Project routemap

- Reproduce the results

- Modify the model to improve the performance

Reproduction

- Reference: *Multi-label classification of patient notes a case study on icd code assignment*

- Problem: assign ICD9 codes to patient's **discharge summaries**.

- Dataset: MIMIC II (22,000+) and MIMIC III (65,000+)

- Labels: 4,000+ and 6,000+ ICD9 codes respectively

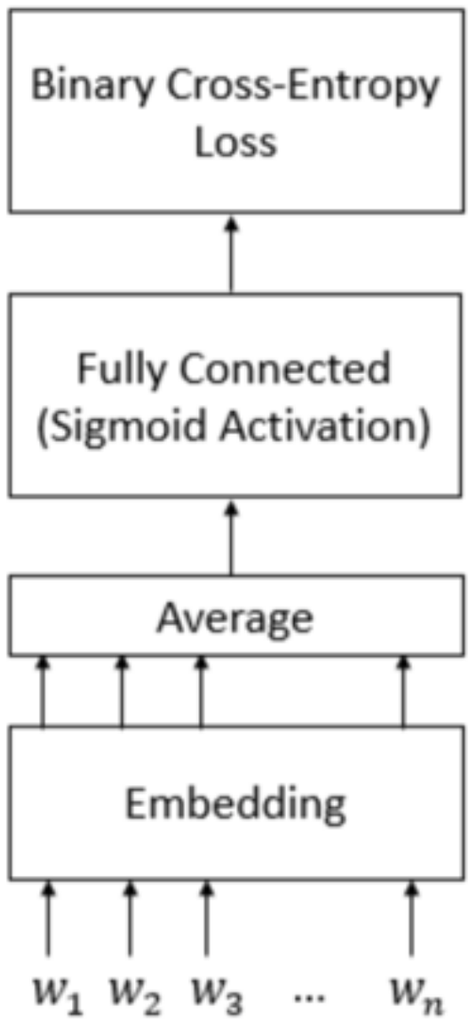
- Evaluation: Macro-F

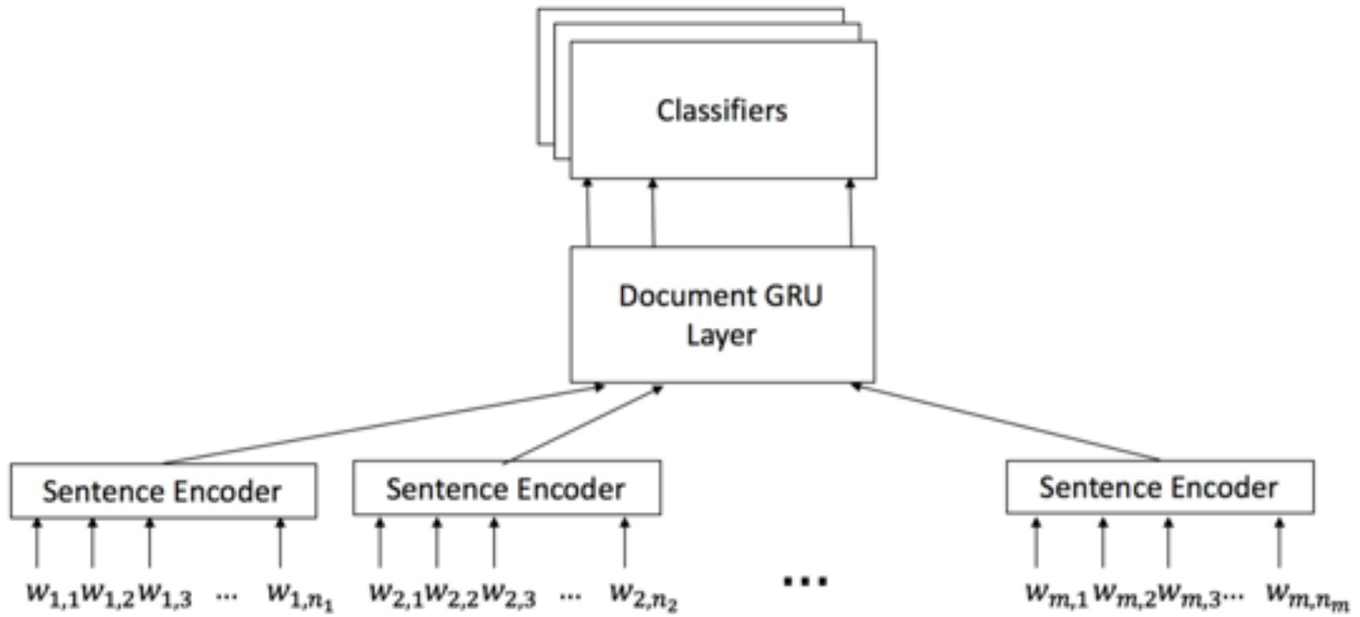
Preprocessing

- Data mainly comes from table **DIAGNOSE_ICD** and **NOTEEVENT**
- Case-sensitive tokenisation
  - How about case-insensitive?
- Digits normalisation
  - Do the numerical values matter? How about non-linear transformation based on numerical values' context?
- Map OOVs to its nearest neighbour by the edit distance
  - How about representing OOV by UNK?
- Hierarchical segmentation: tokens & sentences (LONG DOCUMENTS)
  - How about discourse segmentation

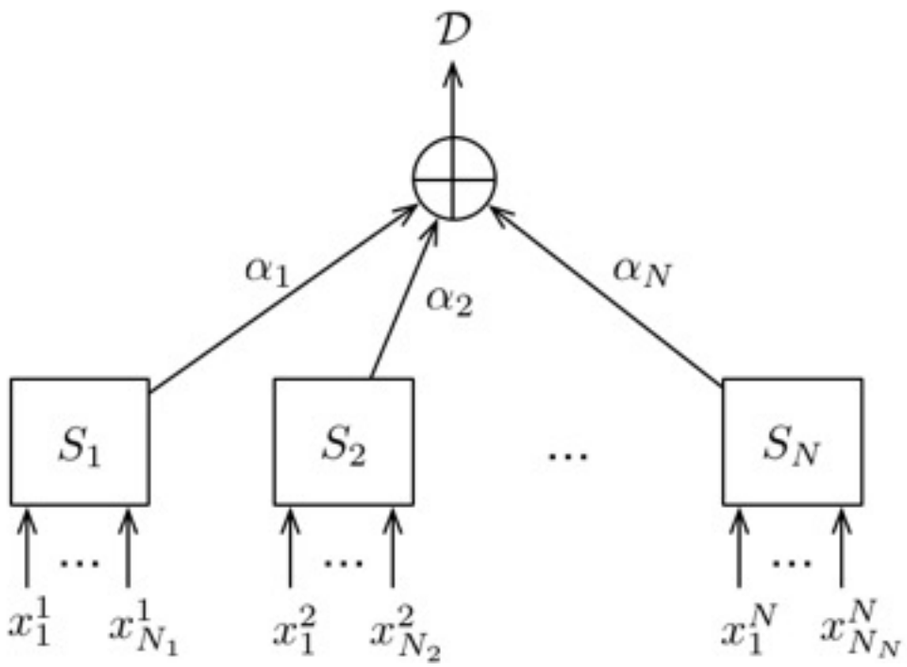
Models

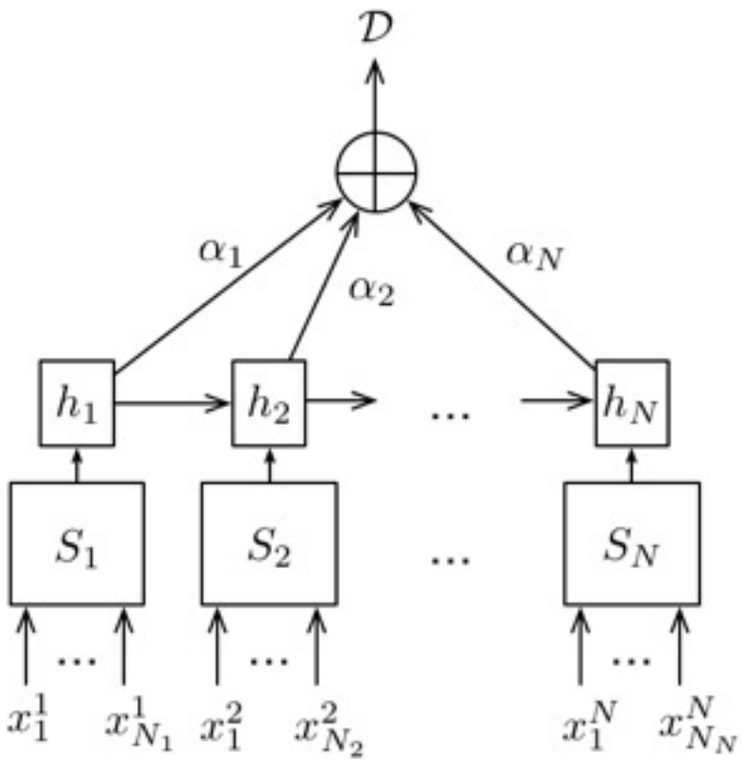
CBOW


Binary Cross-Entropy Loss
Fully Connected (Sigmoid Activation)
Average
Embedding
w_1 w_2 w_3 ... w_n


Classifiers
Document GRU Layer
Sentence Encoder
Sentence Encoder
Sentence Encoder
$w_{1,1}w_{1,2}w_{1,3}$ ... w_{1,n_1}
$w_{2,1}w_{2,2}w_{2,3}$ ... w_{2,n_2}
...
$w_{m,1}w_{m,2}w_{m,3}$... w_{m,n_m}

Our models


$\mathcal{D}$
α_1
α_2
α_N
S_1
S_2
...
S_N
x_1^1 $x_{N_1}^1$ x_1^2 $x_{N_2}^2$ x_1^N $x_{N_N}^N$


$\mathcal{D}$
α_1
α_2
α_N
h_1
h_2
...
h_N
S_1
S_2
...
S_N
x_1^1 $x_{N_1}^1$ x_1^2 $x_{N_2}^2$ x_1^N $x_{N_N}^N$