

# 机器学习与数据挖掘

*Machine Learning & Data Mining*

权小军 教授

中山大学数据科学与计算机学院

[quanxj3@mail.sysu.edu.cn](mailto:quanxj3@mail.sysu.edu.cn)


# 考核方式

## □考核方式

- Mid-term Project: 40%;
- Final Project: 40%;
- 课堂作业: 10% ;
- 课堂考勤: 10%

# 考核方式

▣ Mid-term Project: 40%;

- Ensemble Model (20%, by 10 June)
- Support Vector Machine (20%, by 30 June)
- Deep Learning 

# Final Project

- Final project: machine reading comprehension
- Dataset: SQuAD v1.1
- Paper: *SQuAD: 100,000+ Questions for Machine Comprehension of Text*
- 数据收集方式：文章摘自Wikipedia，问题和答案采用网络众包方式收集
- 任务解析：给定一段文本(context)和一个问题(question)，要求在文本中找到对应问题答案的文本范围(span)

# Example

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

Span #1

Context

What causes precipitation to fall?

Question #1

**gravity**

Answer #1

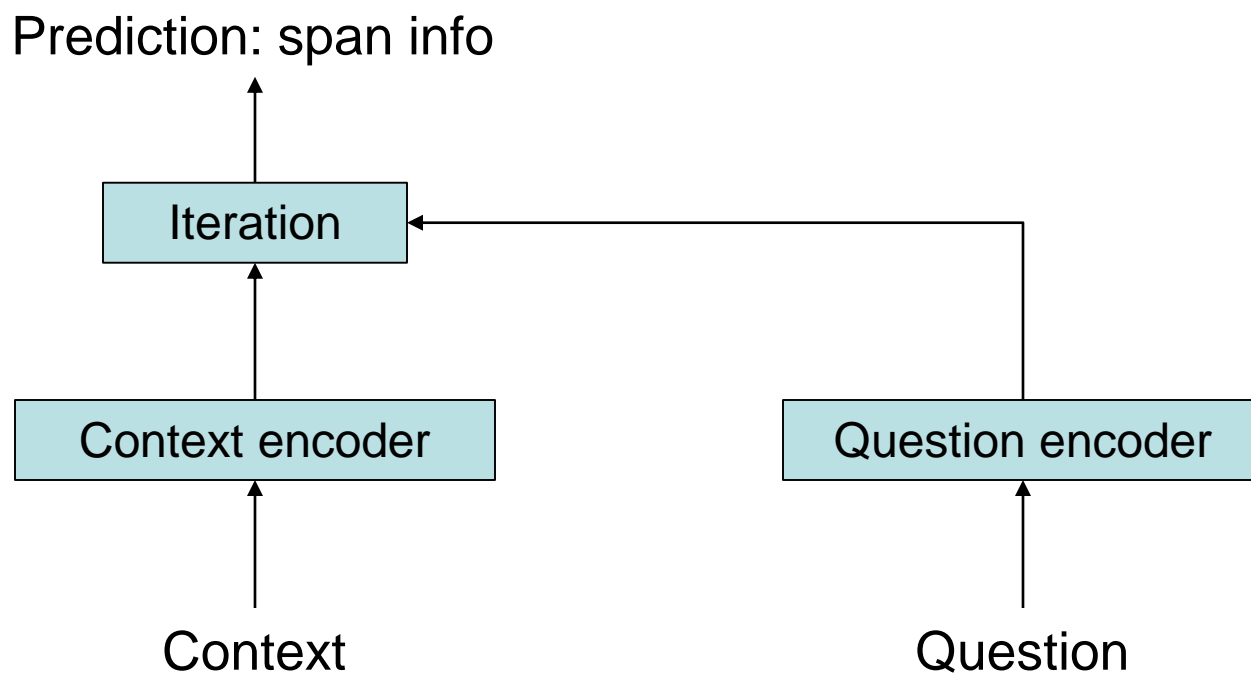
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

**graupel**

Where do water droplets collide with ice crystals to form precipitation?

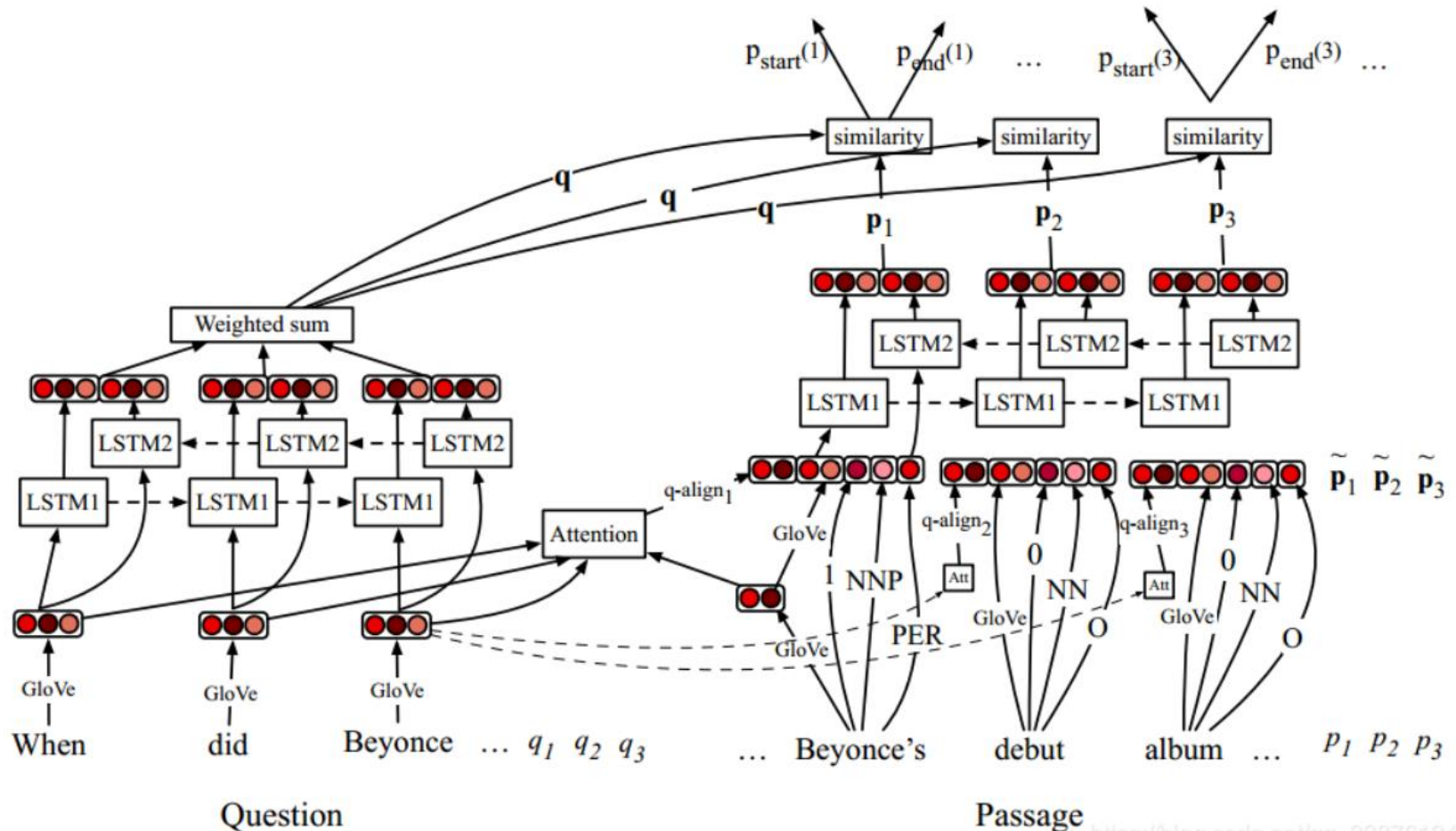
**within a cloud**

# Solution



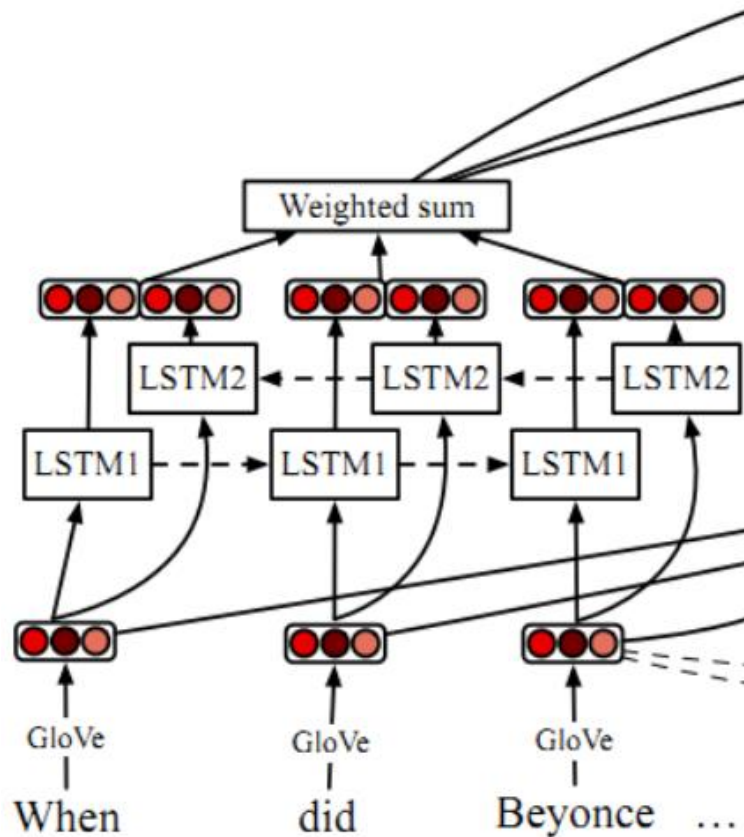
# How models solve SQuAD

## 1. Stanford Attentive Reader (Neural Reading Comprehension and Beyond)



# How models solve SQuAD

## 1. Stanford Attentive Reader (Neural Reading Comprehension and Beyond)



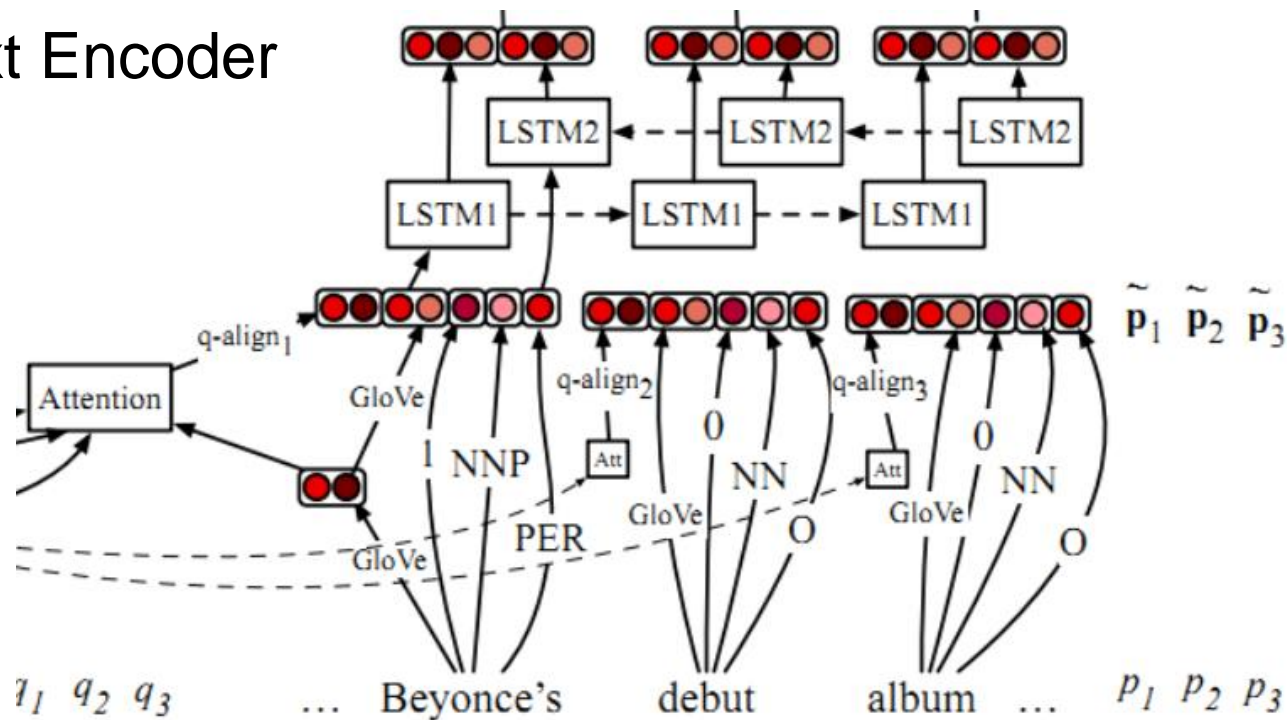
Question Encoder



# How models solve SQuAD

## 1. Stanford Attentive Reader (Neural Reading Comprehension and Beyond)

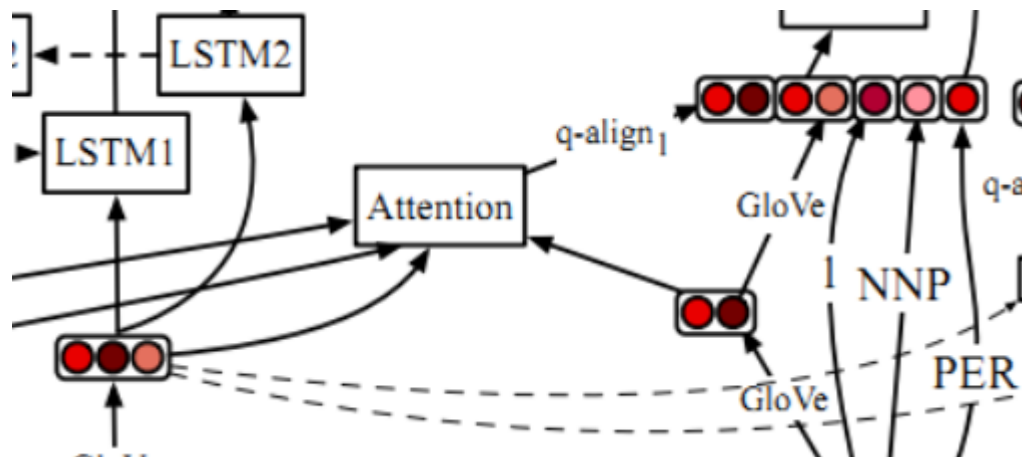
### Context Encoder



# How models solve SQuAD

## 1. Stanford Attentive Reader (Neural Reading Comprehension and Beyond)

### Interaction #1



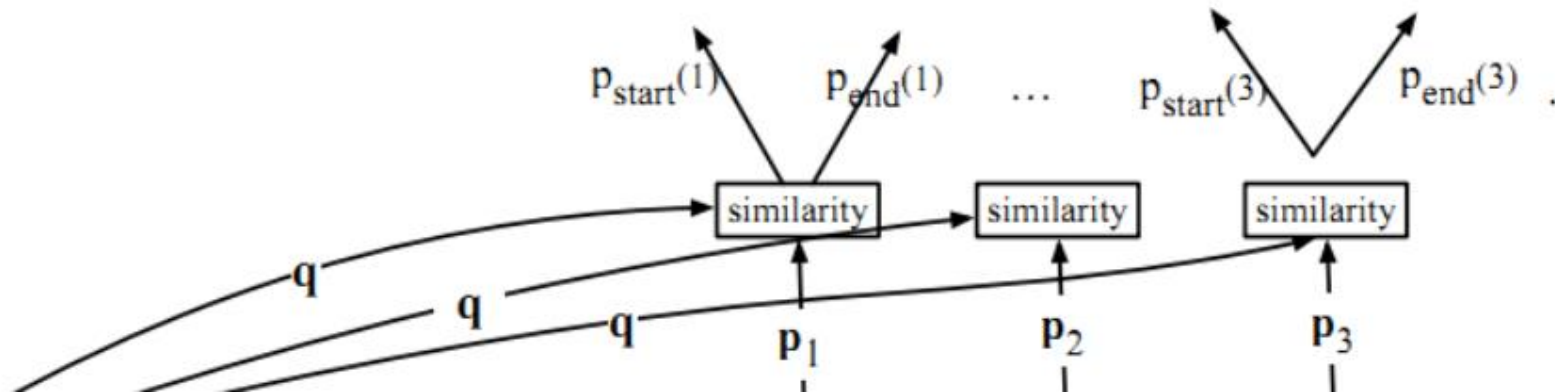
$$f_{align}(p_i) = \sum_j a_{i,j} \mathbf{E}(q_j)$$
$$a_{i,j} = \frac{\exp(MLP(\mathbf{E}(p_i))^T MLP(\mathbf{E}(q_j)))}{\sum_{j'} \exp(MLP(\mathbf{E}(p_i))^T MLP(\mathbf{E}(q_{j'})))}$$

计算文章的glove词向量对问题词向量的注意力，用问题词向量的加权和来丰富文章词向量的特征

# How models solve SQuAD

## 1. Stanford Attentive Reader (Neural Reading Comprehension and Beyond)

### Interaction #2 and prediction



$$P^{(start)}(i) = \frac{\exp(\mathbf{p}_i \mathbf{W}^{(start)} \mathbf{q})}{\sum_j \exp(\mathbf{p}_j \mathbf{W}^{(start)} \mathbf{q})}$$
$$P^{(end)}(i) = \frac{\exp(\mathbf{p}_i \mathbf{W}^{(end)} \mathbf{q})}{\sum_j \exp(\mathbf{p}_j \mathbf{W}^{(end)} \mathbf{q})}$$

对文章的每个位置，使用双线性函数计算文章与问题隐向量的相似性，预测每个位置是答案的开头或结尾的概率

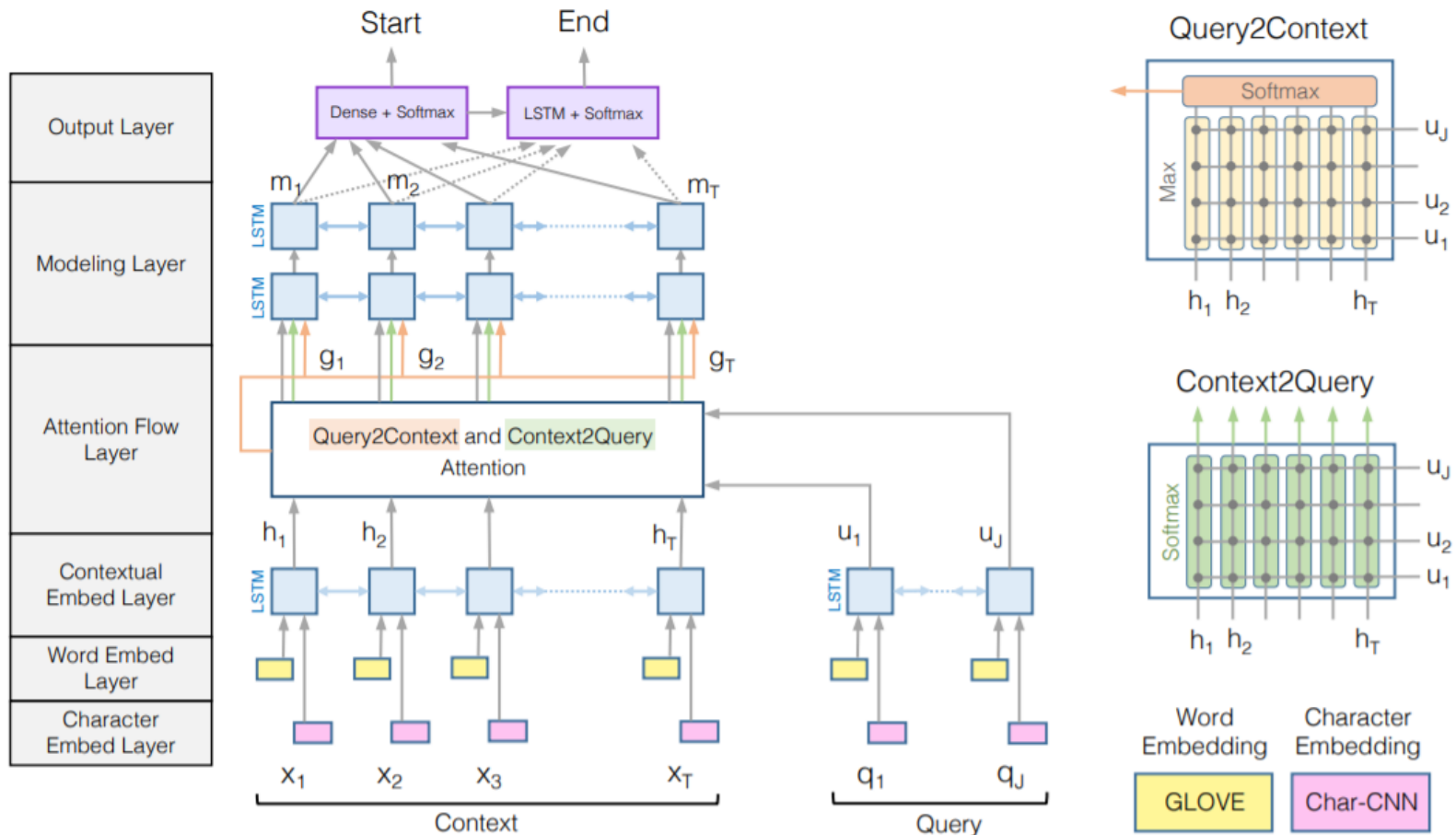
# How models solve SQuAD

## 1. Stanford Attentive Reader (Neural Reading Comprehension and Beyond)

Method	Dev		Test	
	EM	F1	EM	F1
Logistic regression (Rajpurkar et al., 2016)	40.0	51.0	40.4	51.0
Match-LSTM (Wang and Jiang, 2017)	64.1	73.9	64.7	73.7
RaSoR (Lee et al., 2016)	66.4	74.9	67.4	75.5
DCN (Xiong et al., 2017)	65.4	75.6	66.2	75.9
BiDAF (Seo et al., 2017)	67.7	77.3	68.0	77.3
<b>Our model (Chen et al., 2017)</b>	69.5	78.8	70.0	79.0
R-NET (Wang et al., 2017)	71.1	79.5	71.3	79.7
BiDAF + self-attention (Peters et al., 2018)	N/A	N/A	72.1	81.1
FusionNet (Huang et al., 2018b)	N/A	N/A	76.0	83.9
QANet (Yu et al., 2018)	73.6	82.7	N/A	N/A
SAN (Liu et al., 2018)	76.2	84.1	76.8	84.4
BiDAF + self-attention + ELMo (Peters et al., 2018)	N/A	N/A	78.6	85.8
BERT (Devlin et al., 2018)	84.1	90.9	N/A	N/A
Human performance (Rajpurkar et al., 2016)	80.3	90.5	82.3	91.2

# How models solve SQuAD

## 2. BiDAF (Bi-Directional Attention Flow for Machine Comprehension)



# How models solve SQuAD

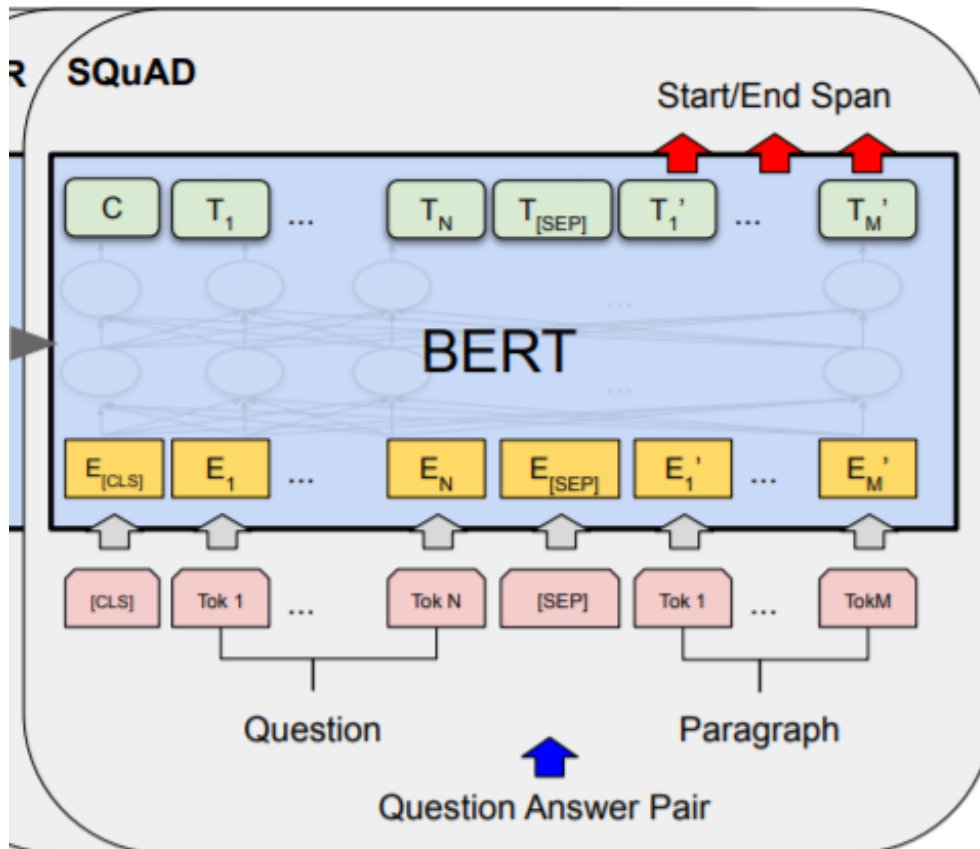
## 2. BiDAF (Bi-Directional Attention Flow for Machine Comprehension)

	Single Model		Ensemble	
	EM	F1	EM	F1
Logistic Regression Baseline <sup>a</sup>	40.4	51.0	-	-
Dynamic Chunk Reader <sup>b</sup>	62.5	71.0	-	-
Fine-Grained Gating <sup>c</sup>	62.5	73.3	-	-
Match-LSTM <sup>d</sup>	64.7	73.7	67.9	77.0
Multi-Perspective Matching <sup>e</sup>	65.5	75.1	68.2	77.2
Dynamic Coattention Networks <sup>f</sup>	66.2	75.9	71.6	80.4
R-Net <sup>g</sup>	<b>68.4</b>	<b>77.5</b>	72.1	79.7
BiDAF (Ours)	68.0	77.3	<b>73.3</b>	<b>81.1</b>



# How models solve SQuAD

## 3. BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding)



以 “[CLS]问题[SEP]文章” 的形式对整个样例进行编码，并在最后一层输出文章的各个位置是问题答案开头/结尾的概率

# How models solve SQuAD

## 3. BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding)

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT <sub>BASE</sub> (Single)	80.8	88.5	-	-
BERT <sub>LARGE</sub> (Single)	84.1	90.9	-	-
BERT <sub>LARGE</sub> (Ensemble)	85.8	91.8	-	-
BERT <sub>LARGE</sub> (Sgl.+TriviaQA)	<b>84.2</b>	<b>91.1</b>	<b>85.1</b>	<b>91.8</b>
BERT <sub>LARGE</sub> (Ens.+TriviaQA)	<b>86.2</b>	<b>92.2</b>	<b>87.4</b>	<b>93.2</b>



# 数据集解析

```
[
  # 整个数据集用json格式组织，是一个序列，序列的每个元素是一个字典，代表一篇文章
  {
    'title': 'Super_Bowl_50', #这是文章标题，与本任务无关
    'paragraphs':[ #pargaraphs是文章的正文部分，每个paragraph也是一个序列，其中的每个元素代表一个context，及对应的问题
      {
        'context': 'Super Bowl 50 was an American football...', #这是一段context
        'qas':[ # 关于这段context的所有问题-答案对
          {
            # 答案，answer_start表示答案在context中的起始位置是第几个单词，text 表示答案的文本
            'answers': [{ 'answer_start': 515, 'text': 'Saint Bernadette Soubirous' }],
            # 问题的文本
            'question': 'To whom did the Virgin Mary allegedly appear in 1858 in Lourdes
                          France?',
            # 这个问题的 id
            'id': '5733be284776f41900661182'}
          }, ...]
        }, ....]
  }, .....]
```

# 大作业要求

- 使用创新的深度学习算法在SQuDA v1.1数据集上进行训练和评估
- 使用数据集train-v1.1.json进行模型训练和超参数调优
- 在数据集dev-v1.1.json上预测结果，并使用evaluate-v1.1.py中代码进行评估，记录在验证集上算法的分数
- 预测结果需转换为dict类型，其中每个元素的键为问题的id，值为该问题的答案文本，需要输出为json格式文件
- 请在train-v1.1.json文件上自行划分训练集和验证集，严禁使用dev-v1.1.json上的数据训练
- 撰写报告，在报告中记录使用算法的原理、背景，模型训练方法，模型最终得分以及其它必要的实验结果

# 大作业要求

- Deadline: July 26,2020;
- 创新
  - 新方法、新思路;
- 抄袭（网络、同学）：0分;

# Thank you!

权小军 中山大学数据科学与计算机学院