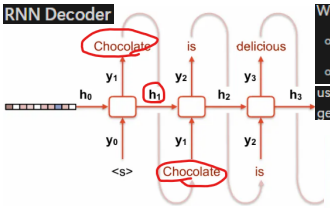
RNN Decoder

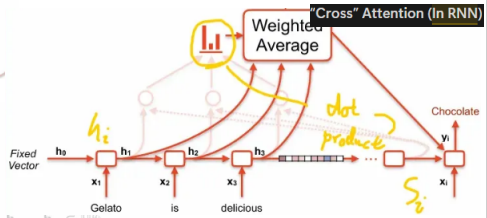


We can also use Greedy Inference with an sequence to sequence model

- How to stop beam search encoder-decoder?
  Keep going until we have N outputs.
- How to score diffrent output with different length?
  usually longer output will have higher length, so we need to normalize the score, to get the average probability per word

We will solve this with 'attention'

Give each output step
a representation of the input
that is most useful for that output decision



Hidden vectors from input $h_1, h_2, h_3 \dots h_N \in \mathbb{R}^{d_h}$

Hidden vector for current output step $s_i \in \mathbb{R}^{d_s}$

Calculate attention scores $e_i = [s_i^T h_1, s_i^T h_2, s_i^T h_3, \dots s_i^T h_N] \in \mathbb{R}^N$

This is dot product attention

Normalise s ax(e_i) $\in \mathbb{R}^N$

Calculate weighted average $a_i = \sum_{j=1}^{N} \alpha_{i,j} h_j \in \mathbb{R}^{d_h}$

chrF — compare character ngrams

We compare TP/TN/FP/FN for all n-gram
FP = FP-1-gram +...+ FP-4-gram

$$F_\beta\text{-score} = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP}$$

smaller β means colser to Precision
larger β means closer to recall

Robust to typos

BLEU — compare word ngrams

Translation quality

1. Calculate precision for unigrams, bigrams, trigrams, 4grams separately
2. Take the geometric mean
3. Apply a penalty if the output is short

Strongly depends on word order
no 4grams matches means a score of 0!

What if sentence cannot be split on whitespace to get token?
start with small units, then combine units

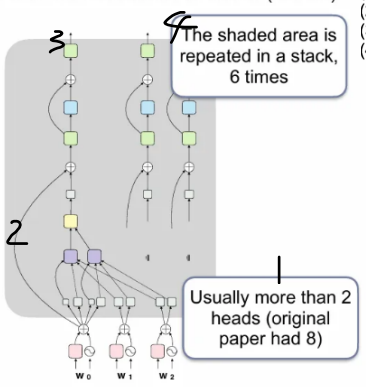
Byte-Pair Encoding (BPE)
1. Set vocabulary to be all characters
2. Find the two vocabulary items that are adjacent most frequently,
3. Create a new vocabulary item for that pair and update data
4. Check if the vocabulary is size K (e.g., 100,000). If not, go to 2.

WordPiece Consider all possible vocabulary pairs, then Choose the one that will decrease perplexity most if added to the model

start with big and small units, then delete units | more meaningful subword units

1. Set vocabulary to be all characters and all frequent character sequences, up to full words.
2. Find vocabulary items to remove using a complex method.
3. Repeat until the vocabulary gets down to the desired size.

This is the Transformer encoder! (Almost)

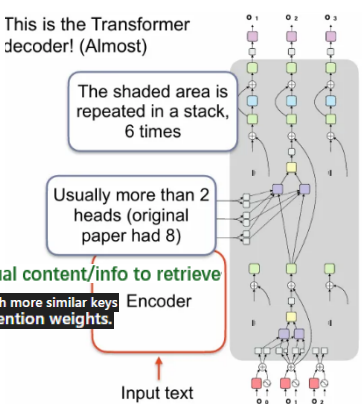


(1) Do attention multiple times at once
(2) Add residual connections
(3) Add layer normalisation
(4) Stack the architecture multiple times

Feed Forward $\max(0, xW_1 + b_1)W_2 + b_2$
Layer Norm
Element-wise sum
Concatenate
Attention: $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
Linear transform
Look up vector
Position as vector

vector with actual content/info to retrieve
Values associated with more similar keys receive higher attention weights.
Encoder

This is the Transformer decoder! (Almost)



Softmax
Feed Forward $\max(0, xW_1 + b_1)W_2 + b_2$
Layer Norm
Element-wise sum
Concatenate
Attention: $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
Linear transform
Look up vector
Position as vector

Layer normalisation helps stabilise training by keeping the output of each layer at a consistent scale (mean and standard deviation are consistent between layers).

Fast Train | Internal Covariate Shift

slight accuracy gains, big efficiency gains.
Faster training means they can scale

BERT Encoder
GPT-1 Decoder

K and V come from encoder
Q comes from decoder

Scaling Laws

estimate the best model size given a certain amount of data and a budget for training

Attention is quadratic, is that a problem?
Lots of clever variants to avoid it!

- PerFormer
- LongFormer
- LinFormer

memory bandwidth

Encoder — BERT, ELECTRA

- masked-token prediction — ***BERT*** Ignore output for unmasked words
- span prediction — SpanBERT mask subword of word or a countinuous token set
- **random-token correction — BERT** correct wrong token (model is not told which token was changed)
- combination of masked-token + random-token
- token edit detection Edits are not random use a small LLM to generate hard edits
- next sentence prediction — BERT judge next sentence (NSP) task.

RoBERTa outperforms BERT by training longer, using more data, and removing the

Encoder-Decoder — T5, BART

- masked sequence prediction — ***BART*** mask might include multiple token
- masked span prediction — ***T5*** mask cannot be empty, include multiple token
- deleted sequence prediction — BART or T5 work out where deletion happened, as well as what was deleted
- permuted sequence prediction — BART rearrange sentence
- Rotated Sequence Prediction — BART To enhance the model's ability to understand changes in sentence order or structure.
- infilling sequence prediction — BART The mask might be empty

Decoder — ELMo, GPT

- next-token prediction — ***GPT***

Advantages	Disadvantages
- Captures full bidirectional context	- Not suited for text generation
- Good for understanding tasks	- No autoregressive decoding

Making a prediction for every word faster training

- Captures full input meaning	- More complex
- Flexible input-output mapping	- Slower inference (2 modules)
- Powerful for seq2seq	- Requires more memory

Training (fine-tuning)

- Done many times
- Can be short or long training
- Can be on limited compute or a lot
- Task specific data

- Autoregressive: good for generation	- Unidirectional: only left context
- Simple architecture	- Not good for understanding tasks

Mean Reciprocal Rank (MRR)
earlier the correct answer appears in the ranking, the higher the score.

pre-training and training

Pre-training

- done once
- Long training
- Trained on lots of GPUs
- Input + Output are based on raw text

Sentiment (Classification) — all
NER (Token predictions) — En & De
Coreference (Structured Output) — Encoder-Decoder
Summarisation (Generation) — Decoder
Translation (Generation, in another language) — Decoder

In-Context Learning ICL
Low perplexity prompts are better
Order of examples matters
unsorted examples are best
more number of shots N → better performance
provide instructions and examples in the input
any label is acceptable
have similar performance in the end
Bad explanations usually lead to the wrong answer!

Url Filtering
Trafilatura text extraction on the raw HTML
FastText LanguageFilter removing any document with en language score lower
Wrong answers that have many matching words
whether a model can do well without the question and/or context
Quality filtering
Minhash deduplication deduplicate
PII Formatting anonymize email and public IP addresses
Train a model-based detector — overpredict/overfilter

Risk of dataset
- Avoid overfitting into one dataset
- Shortcuts / Validity minimal edits true answer make it wrong
- Does our task have statistical power?
- Social Bias
- Is this a task we want to create?

correcting for chance agreement
Cohen's Kappa 0.80-1.0 good
Unclear guidelines? Careless annotators? Ambiguous cases?

Instruction Tuning FLAN-T5 higher generalization Alpaca LIMA
no clear pattern bigger model benefits more from instruction fine tuning
During instruction fine-tuning: Perplexity may increase instead of decreasing.
Filtered data produce better generation quality high quality
Increase training examples does not improve generation quality

Efficiency of LLM

Reducing Model Size

- Distillation use a big model to train a small model
- Adding sparsity removing unnecessary features or connections.

doesn't help on GPUs because sparse

Increasing training memory efficiency

- Parameter-Efficient Fine Tuning (PEFT) — LoRA

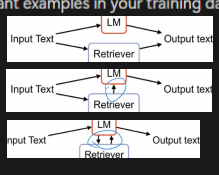
Numerical approximation Reduce weight matrix size

- Quantization — LLM.int8 Outlier 16-bit reduced numerical precision version

Mixtures of models to less important weights

- SMoE divide the model into multiple smaller "expert"

1. Treating retrieval as an LM itself and doing model ensembling
   a. kNN-LM Larger k values generally lead to lower perplexity. Larger datastore size makes relying on external context more reliable. Smaller λ means more dependence on the LM's internal (in-context) information.
2. Provide the retrieved content in the prompt
   Retrieval ICL = ICL + auto-select relevant examples for each input
   a. Retrieval ICL Find the most relevant examples in your training data and use them in the prompt.
   b. RAG
   c. RETRO
   d. REALM
   e. FLARE



Retrieval Augmentation

3. Interactive methods (related to Agents, which we will see later)

Preference Optimization

RLHF Approach Without making too big change reward model
DPO directly train from preference data
Avoids training a reward model, which may be unreliable

Agent
Reasoning: happened before generating prediction
Self-consistency voting with multiple outputs
Refletion: consider about previous part
Tree of thoughts exploring multiple possible "thought paths"
Acting: doing something beyond the generation; RAG