GROUP 6

Improving Question Answering with an Ensemble Approach

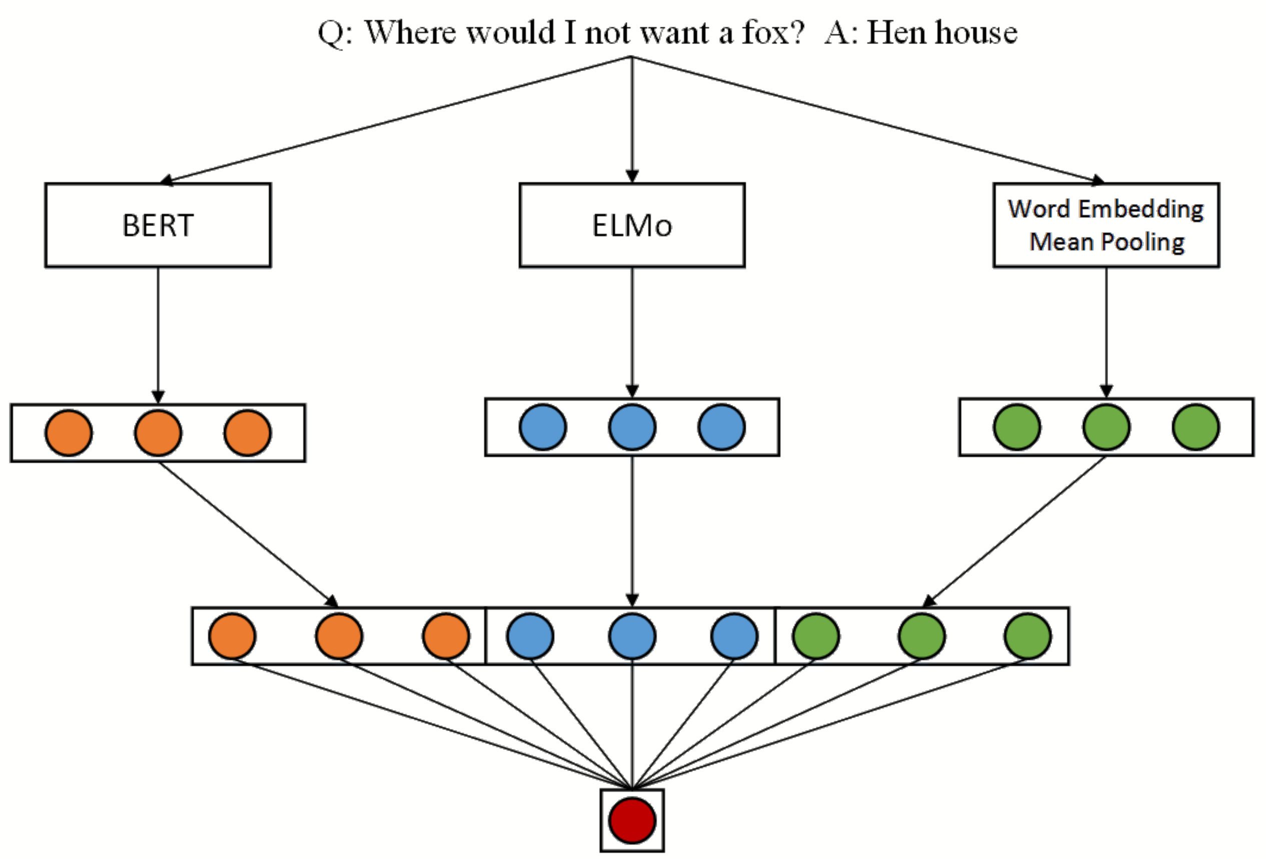
CONG BAO, YUAN GAO, YUAN LI

DEPARTMENT OF COMPUTER SCIENCE

INTRODUCTION

Commonsense reasoning is a major challenge in Question Answering (QA) tasks. In our project, we proposed a ensemble that combining three pre-trained models, which are BERT [1], ELMo [2], and a mean pooling over pre-trained word embeddings such as GloVe [3]. Three models are fine-tuned separately and the feature layers are combined as input for down stream classification task. Several experiments are performed and the results indicates the ensemble model outperforms individual models.

MODLE ARCHITECTURE



Model Description

We combine three pre-trained models in our ensemble model, which are BERT [1], ELMo [2], and a mean pooling over pre-trained word embeddings such as GloVe [3]. After fine-tuning three models individually, the feature vectors of three models are concatenated as a single feature vector as the input of down stream classification task. The model generates a single logit unit for each question-answer pair, which is then fed into choice model for predication.

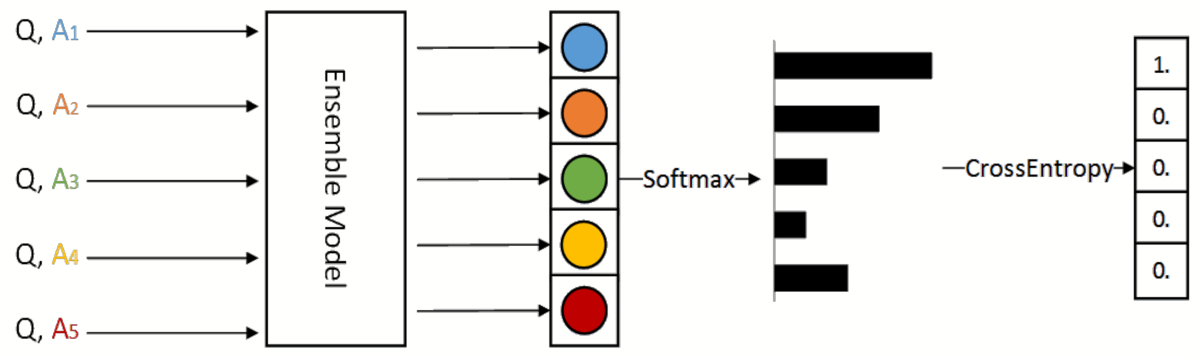
Feature extraction methods for each individual model:

Feature of mean pooling: $f_{AVG} = \frac{1}{|S|}\sum_{t\in S} S(t)$

Feature of ELMo: $f_{ELMo} = W\left(\sum_{i=1}^{L} \gamma_i h_i^{ELMo}\right) + b$

Feature of BERT: $f_{BERT} = W\left(h^{[CLS]}\right) + b$

QA MULTIPLE CHOICE MODEL



Multiple Choice Model Description

The multiple choice model is a special use case introduced in OpenAI GPT [4]. However, instead of using different linear layers for each question-answer pair, we use a common linear layer among all question-answer pairs. The logit units are then concatenated and fed into a softmax function to yield a distribution of prediction. The loss is computed as the cross-entropy between predicted distribution and the true distribution (labels). Weights in ensemble model are updated by back-propagation.

PARAMETERS

	AVG	ELMo	BERT	ELMo+ BERT	ALL
bs	64	64	32	32	32
lr	1e-5	1e-3	1e-5	1e-5	1e-5
epochs	9	9	1	1	1
dropout	0.1	0.1	0.1	0.2	0.2

REFERENCES

1Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR , abs/1810.04805.
2Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Ken- ton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. CoRR, abs/1802.05365.
3Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages1532–1543.
4 Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. CoRR , abs/1811.00937.
5 Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge.

DATASET

CommonsenseQA [5], raised by Tel-Aviv University, is regarded as a novel multiple-choice question answering dataset, aiming to predict the right answer, which is in need of profound knowledge of commonsense. It is composed of 12,247 questions, with a correct answer and four distractor answers. Questions and their answers within the dataset came from ConceptNet [5]. To capture the commonsense beyond associations, each question dis-criminates between three target concepts that all share the same relationship to a single source drawn from ConceptNet.

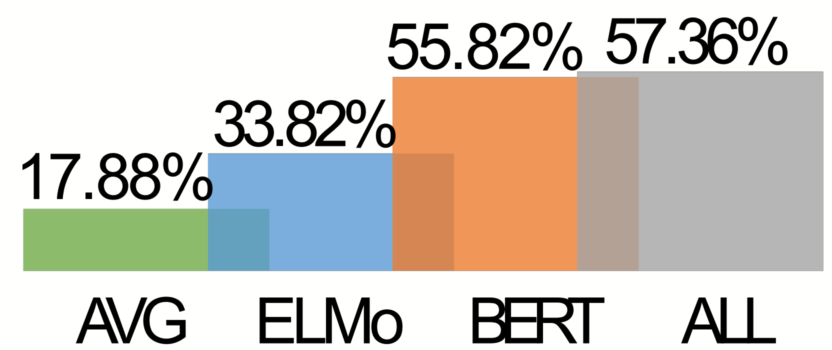
Data Example

Q: Where would I not want a fox?

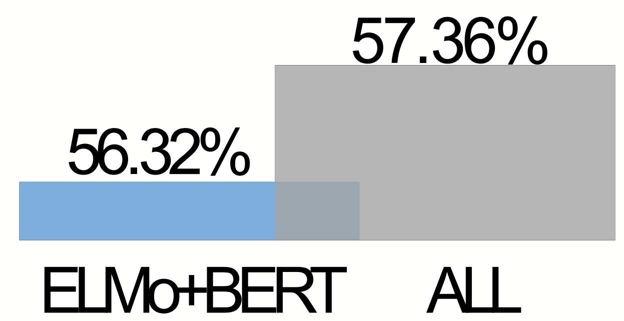
A:

- ✓ Hen house
- ✗ England
- ✗ Mountains
- ✗ English hunt
- ✗ California

EVALUATIONS



Accuracy on development set of three individual models and ensemble model



Accuracy on development set of ensemble models with different combinations