CS378: Natural Language Processing
Lecture 19: MT 3, Transformers

Greg Durrett

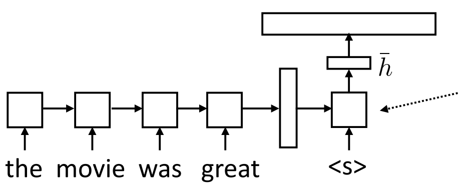
TEXAS
The University of Texas at Austin

Announcements

- A3 back soon
- A4 due today
- A5 released today
- Final project released next week

Recall: Seq2seq Model

- Generate next word conditioned on previous word as well as hidden state
- W size is |vocab| x |hidden state|, softmax over entire vocabulary



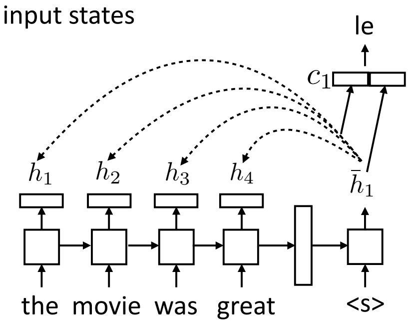
$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W\bar{h})$$

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1})$$

Decoder has separate parameters from encoder, so this can learn to be a language model (produce a plausible next word given current one)

Recall: Attention

- For each decoder state, compute weighted sum of input states



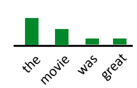
- No attn: $P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W\bar{h}_i)$

$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W[c_i; \bar{h}_i])$$

$$c_i = \sum_j \alpha_{ij} h_j$$

- Weighted sum of input hidden states (vector)

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})}$$



$$e_{ij} = f(\bar{h}_i, h_j)$$

- Some function f

Neural MT

Results: WMT English-French

- 12M sentence pairs

Classic PBMT system: ~**33** BLEU, uses additional target-language data

PBMT + rerank w/LSTMs: **36.5** BLEU (long line of work here; Devlin+ 2014)

Sutskever+ (2014) seq2seq single: **30.6** BLEU (input reversed)

Sutskever+ (2014) seq2seq ensemble: **34.8** BLEU

Luong+ (2015) seq2seq ensemble with attention and rare word handling: **37.5** BLEU

- But English-French is a really easy language pair and there's *tons* of data for it! Does this approach work for anything harder?

Results: WMT English-German

- 4.5M sentence pairs

Classic phrase-based system: **20.7** BLEU

Luong+ (2014) seq2seq: **14** BLEU

Luong+ (2015) seq2seq ensemble with rare word handling: **23.0** BLEU

- Not nearly as good in absolute BLEU, but BLEU scores aren't really comparable across languages
- French, Spanish = easiest
  German, Czech = harder
  Japanese, Russian = hard (grammatically different, lots of morphology…)

MT Examples

src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
ref	However , in an interview , Bloom has said that he and *Kerr* still love each other .
best	In an interview , however , Bloom said that he and *Kerr* still love .
base	However , in an interview , Bloom said that he and **Tina** were still <unk> .

- best = with attention, base = no attention
- NMT systems can hallucinate words, especially when not using attention — phrase-based doesn't do this

Luong et al. (2015)

MT Examples

src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The *austerity imposed by Berlin and the European Central Bank , coupled with the straitjacket* imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict *austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket* in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure **imposed by the European Central Bank and the Federal Central Bank with the strict austerity** imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

- best = with attention, base = no attention

Luong et al. (2015)

Handling Rare Words

- Words are a difficult unit to work with: copying can be cumbersome, word vocabularies get very large
- Character-level models don't work well
- Compromise solution: use thousands of "word pieces" (which may be full words but may also be parts of words)

Input: _the **_eco tax** _port i co _in _Po nt - de - Bu is ...

Output: _le _port ique **_éco taxe** _de _Pont - de - Bui s

- Can achieve transliteration with this, subword structure makes some translations easier to achieve

Sennrich et al. (2016)

Byte Pair Encoding (BPE)

- Start with every individual byte (basically character) as its own symbol

```
for i in range(num_merges):
  pairs = get_stats(vocab)
  best = max(pairs, key=pairs.get)
  vocab = merge_vocab(best, vocab)
```

- Count bigram character cooccurrences
- Merge the most frequent pair of adjacent characters
- Doing 8k merges => vocabulary of around 8000 word pieces. Includes many whole words
- Most SOTA NMT systems use this on both source + target

Sennrich et al. (2016)

Byte Pair Encoding (BPE)

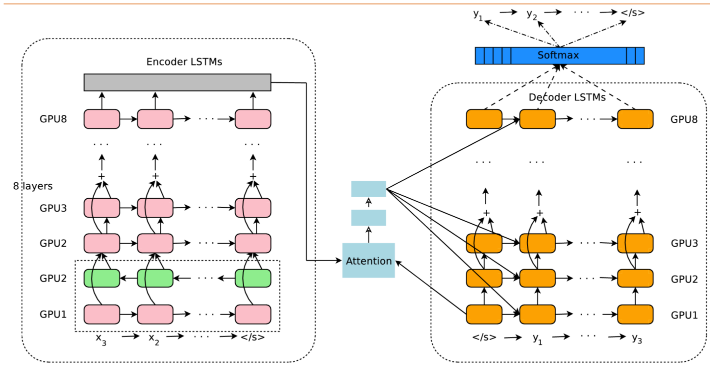
(a)	**Original:**	furiously	(b)	**Original:**	tricycles
	BPE:	_fur | iously		**BPE:**	_t | ric | y | cles
	Unigram LM:	_fur | ious | ly		**Unigram LM:**	_tri | cycle | s
(c)	**Original:**	Completely preposterous suggestions			
	BPE:	_Comple | t | ely | _prep | ost | erous | _suggest | ions			
	Unigram LM:	_Complete | ly | _pre | post | er | ous | _suggestion | s			

- BPE produces less linguistically plausible units than another technique based on a unigram language model

Bostrom and Durrett (2020)

Google NMT

Google's NMT System (2016)



- 8-layer LSTM encoder-decoder with attention, word piece vocabulary of 8k-32k

Wu et al. (2016)

Google's NMT System (2016)

English-French:

Google's phrase-based system: 37.0 BLEU

Luong+ (2015) seq2seq ensemble with rare word handling: 37.5 BLEU

Google's 32k word pieces: 38.95 BLEU

English-German:

Google's phrase-based system: 20.7 BLEU

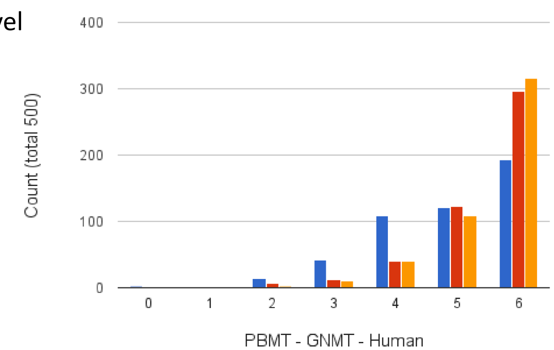
Luong+ (2015) seq2seq ensemble with rare word handling: 23.0 BLEU

Google's 32k word pieces: 24.2 BLEU

Wu et al. (2016)

Human Evaluation (En-Es)

- Similar to human-level performance *on English-Spanish*



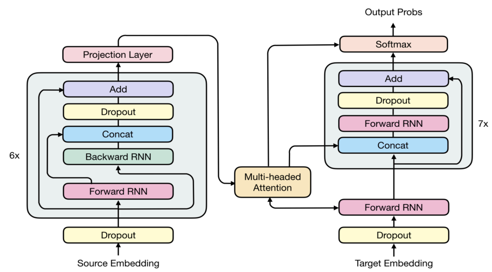
Wu et al. (2016)

Updated Version

- RNMT+ model — better RNNs in response to Transformers

Model	RNMT+	Trans. Big
Baseline	41.00	40.73
- Label Smoothing	40.33	40.49
- Multi-head Attention	40.44	39.83
- Layer Norm.	*	*
- Sync. Training	39.68	*

Table 4: Ablation results of RNMT+ and the Transformer Big model on WMT'14 En → Fr. We report average BLEU scores on the test set. An asterisk '*' indicates an unstable training run (training halts due to non-finite elements).



- RNMT+ is a bit better than Transformers, but also uses multi-head attention

Chen et al. (2018)

Frontiers in MT: Small Data

ID	system	BLEU 100k	BLEU 3.2M
1	phrase-based SMT	15.87 ± 0.19	26.60 ± 0.00
2	NMT baseline	0.00 ± 0.00	25.70 ± 0.33
3	2 + "mainstream improvements" (dropout, tied embeddings, layer normalization, bideep RNN, label smoothing)	7.20 ± 0.62	31.93 ± 0.05
4	3 + reduce BPE vocabulary (14k → 2k symbols)	12.10 ± 0.16	-
5	4 + reduce batch size (4k → 1k tokens)	12.40 ± 0.08	31.97 ± 0.26
6	5 + lexical model	13.03 ± 0.49	31.80 ± 0.22
7	5 + aggressive (word) dropout	15.87 ± 0.09	**33.60** ± 0.14
8	7 + other hyperparameter tuning (learning rate, model depth, label smoothing rate)	**16.57** ± 0.26	32.80 ± 0.08
9	8 + lexical model	16.10 ± 0.29	33.30 ± 0.08

- Synthetic small data setting: German -> English

Sennrich and Zhang (2019)

Frontiers in MT: Low-Resource

- Particular interest in deploying MT systems for languages with little or no parallel data

- BPE allows us to transfer models even without training on a specific language

- Pre-trained models can help further

Burmese, Indonesian, Turkish

Transfer	BLEU My→En	BLEU Id→En	BLEU Tr→En
baseline (no transfer)	4.0	20.6	19.0
transfer, train	17.8	27.4	20.3
transfer, train, reset emb, train	13.3	25.0	20.0
transfer, train, reset inner, train	3.6	18.0	19.1

Table 3: Investigating the model's capability to restore its quality if we reset the parameters. We use En→De as the parent.

Aji et al. (2020)

Transformers for MT

Self-attention Intro (notes)

Multi-Head Self Attention

- Multiple “heads” analogous to different convolutional filters
- Let X = [sent len, embedding dim] be the input sentence
- Query $Q = W^Q X$: these are like the **decoder hidden state** in attention
- Keys $K = W^K X$: these control what gets attended to, along with the query
- Values $V = W^V X$: these vectors get summed up to form the output

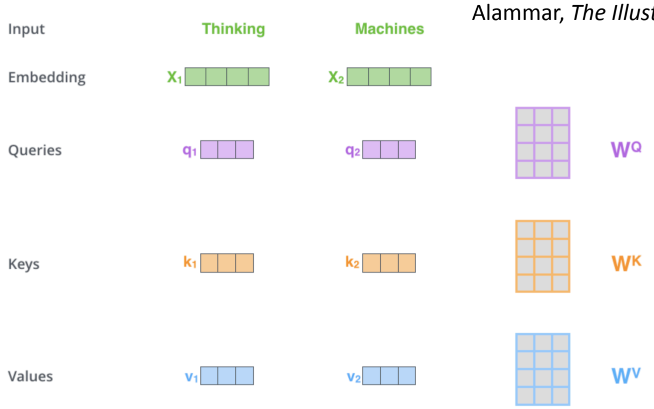
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

dim of keys

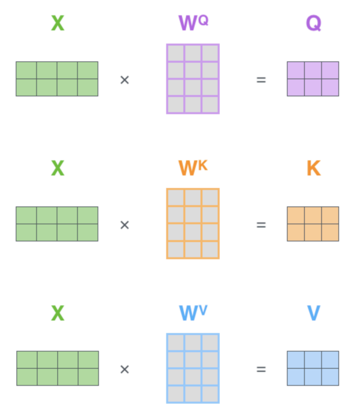
Vaswani et al. (2017)

Multi-Head Self Attention

Alammar, *The Illustrated Transformer*

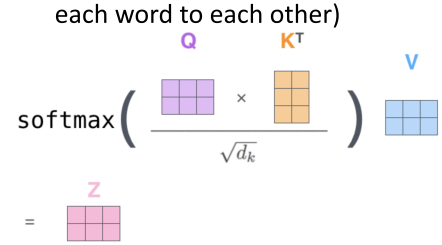


Multi-Head Self Attention



Alammar, *The Illustrated Transformer*

sent len x sent len (attn for each word to each other)



sent len x hidden dim

Z is a weighted combination of V rows

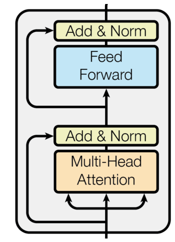
Properties of Self-Attention

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

- *n* = sentence length, *d* = hidden dim, *k* = kernel size, *r* = restricted neighborhood size
- **Quadratic complexity**, but O(1) sequential operations (not linear like in RNNs) and O(1) "path" for words to inform each other

Vaswani et al. (2017)

Transformers



- Alternate multi-head self-attention layers and feedforward layers
- Residual connections let the model "skip" each layer — these are particularly useful for training deep networks

Vaswani et al. (2017)

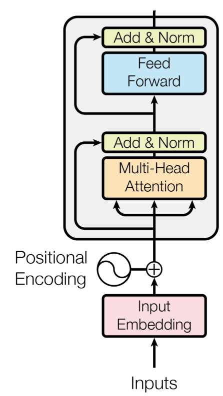
Transformers: Position Sensitivity

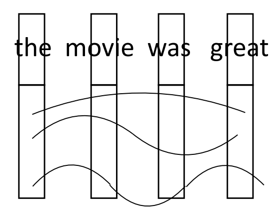
The ballerina is very excited that she will dance in the show.

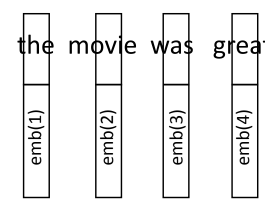
- If this is in a longer context, we want words to attend *locally*
- But transformers have *no notion of position* by default

Vaswani et al. (2017)

Transformers: Position Sensitivity





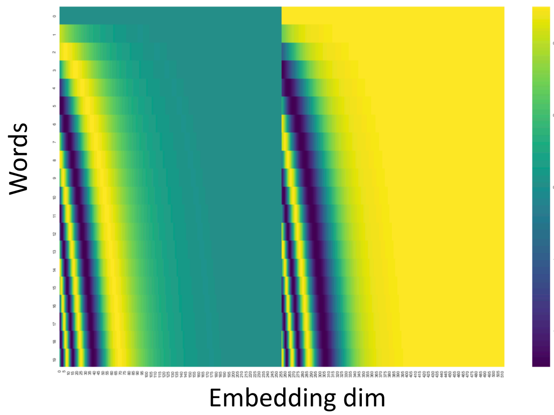


- Augment word embedding with position embeddings, each dim is a sine/cosine wave of a different frequency. Closer points = higher dot products
- Works essentially as well as just encoding position as a one-hot vector

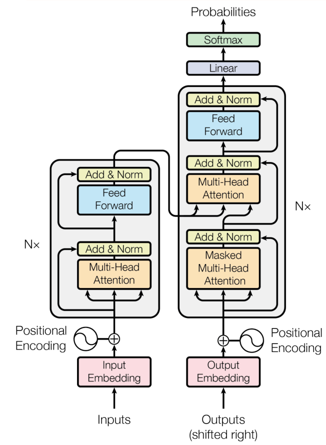
Vaswani et al. (2017)

Transformers

Alammar, *The Illustrated Transformer*



Transformers: Complete Model



- Encoder and decoder are both transformers
- Decoder alternates attention over the output and attention over the input as well
- Decoder consumes the previous generated tokens but has *no recurrent state*

Vaswani et al. (2017)

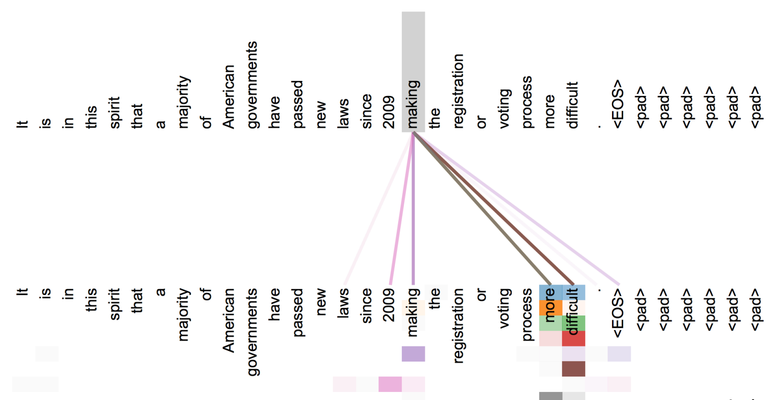
Transformers

Model	BLEU	
	EN-DE	EN-FR
ByteNet [18]	23.75	
Deep-Att + PosUnk [39]		39.2
GNMT + RL [38]	24.6	39.92
ConvS2S [9]	25.16	40.46
MoE [32]	26.03	40.56
Deep-Att + PosUnk Ensemble [39]		40.4
GNMT + RL Ensemble [38]	26.30	41.16
ConvS2S Ensemble [9]	26.36	**41.29**
Transformer (base model)	27.3	38.1
Transformer (big)	**28.4**	**41.8**

- Big = 6 layers, 1000 dim for each token, 16 heads, base = 6 layers + other params halved

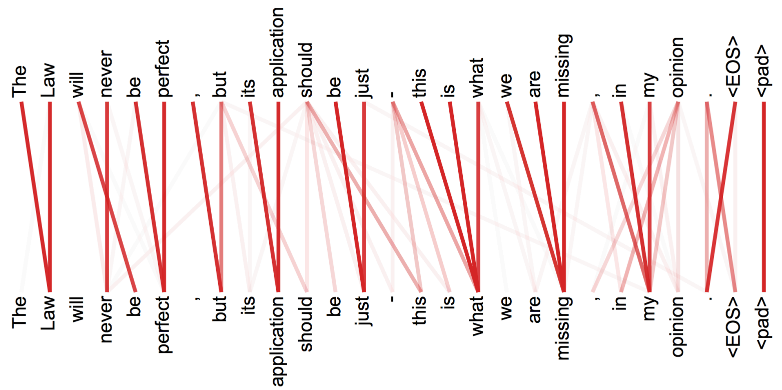
Vaswani et al. (2017)

Visualization



Vaswani et al. (2017)

Visualization



Vaswani et al. (2017)

Takeaways

- Transformers are powerful seq2seq models, can also replace RNN encoders
- When you have massive datasets like for machine translation, transformers work very well
- Next: **pre-training** with transformer language models