Lec 4 Sequence Modelling

Motivation

Lots of seqs in NLP
- words
- characters
- documents

Long-distance dependencies
- gender agreement
- semantic/factual dependencies
- what 'it' refers to

Winograd schema ← linguistic challenge with minimal pairs

trophy won't fit in bag, it is too small
trophy won't fit in bag, it is too big

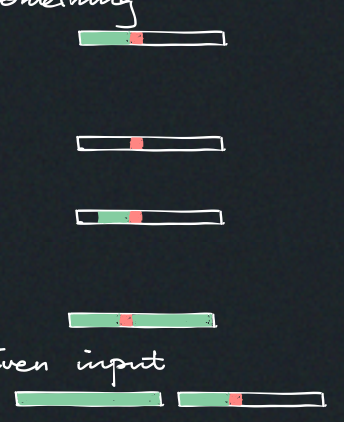
→ Figurative language

Sequential prediction problems

- Binary
- Multi-class
- Structured
  - part of speech labelling
    I hate this movie → PRP VBP DT NN
  - translation
    I hate this movie → kono eiga ga kirai

predicting
condition

- Unconditioned – generate something
  - left-to-right autoregressive
    - RNN
    - Trans LM
  - independent
    - unigram
  - left-to-right Markov
    - n-grams
  - bidirectional
    - masked LM
- Conditioned – generate given input
  - autoregressive
  - non-autoregressive



Paradigm: extract feature → predict

▷ Sequence Labelling

$X \to Y$ s.t. $|Y| = |X|$

- Part of speech
- Lemmatisation
- Morphological tagging
  saw → tense: past

{ stemming only chops things off

▷ Span Labelling

- NER
  Carnegie Mellon University → ORG
- Syntactic Chunking
  VP, NP, etc.
- Semantic Role Labelling
  actor, predicate, location

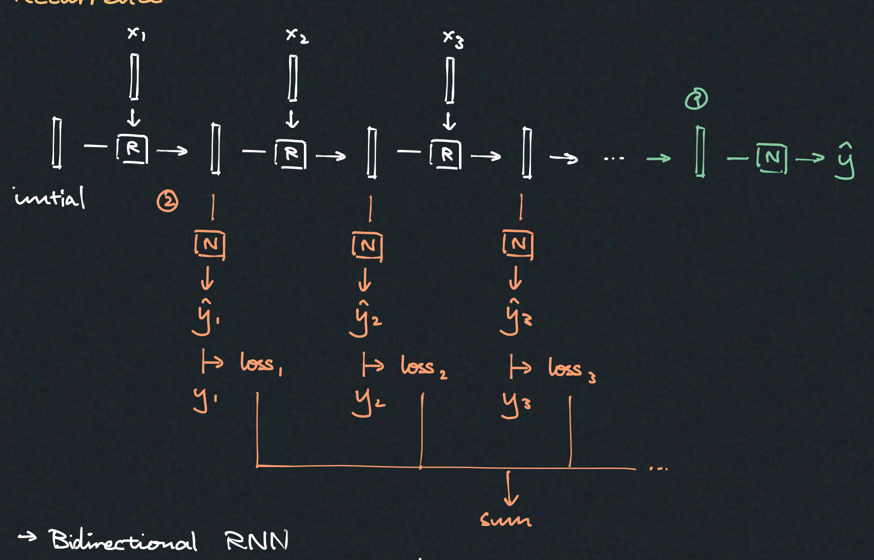
This can reduce to tagging by labelling tkns with Begin | In | Out

Sequence Labels

For seq length n

- Recurrence – condition on history
  $O(n)$, sequential
- Convolution – convolve on local context with window size w
  $O(nw)$, parallelisable
- Attention – weighted average of all tkns
  $O(n^2)$, parallelisable

Recurrence



→ Bidirectional RNN
→ Vanishing / exploding gradient if long recurrence
  - try to put more direct path from important info to loss and unimportant ones farther
  - residual connections in other situations

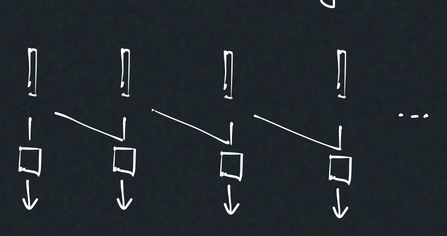
LSTM – useful for very long sequences

Idea: addictive input between every time step

Convolution – useful in speech / image proc

↑ many frames ↑ many pixels

→ Convolution auto regressive

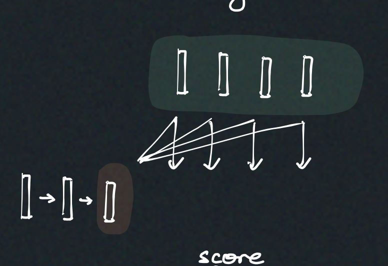


Attention

- Cross attention: attend to another seq
- Self attention: attend to context in same seq

Calculation:

query := current decoder state
key := encoder states
values := things to combine together



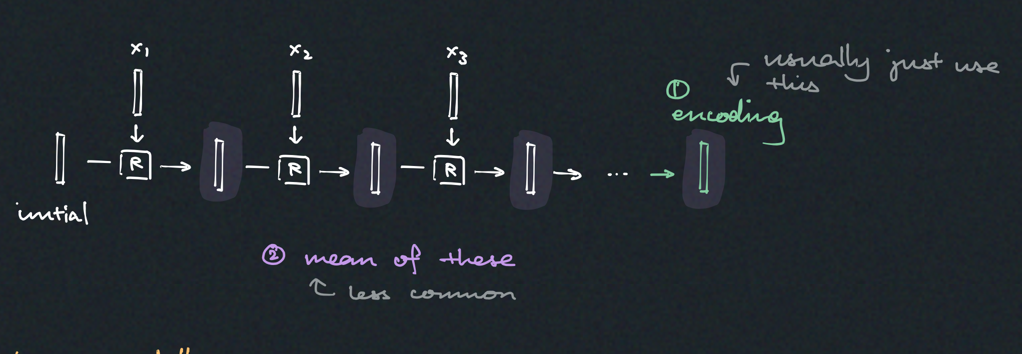
key, query $\xrightarrow{\text{score func}}$ score $\xrightarrow{\text{softmax}}$ weight
values

The score func a:
- MLP (original paper) — expressive, expensive
  $$a(q,k) = w_2^T \tanh(w_1[q;k])$$
- bilinear $a(q,k) = q^T W k$
- dot product $a(q,k) = q^T k$
- scaled dot prod $a(q,k) = \frac{q^T k}{\sqrt{|k|}}$ ← normalise

⋮

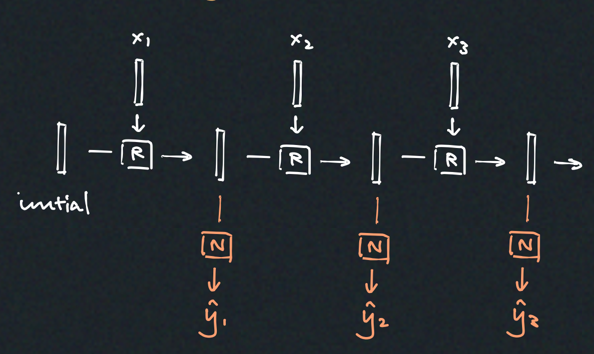
Note we sometimes want the model to not look at future information. We can make attention mask to make attention score to future token $-\infty$ (softmax turns them 0)

▷ Seq encoding



② mean of these
↑ less common

▷ Lang modelling



← use these, autoregressive

Usually:
- RNN could be harder to change if long distance
- attention has direct connection so learns long distance quickly
- attention models becomes sequential when generating

Exists: translation with strong Trans encoder + fast RNN decoder

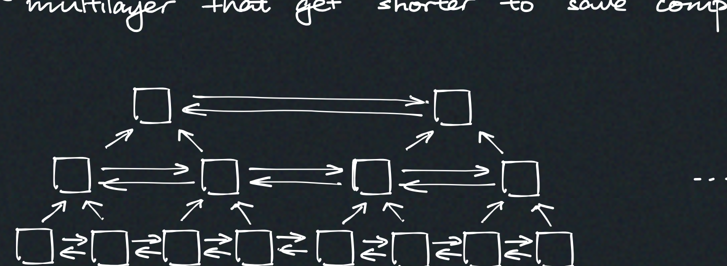
Efficiency Considerations

Sequences have different length :C

→ Padding and masking
  - when calculating loss, multiply pad locs by 0

→ Bucketing / sorting
  - put similar length in same batch
  - BUT this disrupts random data distribution

→ Strided Archi aka Pyrimital RNN aka sparse attention
  - multilayer that get shorter to save compute



→ Truncated BPTT
  - break into smaller seqs, subsequent passes take prev pass out
  - backprop into shorter distances