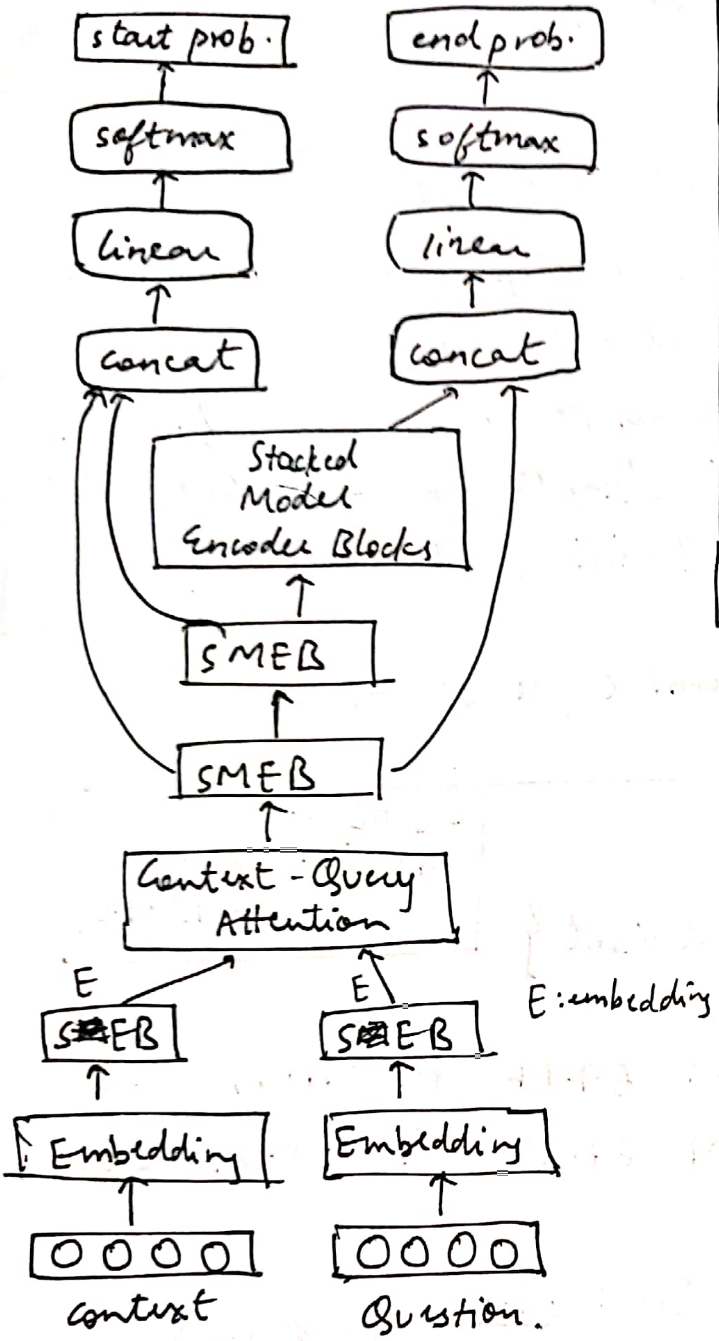
QANet.

Model



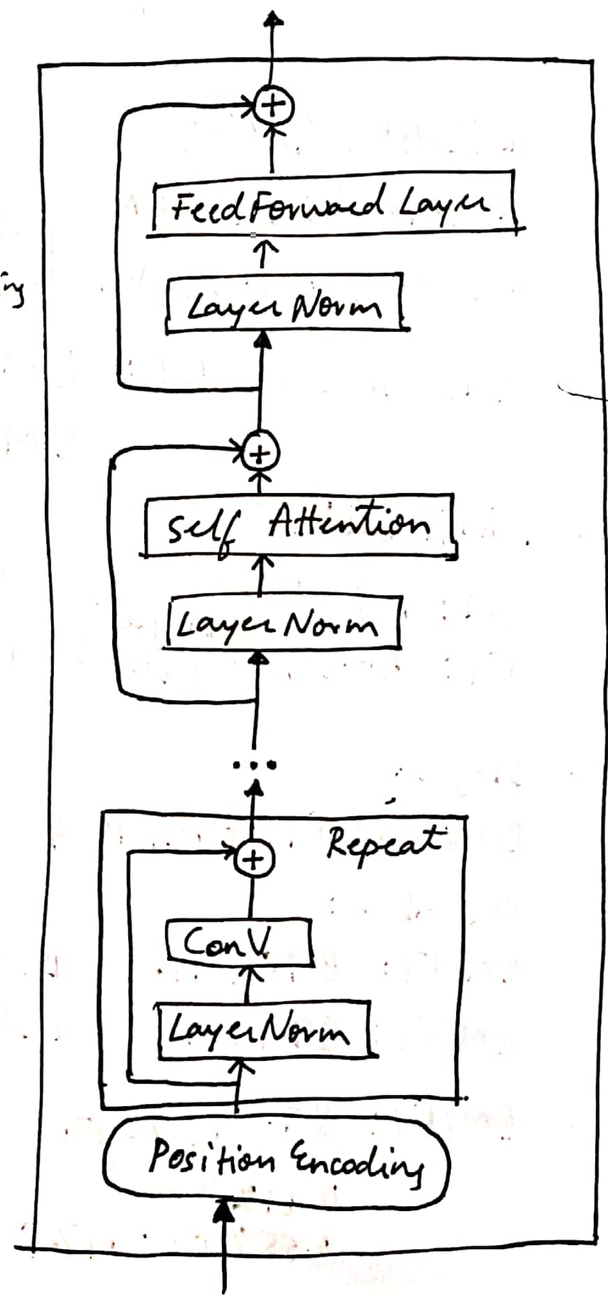
Model Layers:

1. Input Embedding L.
2. Embedding Encoder L.
3. Content-Query Attention L.
4. Model Encoder L.
5. Output L.

- 77.0 F1 score → 3hrs training
  (BiDAF: 15hrs)
- 82.7 F1 (dev) → 18hrs.

- SQuAD → 84.6 F1 (test)
  (81.8 SoTA 2017)
- 2018:
  SQuAD EM/F1
  82.2/88.6 (single)
  83.9/89.7 (ensemble)
  82.3 (human)

One Encoder Block.



QANET

context paragraph : $C = \{c_1, \dots, c_n\}$

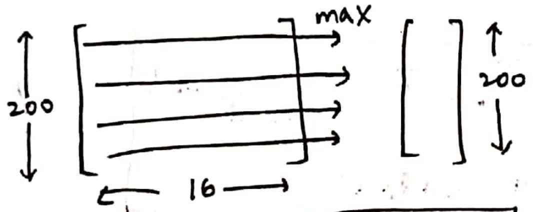
query sentence : $Q = \{q_1, \dots, q_m\}$

o/p span : $S = \{c_i, \dots, c_{i+j}\}$

$\boxed{x \in C, Q}$

① Input Embedding Layer:

- word emb : fixed
  - $P_1 = 300$ D (GloVe)
  - <unk> → trainable c̄ random init
- char emb :
  - $P_2 = 200$ D
  - max-word-len = 16
  - <pad>



- $\boxed{x = [x_w ; x_c] \in \mathbb{R}^{P_1+P_2} \in \mathbb{R}^{500}}$

② Embedding Encoder Layer

- stack of the foll^n block :
  [conv layer x # + self attention layer + feed forward layer]
- depthwise separable conv. (mem. efficient)
  k = 7 # conv layers within 1 block = 4.
  d = 128 filters
  } conv.
- multiheaded attention
  - query, key ~~inside a residual block~~.
  - no. of heads = 8
  } self atten.
- conv / attent / feed forw
  - inside a residual block
  - f(layernorm(x) + x)
  - no of encoder blocks = 1
  } General
- i/p = 500 D → 1D conv → 128 D *
  o/p = 128 D
  } i/p o/p

③ $\boxed{\text{Context-Query Attention Layer.}}$

- Similarity matrix $S \in \mathbb{R}^{n\times m}$: sim b/w $\underset{(n)}{C}$ & $\underset{(m)}{Q}$.
- $\boxed{S \xrightarrow{\text{softmax}} \bar{S}}$ (row normalized)
  $\Rightarrow$ dim=1
- Attention (C→Q):

$$\boxed{A = \underset{(n\times m)}{\bar{S}} \cdot \underset{(m\times d)}{Q^T} \in \mathbb{R}^{n\times d}}$$

- Similarity fn : (trilinear)

$$\boxed{f(q,c) = W_0[q, c, q\odot c]}$$

* Query – Context Attention.

- (DCN)
- column normalized $\boxed{S \xrightarrow{\text{soft}} \bar{\bar{S}}}$ $\Rightarrow$ dim = 0

$$\boxed{B = \underset{(n\times m)}{\bar{S}} \cdot \underset{(m\times n)}{\bar{\bar{S}}^T} \cdot \underset{(n\times d)}{C^T} \in \mathbb{R}^{n\times d}}$$

? $(m\times d)$

④ $\boxed{\text{Model Encoder Layer}}$

- i/p : $[c, a, c\odot a, c\odot b]$ a, b : row of A & B
- parameters same as before except:
  #conv layers = 2 in a block
  #blocks = 7

⇒ weights are shared b/w each of the 3 repitions of this Layer.

⑤ $\boxed{\text{Output Layer}}$

$$\boxed{p^1 = \text{softmax}(W_1[M_0; M_1]) \quad p^2 = \text{softmax}(W_2[M_0; M_2]).}$$

$$\boxed{\text{score} = p^1 \times p^2}$$

$$\boxed{L(\theta) = -\frac{1}{N}\sum_{i}^{N}\left[\log(p^1_{y_i^1}) + \log(p^2_{y_i^2})\right]}$$

M_2
M_1
M_0

Inference : $p^1_s p^2_e \longrightarrow$ max $s \leq e$.
(DP can solve in linear time)

EXPERIMENTS

(SQuAD 1)

- **Data Preprocessing**
  - NLTK tokenizer
  - max context len = 400
    (para longer → discard)
  - max ans len = 30
    (300-D GLoVe)

- <PAD> : short
- <UNK> → trained
- 200 D char emb → trained

- **Training Details**

  ① L2 weight decay
  → all trainable variables
  → $\lambda = 3 \times 10^{-7}$

  ② Dropout
  → word emb } ~~0.1~~ → 0.1
  → char emb → 0.05
  → b/w layers → 0.1

- Stochastic Depth Method (layer dropout)
  → within each embedding / model encoder layer.
  → sublayer l : survival prob

$$\boxed{P_l = 1 - \frac{l}{L}(1 - P_L)}$$

$P_L = 0.9$ L = last layer.

- hidden size } 128
  conv filter }

- batch size = 32   steps = 150K
- conv : emb = 4   k = 7   # = 1
- conv : model = 2   k = 5   # = 7

- ADAM
  $\beta_1 = 0.8$ $\varepsilon = 10^{-7}$
  $\beta_2 = 0.999$
  - LR warm up scheme
    0.0 → 0.001 (1000 steps)
    then constant.

- Exponential moving avg is applied on all trainable variables
  decay rate = 0.9999.

NVIDIA p100 GPU.

	EM	F1
no data aug :	73.6 /	82.7

(SQuAD 1)