Neural Machine Translation (Tutorial)

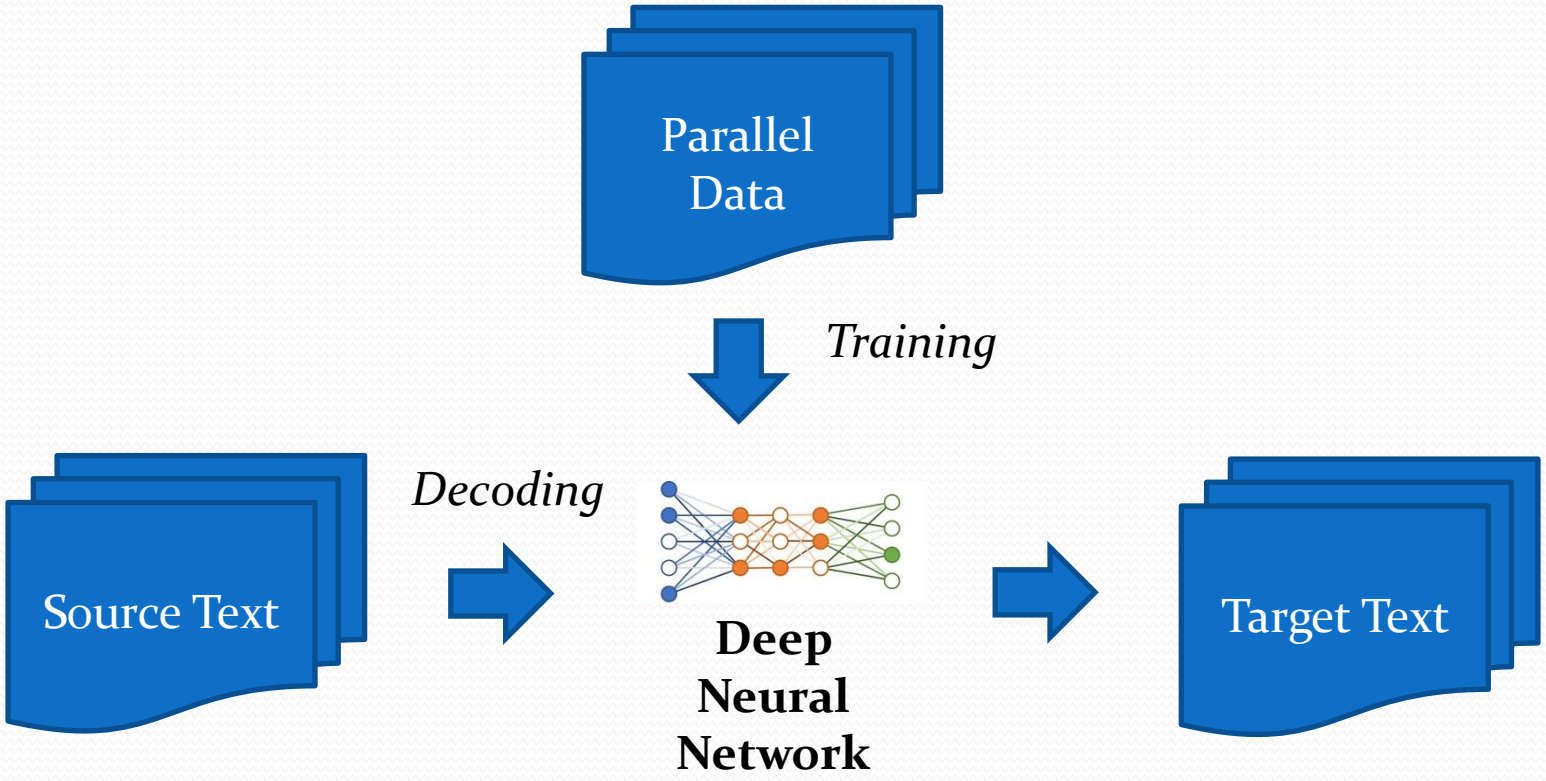
Abdul Rafae Khan
akhan4@stevens.edu

Machine Translation

- Automatically translate a sentence from one human language (English) to another human language (German)

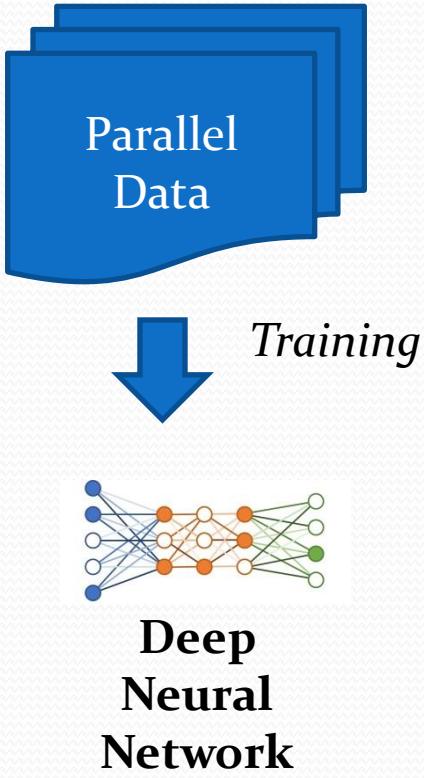
Neural Machine Translation

- Use neural networks to train the translation system



Training the Model

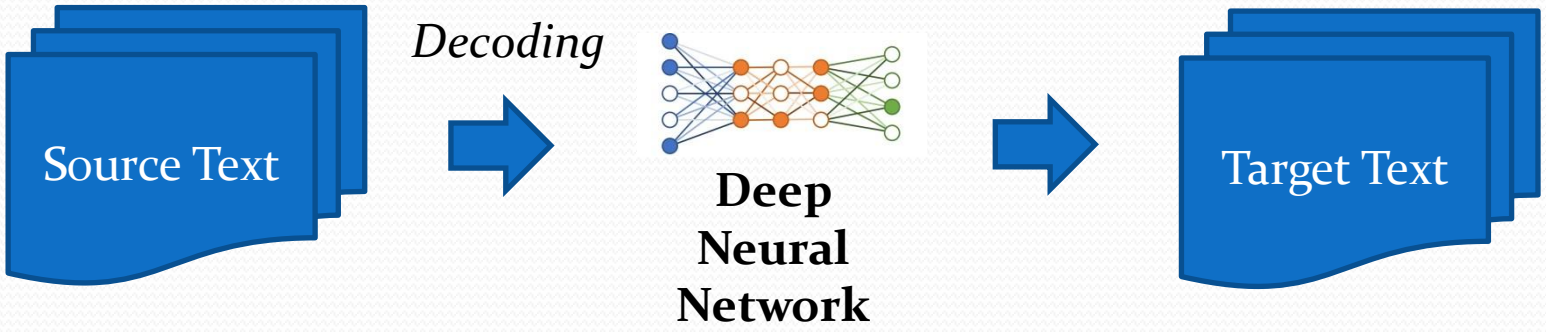
- Train model on:
  - Training data
  - Validation data



Decoding the Model

- Test model on:
  - Test data

- No overlap between test and training/validation data



(1) Download Parallel Data

- Aligned sentences
- Also called Bilingual data

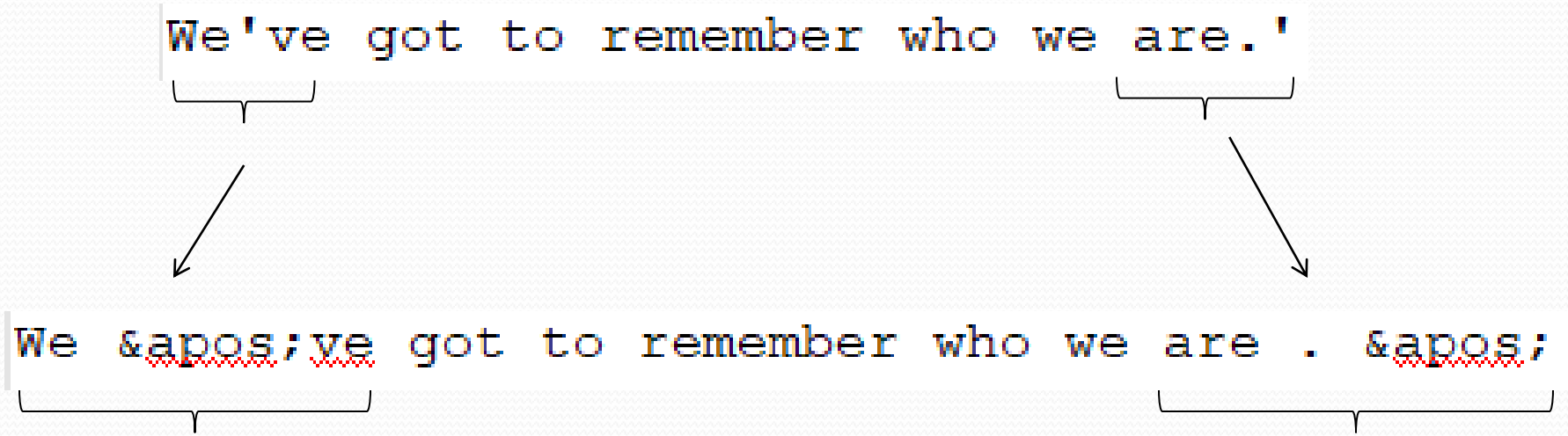
English (source) Data		**German (target) Data**
This was reported by broadcaster SABC.	→	Dies berichtete der Sender SABC.
The crime had caused nationwide horror.	→	Die Tat hatte landesweit Entsetzen ausgelöst.
We've got to remember who we are.'	→	Wir müssen uns daran erinnern, wer wir sind.
I'm reading a terribly sad book these days.	→	Ich lese derzeit ein furchtbar trauriges Buch.

(2) Extract Raw Data

```
<srcset setid="newstest2014" srclang="any">
<doc docid="007c7b2cf0eaeb1b05efd7ef4871b255" genre="news" origlang="xx">
<seg id="1">This was reported by broadcaster SABC.</seg>
<seg id="2">The crime had caused nationwide horror.</seg>
<seg id="3">We've got to remember who we are.'</seg>
<seg id="4">I'm reading a terribly sad book these days.</seg>
</doc>
```

```
This was reported by broadcaster SABC.
The crime had caused nationwide horror.
We've got to remember who we are.'
I'm reading a terribly sad book these days.
```

(3) Tokenize Data



(4) Truecase & Clean Data

- Case changes meaning

```
Later, newcomer Nick Jonas decided to google the answer.
```

```
In 2007, Google researchers estimated there
were one hundred trillion words on the Internet.
```

- Learn a truecase model on train data

- Remove very long sentences
  - Ratio of source sentence length vs target sentence length

(5) Byte Pair Encoding

- No 100% overlap in train and target vocabulary
- A lot of unknown words in test data
- BPE reduces the number of unknown words
- Convert words to sub-words (delimiter for separation)

Training Vocabulary	unknown	-> un@@ known
	fortunate	-> fortunate
	slowly	-> slow@@ ly
Unknown word in test	unfortunately	-> un@@ fortunate@@ ly

(5) Word to Integer Sequence

- Create vocabulary
- Convert words to indices and sentences to sequence of indices
- Easy to use during NMT training

(5) Word to Integer Sequence

(i) Create Vocabulary

```
so finally , we used a 3d surface scann@@ er .
so i was train@@ ed as a histori@@ an .
```

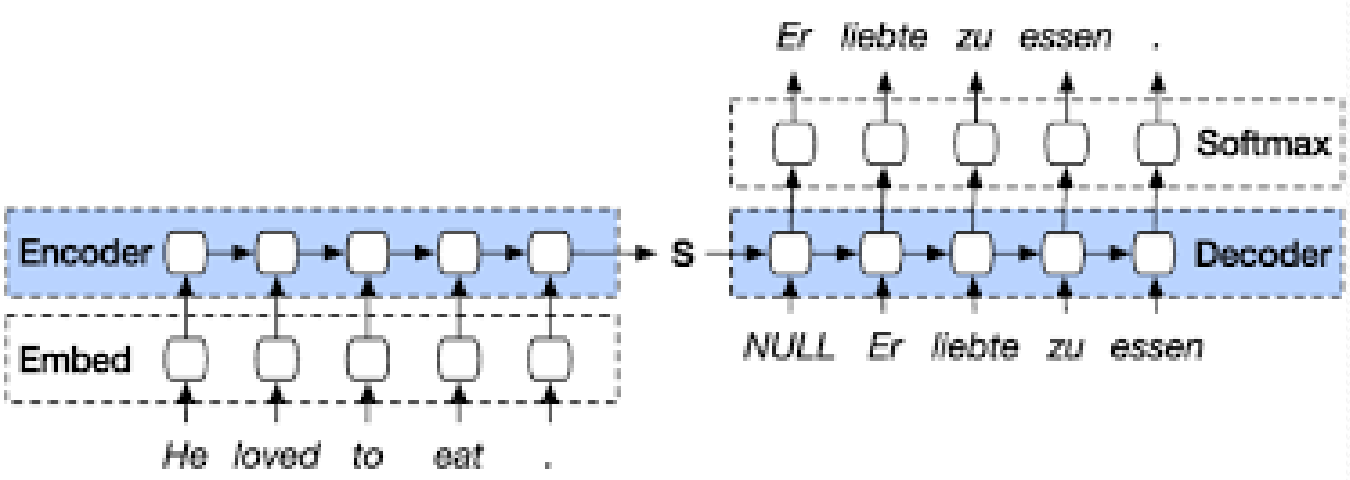
we	0
,	1
was	2
an	3
er	4
ed	5
as	6
.	7
train@@	8
surface	9
so	10
histori@@	11
scann@@	12
finally	13
3d	14
i	15
a	16
used	17

(ii) Convert to Indices

```
10 13 1 0 17 16 14 9 12 4 7
10 15 2 8 5 6 16 11 3 7
```

(6) Train NMT Model

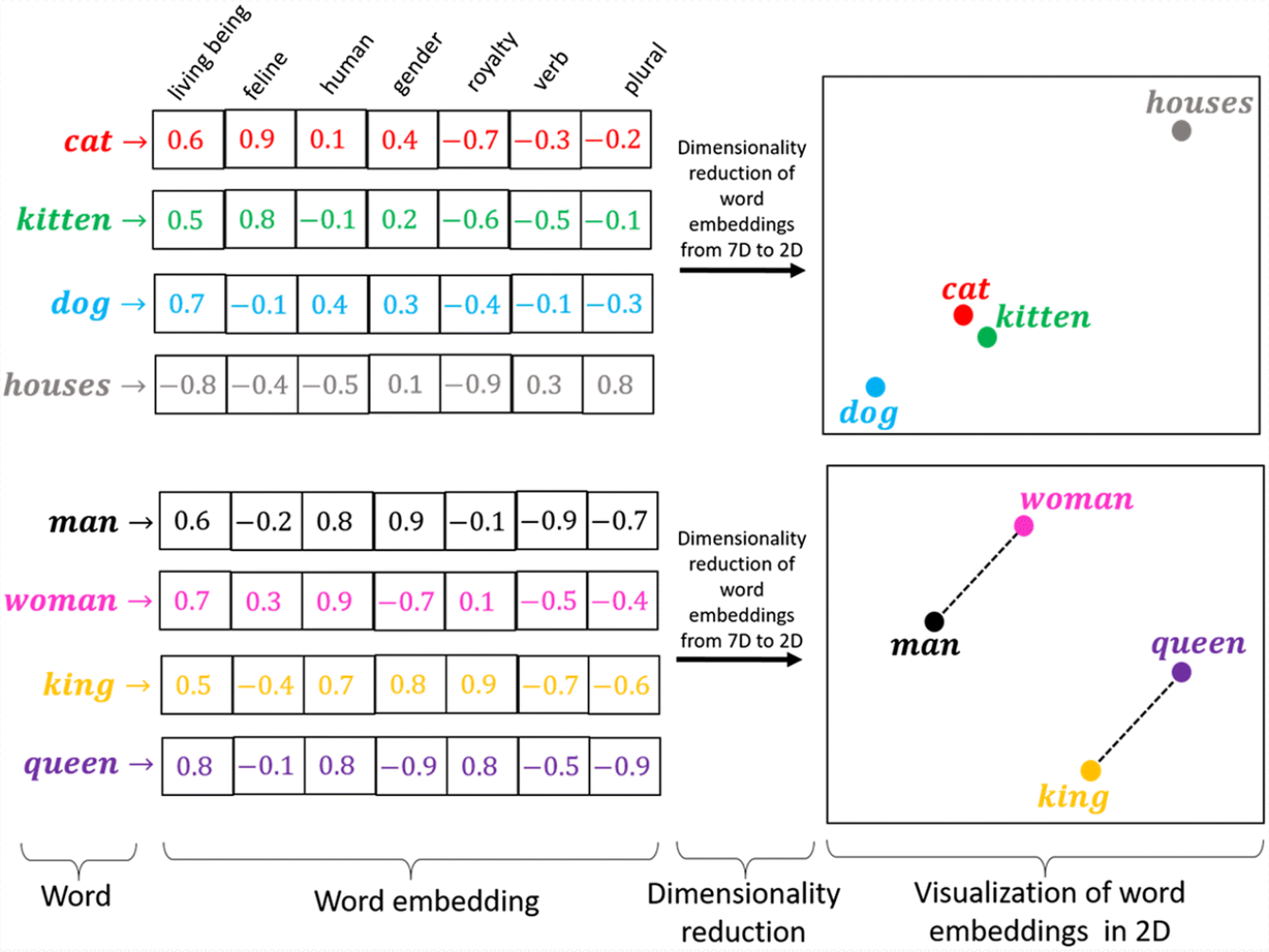
- Recurrent Neural Network
  - Basic Sequence-to-Sequence (Seq2Seq) Model
  - Encoder/Decoder framework
  - Word Embeddings



Word Embeddings

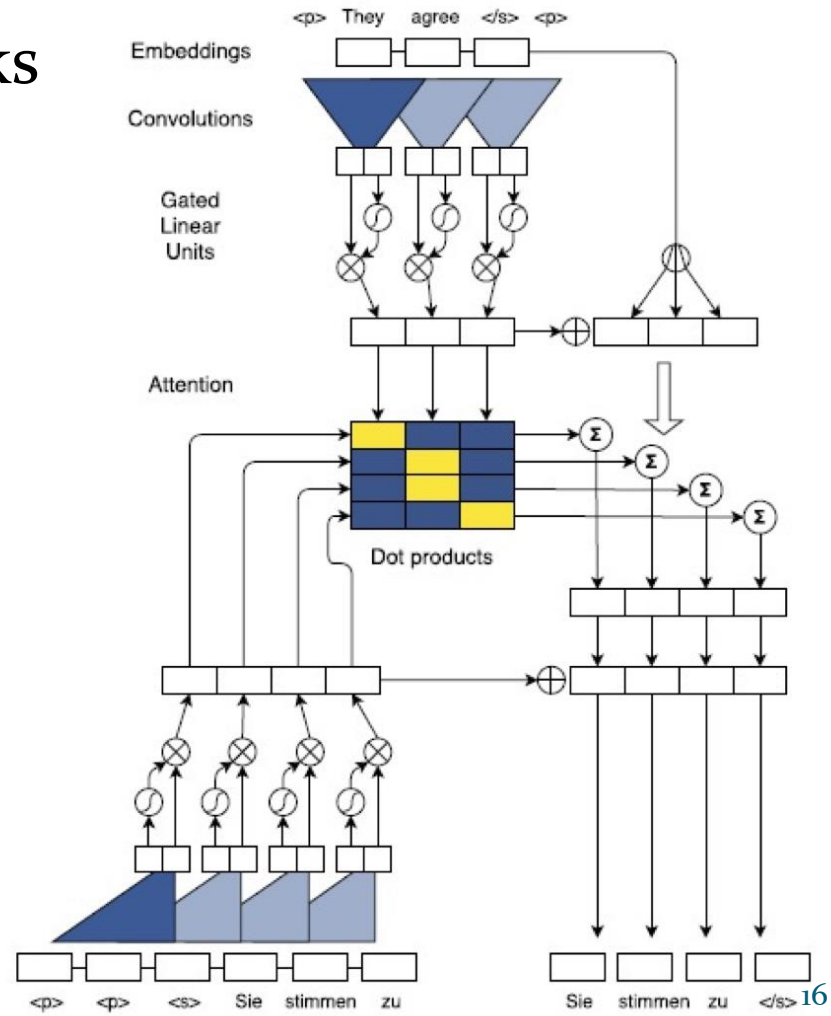
- Neural Networks use vectors as inputs
- Convert a word to a fixed-length vector
- Semantic meaning of words is preserved

Word Embeddings



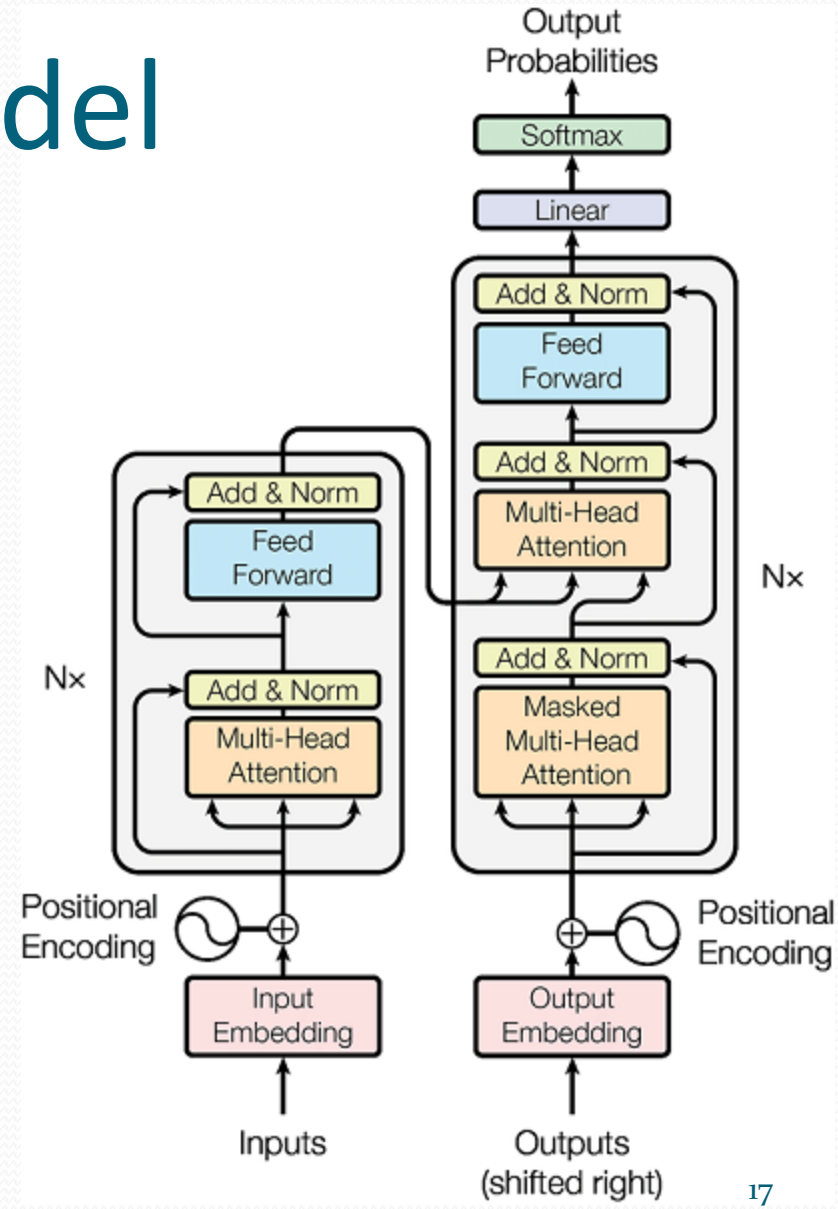
(6) Train NMT Model

- More sophisticated networks
- Convolutional Seq2Seq
  - Convolutions
  - Position Embeddings
  - Attention



(6) Train NMT Model

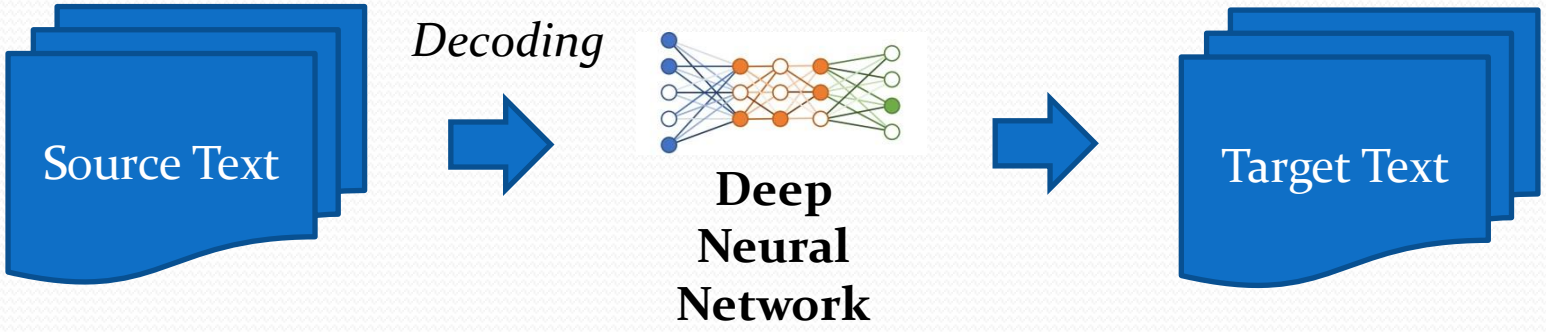
- More sophisticated networks
- Transformer
  - Self-Attention
  - Multi-Head Attention



(6) Train NMT Model

- A number of model parameters (hyperparameters) can be tuned
- GPU Memory
  - Size of word embeddings (encoder/decoder dimensions)
  - Number of sentences to learn simultaneously (batch size)
- Training Time
  - Maximum number of runs (epochs)
  - When to stop training? (patience)

(7) Decoding NMT Model



(8) Post-processing

- Convert BPE sub-words into words
- Remove truecasing
- Remove tokenization

(9) Automatic Evaluation

- Compare translation outputs (hypothesis) with the original translations (reference)
- **Bi**Lingual **E**valuation **U**nderstudy (BLEU)
- From 0 to 1
  - 1 for identical hypothesis and reference
- Usually represented in percentage (0-100%)
- Higher than 30% considered a good BLEU score
- Sacrebleu
- Mteval

(9) Automatic Evaluation

- **BLEU** Cased & Un-cased
- **W**ord **E**rror **R**ate (WER)
  - Similar to Edit distance
- **M**etric for **E**valuation of **T**ranslation with **E**xplicit **Or**dering (METEOR)
- Round-trip Translation
  - Translate the translation to source language and compare original source
  - Comparing not 1 but 2 systems
  - Used mostly in unsupervised translation (no parallel data)

(10) Manual Evaluation

- By human translators
  - Adequacy
  - Fluency

	Fluency	Adequacy
1	incomprehensible	none
2	disfluent English	little meaning
3	non-native English	much meaning
4	good English	most meaning
5	flawless English	all meaning

Thank You!