CS 182/282A: Designing, Visualizing and Understanding Deep Neural Networks **Spring 2019**

Section 8: Question Answering and Pretraining in NLP

Notes by: Philippe Laban

8.1 Course Logistics

- Hopefully you had a nice Spring Break.
- HW03 is due on Wednesday, and HW04 (the last homework, about deep reinforcement learning), will be released this week.
- Don't forget about Midterm 2. See the course webpage, it is scheduled for April 10.
- Projects: you should be moving along on the projects. Don't hesitate to contact your project-assigned TA if you are stuck or need help.

8.2 Question Answering

Before Spring Break, you covered two types of Question Answering systems in lecture: bAbI dataset and SQuAD.

8.2.1 bAbI tasks [1]

Given a sequence of short, simple (baby) sentences, and a question, produce a one-word answer, from a fixed example. The different types of questions are broken down into 20 categories. Figure 8.1 shows ten examples.

Several points of difficulty:

1. Some of the information of the statement is irrelevant.
2. Some answers require assembling information from several statements (look at the second and third examples in Figure 8.1).

Memory Networks (MemNNs) perform well on bAbI tasks. Figure 8.2 shows a diagram of a standard MemNN applied to bAbI.

The procedure is very similar to the QKV attention of a Transformer. In each layer:

- In each layer: the question (or query, q) is embedded using the query projection (matrix B).
- Each input sentence (x_i) is embedded into a key projection (matrix A), and compared (inner product) to the query. This enables the network to "select" which sentences in the input are relevant to the query.

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations
Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions
John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? A:no
Is Daniel in the bathroom? A:yes

Task 7: Counting
Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets
Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation
Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge
John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

Figure 8.1: Sample statements and questions from tasks 1 to 10 of the bAbI dataset [1].

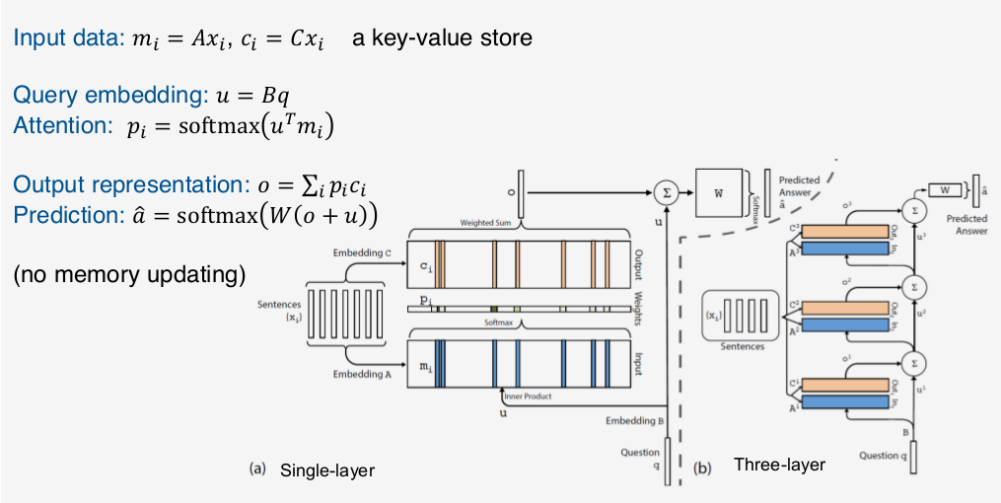

Figure 8.2: Standard Memory Network applied to bAbI task.

- Each input sentence (x_i) is separately embedded into a value projection (matrix C), this esentially can "grab" the useful information from that sentence that we would like to propagate to the output, if the sentence is judged relevant.
- The output is the weighed sum of the input value projections, according to the similarity match obtained in the key-query comparison.
- (Optionally): Can perform "multi-hop" reasoning, use the output from this step as the next query input to a similar next-layer. Multi-hop reasoning often uses 3-layers.
- Finally: produce an answer candidate. The answer is a single word from a fixed-vocabulary (classification). Project the output using a matrix W into a vector the size of the vocabulary, and perform classification (cross-entropy loss).

8.2.2 SQuAD dataset

Some of the limitations of bAbI is that the English is really simple, therefore using Bag-Of-Words on the input side is sufficient, and the answer is a one-word class. The SQuAD setting is slightly different: given a short passage (under 400 words), and a question, produce an answer that is a small interval of words from the short passage. An example of a passage and several questions is shown below.

In SQuAD 1.0, all questions had a possible answer in the text, making the networks trained on them over-confident that an answer is always present. SQuAD 2.0 introduces decoy questions that do not have answers in the passage, expecting the neural network to know when it can and cannot answer a question.

A common competitive approach to SQuAD is to do a QANet, which is a slightly modified version of the Transformer.

8.3 Pre-training in Deep learning.

In Computer Vision, we've seen that CNNs trained on large generic datasets (such as ImageNet) can be fine-tuned on smaller, specific datasets to yield good performance. In the fine-tuning step, only a subset of the weights are retrained, and typically the first layers are "frozen". In 2018 and 2019, the similar idea of pre-training and fine-tuning has had a large effect on common NLP tasks. The generic task that is performed is language modeling, where given a large, unannotated corpus of text, the neural network has to learn to produce next words, or fill-in missing words. Two common approaches are: bidirectional Masked-Language modeling (BERT) and Left-to-right, language modeling (such as GPT-2). These networks are trained on very large corpora of text pulled from the internet, and then finetuned on small datasets.

8.3.1 BERT

BERT uses a Transformer Encoder architecture to read text. Some fraction of the words are replaced by a MASK token (usually 15%) and the network is tasked to recover these words, given all the others. The network has a special token *CLS* that can be used at fine-tuning to use the model for classification purposes. The Figure below shows how the Transformer Encoder trained on the Language modeling task is then adapted to other NLP tasks such as sentence pair classification, SQuAD, etc. The Table afterwards shows how BERT helped outperform highest achieving models on almost all of the tasks, by leveraging the pre-training.

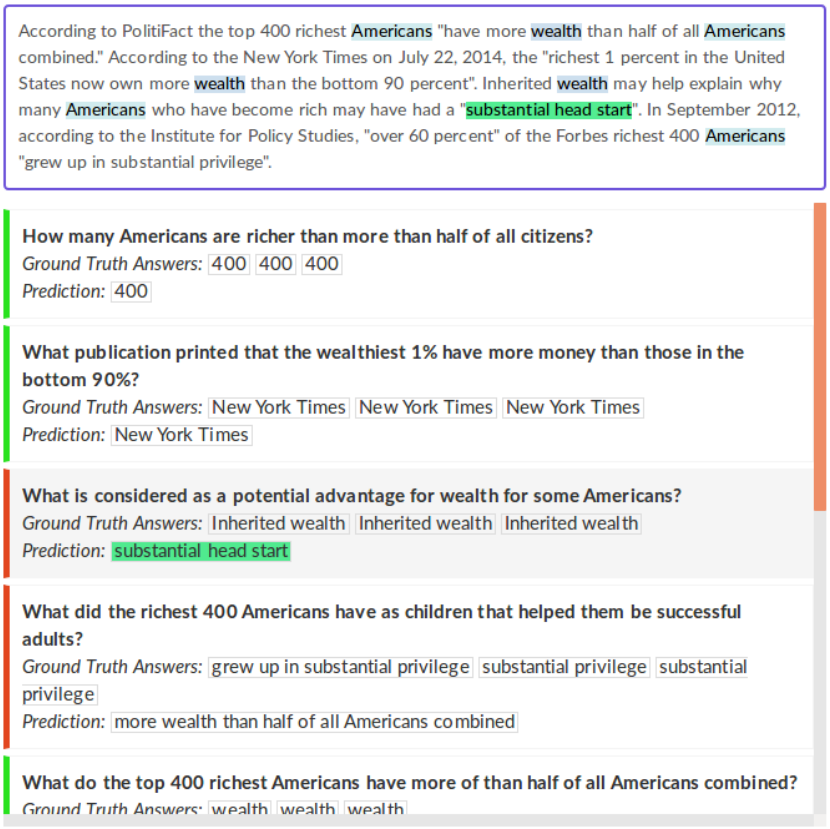

Figure 8.3: Example short passage from SQuAD 1.0, paired with a set of questions and answers from the text.

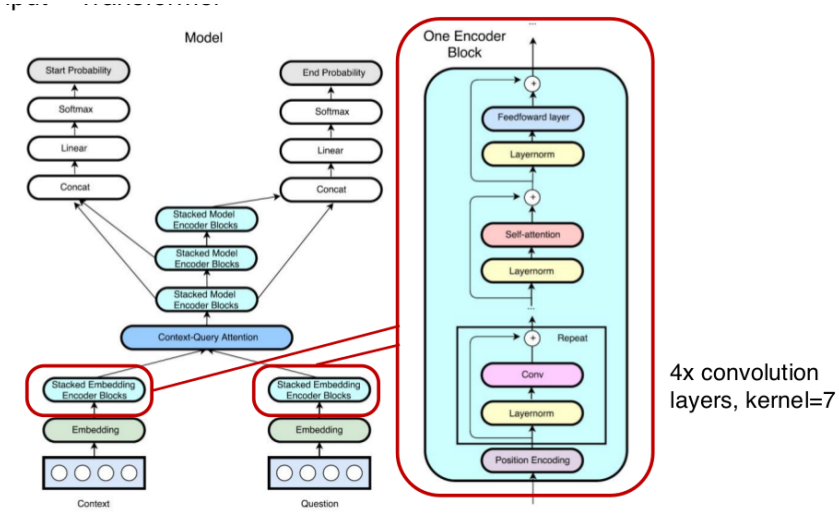

Figure 8.4: QANet, a similar architecture to the Transformer, applied to the SQuAD 1.0 dataset. The model output produces a probability distribution over the words on the input to decide what word is the beginning of the answer, and a probability distribution to decide where the answer ends. In SQuAD 2.0, there is a third classification that decides whether or not the neural network can answer the question.

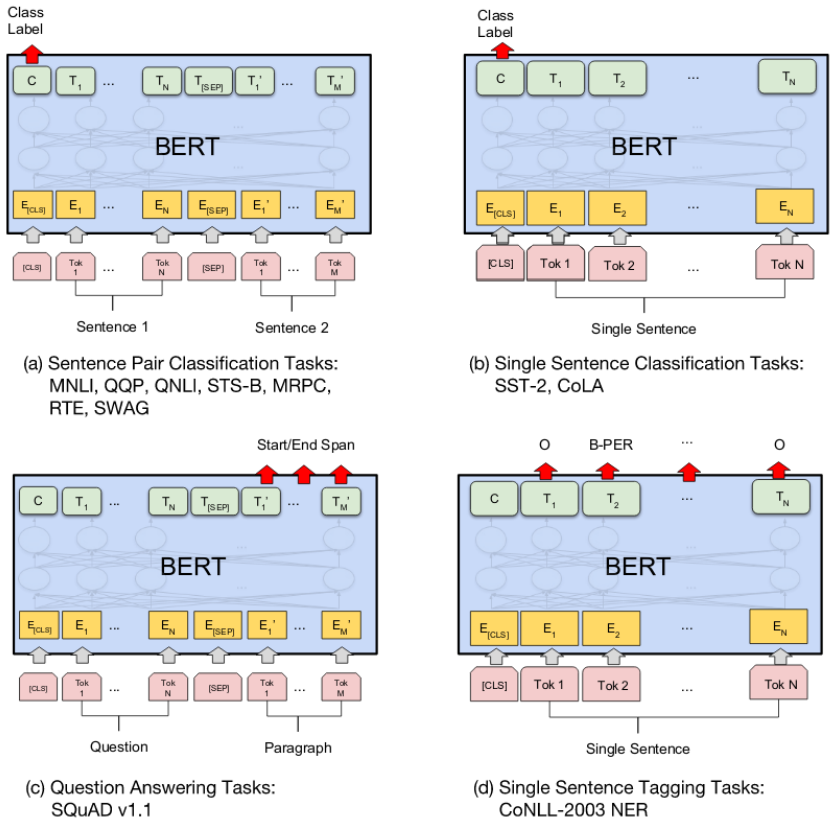

Figure 8.5: BERT model adapted to 4 different types of tasks.

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); $\text{BERT}_{\text{BASE}}$ = (L=12, H=768, A=12); $\text{BERT}_{\text{LARGE}}$ = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from https://gluebenchmark.com/leaderboard and https://blog.openai.com/language-unsupervised/.

Figure 8.6: Once BERT is pretrained, it can be finetuned on common NLP tasks and achieves strong results.

8.3.2 GPT-2

The pretraining task in BERT is Masked Language modelling: the network has to recover a subset of masked words, based on all others. This cannot easily be used for "generation" tasks, as in generation, the words are typically produced in a left-to-right order. GPT-2 propose to pre-train on the standard task of language modeling: predicting the next word in the sequence, and show that as in BERT, they are able to produce state-of-the-art models with finetuning. Their models however could be used on text-generation tasks, such as summarization or translation. Because it is a standard language model, it can also be used to produce text based on a prompt. Here is an example of text the network generated given a prompt:

> **Prompt:** Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.
>
> The orcs response was a deafening onslaught of claws, claws, and claws; even Elrond was forced to retreat. You are in good hands, dwarf, said Gimli, who had been among the first to charge at the orcs; it took only two words before their opponents were reduced to a blood-soaked quagmire, and the dwarf took his first kill of the night. The battle lasted for hours until two of the largest Orcs attempted to overwhelm Aragorn. When they finally stopped, they lay defeated and lifeless for miles and miles.
>
> I take nothing, said Aragorn. But I give my word, at my peril and mine, that I will never forget this day of horror. None of us will forget. Ever!
>
> Ill never forget it! cried Gimli, who had been in the thick of the battle but hadnt taken part in it. One of the wounded orcs he had carried off, he was the only one of the survivors who remained uninjured. Well keep the memory of that day of evil, and the war with it, alive as long as we live, my friends!
>
> Then well keep it alive as long as we live, added Legolas. And we wont forget the first great battle of the night, even if we may have forgotten the final defeat.
>
> I agree, Gandalf said, but we will all remember it as the last battle in Middle-earth, and the first great battle of the new day.
>
> Aragorn drew his sword, and the Battle of Fangorn was won. As they marched out through the thicket the morning mist cleared, and the day turned to dusk.
>
> The Two Rings were returned to Rivendell. Frodo and Sam woke up alone in their room, and Frodo found a note on his pillow. He opened it and read:

May the Power of the Ring be with you always, and may its light never fade. I am not sure if it matters which of the two rings we accept this day but, as you asked me, I have chosen mine. I am sorry to leave you, Frodo, but know that we are very close to the end, and that you are with us forever. May this letter find you safely in Rivendell; and if it does not, then I will accept the ring in your stead. If by any chance you find or give this letter to the enemy, may they learn the strength of the ring and may the Two Rings never be broken!

References

[1] Jason Weston et al. “Towards ai-complete question answering: A set of prerequisite toy tasks”. In: *arXiv preprint arXiv:1502.05698* (2015).