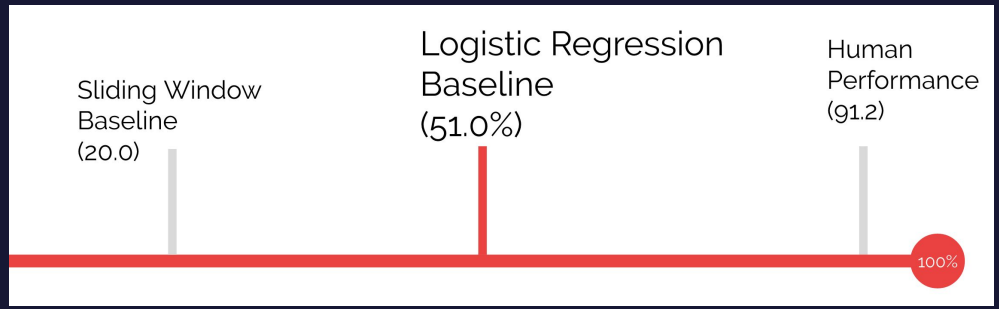
Precision : 0.83

Recall : 0.40

F1- score: 0.54

Discussion - best results

- Our best results (EM: 61.18, F1: 64.39) differ from the author's reported results (EM: 69.54 F1: 72.66) by about 8% but they significantly outperform the regression baseline model by about 13%
- These are reported for Joint SAN + classifier on the dev dataset



F1 baselines and human level performance on Squad v2.0

source: https://rajpurkar.github.io/mlx/qa-and-squad/

Discussion - Reasons for the 8% difference

We think this difference may be due to:

- *The author has used pickled files for pre-processing steps while we have created these files from the train and dev sets, the author's file could have been trained on a larger vocabulary leading to better results*
- *The author has performed a grid search on the number of hidden layers and choosing the best while we fixed these before the training*
- *We were forced to use smaller batch sizes (for gradient descent) as we had access to limited GPUs (on Colab)*

Predicting on custom samples

- You can see our model in action using the predict_sample script
- Enter a paragraph in paragraph.txt and the corresponding questions (1 per line) in questions.txt
- Run ./predict_sample.sh
- We would have loved to show this part interactively in the demo but since that is not possible, we have attached a few sample outputs tested on random wikipedia articles in the next few slides

Predicting on custom samples

- You can see our model in action using the predict_sample script
- Enter a paragraph in paragraph.txt and the corresponding questions (1 per line) in questions.txt
- Run ./predict_sample.sh
- We would have loved to show this part interactively in the demo but since that is not possible, we have attached a few sample outputs tested on random wikipedia articles in the next few slides

Predicting on custom samples

- You can see our model in action using the predict_sample script
- Enter a paragraph in paragraph.txt and the corresponding questions (1 per line) in questions.txt
- Run ./predict_sample.sh
- We would have loved to show this part interactively in the demo but since that is not possible, we have attached a few sample outputs tested on random wikipedia articles in the next few slides

Example 1 - passage

To'ak Chocolate

From Wikipedia, the free encyclopedia

To'ak Chocolate is an Ecuadorian luxury chocolate company that was founded in 2013 by Jerry Toth, Carl Schweizer and Dennise Valencia. To'ak is pronounced Toe-Ahk.[1] The luxury brand To'ak, produces chocolate with the very rare variety of Arriba cacao called Nacional cocoa,[2] which some experts formerly believed to be extinct.[3] It is said to have more floral notes and richness of flavors than any other cacao variety.[2][4] The Heirloom Nacional cacao bar produced by To'ak Chocolate is considered to be the most expensive chocolate bar in the world[5]

Example 1 - Results of our model (lambda = 1.5)

```
------------ Results ------------

Toak Chocolate is an Ecuadorian luxury chocolate company that was founded in 2013 by Jerry Toth, Carl Schweizer
and Dennise Valencia. Toak is pronounced Toe-Ahk. The luxury brand Toak, produces chocolate with the very rare v
ariety of Arriba cacao called Nacional cocoa, which some experts formerly believed to be extinct. It is said to
have more floral notes and richness of flavors than any other cacao variety. The Heirloom Nacional cacao bar pro
duced by Toak Chocolate is considered to be the most expensive chocolate bar in the world.

Ques: Toak chocolate was founded in which year?
Ans : 2013 (confidence = 44.315 %)


Ques: Which cocoa bar is the most expensive chocolate in the world?
Ans : Heirloom Nacional cacao bar (confidence = 41.738 %)


Ques: What is Toak?
Ans : an Ecuadorian luxury chocolate company (confidence = 45.075 %)


---------------------------------
```

Example 2 - passage (more complex)

1982 Individual Speedway World Championship

From Wikipedia, the free encyclopedia

The **1982 Individual Speedway World Championship**.

The 1982 World Final was held in Los Angeles in the United States. This was the only time the Individual World Final was held outside of England or Europe before the advent of the Speedway Grand Prix series in 1995.

The 400 metres (440 yards) speedway track for the Final was laid out over the Los Angeles Memorial Coliseum's existing athletics track. The event was held in front of a crowd of approximately 40,000 people, the largest ever motorcycle speedway attendance in the United States.

1982 Individual Speedway World Championship

Previous: 1981 Next: 1983

Controversy [edit]

Heat 14 of the championship proved to be the most controversial race of the night. After a slow start which saw defending champion Bruce Penhall, and England's Kenny Carter in 3rd and 4th places behind 1976 World Champion Peter Collins and Australian Phil Crump, both riders fought their way past Crump and into 2nd and 3rd behind Collins. Carter moved under Penhall into 2nd place at the end of the 2nd lap, and both riders proceeded to bump each other down the front straight with Carter emerging in front. Carter then went wide through turn 1 which allowed Penhall to come back underneath him. Carter then fell coming onto the back straight and went through the fence causing the race to be stopped. Norwegian referee Torrie Kittlesen then excluded Carter from the race for causing the stoppage. Carter protested claiming that Penhall had hit him in the corner causing him to come off his bike and walked back to the start line in an effort to stop the re-run going ahead without him. Officials and his manager Ivan Mauger were then forced to remove Carter from the track.

Although not shown in the television broadcast of the event, amateur video footage shot from the stands in turns 1 & 2 vindicated Kittlesen's decision. The footage showed that Penhall and Carter did not touch in the turn and that the Englishman had gone down on his own, though years later the debate still rages on about who was at fault.[1] In a television interview with American broadcaster Ken Squier soon after the heat, Kittlesen told that he excluded Carter as he believed the Englishman had fallen without help from Penhall. He also said that the rough riding such as seen from Penhall and Carter on the front straight was to be expected in a World Final. Phil Crump, who had the best view of the incident as he was directly behind the pair, allegedly agreed with the decision to exclude Carter.

Penhall went on to win the re-run from Collins and Crump. In a twist, the result in the re-run ultimately cost Collins' younger brother Les the World Championship in what was his first and only World Final appearance. Had Penhall finished second in Heat 14 behind Peter Collins, and with later results, he and Les Collins could have finished with 13 points each which would have seen the pair in a runoff for the championship. Les Collins had inflicted Penhall's only loss of the meeting when he out-rode the American in Heat 4 in what many believe was a major upset. Additionally, if Penhall had been excluded from Heat 14 and not Carter, Les Collins would likely have won the title as he had finished with 2 point lead over third placed American Dennis Sigalos. Had Penhall only finished on 11 points he would have had a runoff with fellow American Kelly Moran for third place.

In another controversial decision, two races later Kittlesen excluded Czechoslovakia's Václav Verner after a clash with West Germany's Georg Hack, the incident being almost a carbon copy of the Penhall / Carter incident. However, on this occasion it was the rider who stayed on his bike, Verner, who was excluded (though video evidence available to Kittlesen at the time clearly showed Verner's back wheel taking out Hack's front wheel).

Example 2 - Results

```
Ques: Which race was the most controversial?
Ans : Heat 14 of the championship (confidence = 39.438 %)


Ques: Who won the re-run from Collins and Crump?
Ans : Phil Crump (confidence = 45.039 %)


Ques: Who came 1st?
Ans : Could not find (confidence = 46.382 %)


Ques: who came in 2nd place?
Ans : Carter (confidence = 45.809 %)


Ques: who as the person that came third?
Ans : Kelly Moran (confidence = 44.188 %)
```

```
Ques: What was the length of the speedway track?
Ans : 400 metres (440 yards) (confidence = 44.459 %)


Ques: Where was the world final held?
Ans : Los Angeles (confidence = 42.713 %)


Ques: What was the name of the referee?
Ans : Torrie Kittlesen (confidence = 42.99 %)


Ques: What is the debate?
Ans : who was at fault (confidence = 43.371 %)


--------------------------------
```

Example 3 - passage

Birla Institute of Technology and Science, Pilani

From Wikipedia, the free encyclopedia

Coordinates: 28°21′49.96″N 75°35′13.26″E

This article is about the institute in Pilani. For the similarly named institute in Ranchi, see Birla Institute of Technology, Mesra.

Birla Institute of Technology & Science, Pilani (**BITS Pilani**) is a deemed university and Institute of Eminence under Section 3 of the UGC Act 1956 for higher education and research in India.[12] The institute is backed by the Aditya Birla Group and is one of the first six institutes to be awarded the Institute of Eminence status in 2018.[13][14] After expansion to a campus in Dubai, it has become the first international deemed university, spearheading in science, engineering and research with 4 established campuses and 15 academic departments. It focuses primarily on higher education and research in engineering and sciences.[15] Its history, influence, wealth and endowments have made it one of India's most prestigious universities.[16][17][18]

The institute was established in its present form in 1964. During this period, the institute's transformation from a regional engineering college to a national university was backed by G.D. Birla. It is the only deemed university that has been on par with IITs and NITs/RECs and has expanded its campuses from Pilani to Goa, Hyderabad and Dubai. BITS alumni have continued to strongly shape the Indian economy and culture. Through its highly successful and widespread alumni network spanning globally across varied fields, BITS Pilani has made a significant impact on corporates, academia, research, entrepreneurship, arts and social activism.[19][20]

BITS conducts the All-India computerized entrance examination, BITSAT (BITS Admission Test).[21][22] Admission is purely merit-based, as assessed by the BITSAT examination.[23][24] The fully residential institute is privately supported.[25]

Birla Institute of Technology & Science, Pilani

Motto	*jñānaṁ paramaṁ balam* (Sanskrit)
Motto in English	Knowledge is Supreme Power

Example 3 - Results

```
Ques: BITS Pilani is backed by which group?
Ans : Aditya Birla Group (confidence = 38.6 %)


Ques: How many campuses does BITS have?
Ans : 4 (confidence = 44.607 %)


Ques: When was BITS established?
Ans : 1964 (confidence = 46.339 %)


Ques: Why is it one of the most prestigious institutes?
Ans : Its history, influence, wealth and  endowments (confidence = 42.825 %)


Ques: In which areas have BITS alumni made the most impact?
Ans : Could not find (confidence = 47.871 %)


Ques: Which examination do students need to give to get entry into BITS?
Ans : Could not find (confidence = 47.287 %)


Ques: Is BITS privately supported or government funded?
Ans : privately supported (confidence = 46.8 %)
```

Learning Outcomes

1. We learnt how to read and understand research papers and how to convert the theory presented in the paper to code
2. Learnt how to use Google Colab effectively and to quickly train deep learning models along with how to interpret our results (referring to the author's interpretation)
3. We learned practical application of theory concepts learned in lectures

Challenges faced

- Difficult to implement the entire code as the author has used many components that only work with his code (pickle files, directory structures)

- Training the model took a lot of time (~12 hours for 50 epochs per model on Google Colab). We trained about 10-12 models in total

- Time frame was very less, not counting Test-3 we only had about 9 days (after the 1st evaluation) to finish the entire project, and just training the models took up about 5-6 days. It did not leave much scope to try out any new features

Future Scope

1. We can use ELMo vectors instead of the standard GloVe vectors as the ELMo vectors give ~25% better accuracy[1] on the SQuAD dataset

2. Self-Attention can be visualized in the sentences to get a better understanding of how the model processes and understands the questions

[1]: https://allennlp.org/elmo