CMPT 880 Spring 2019 Project Presentation
April 15, 2019

Exploring the Generalizability of Sequence-to-Sequence Architecture

Kumar Abhishek and **Nishant Kambhatla**

Introduction

Generalizability

- The extent to which research findings can be applied to settings other than that in which they were originally tested.

- ~~One task → many architectures~~ One architecture → many tasks

Motivation

- The ultimate aim of AI is to reach Artificial General Intelligence (AGI).
- Deep learning has improved the performance on many NLP tasks individually.
- But the generalization of NLP models remains a hard problem within an approach that focuses on the particularities of a single metric, dataset, and task.

Motivation 1

The Natural Language Decathlon: Multitask Learning as Question Answering

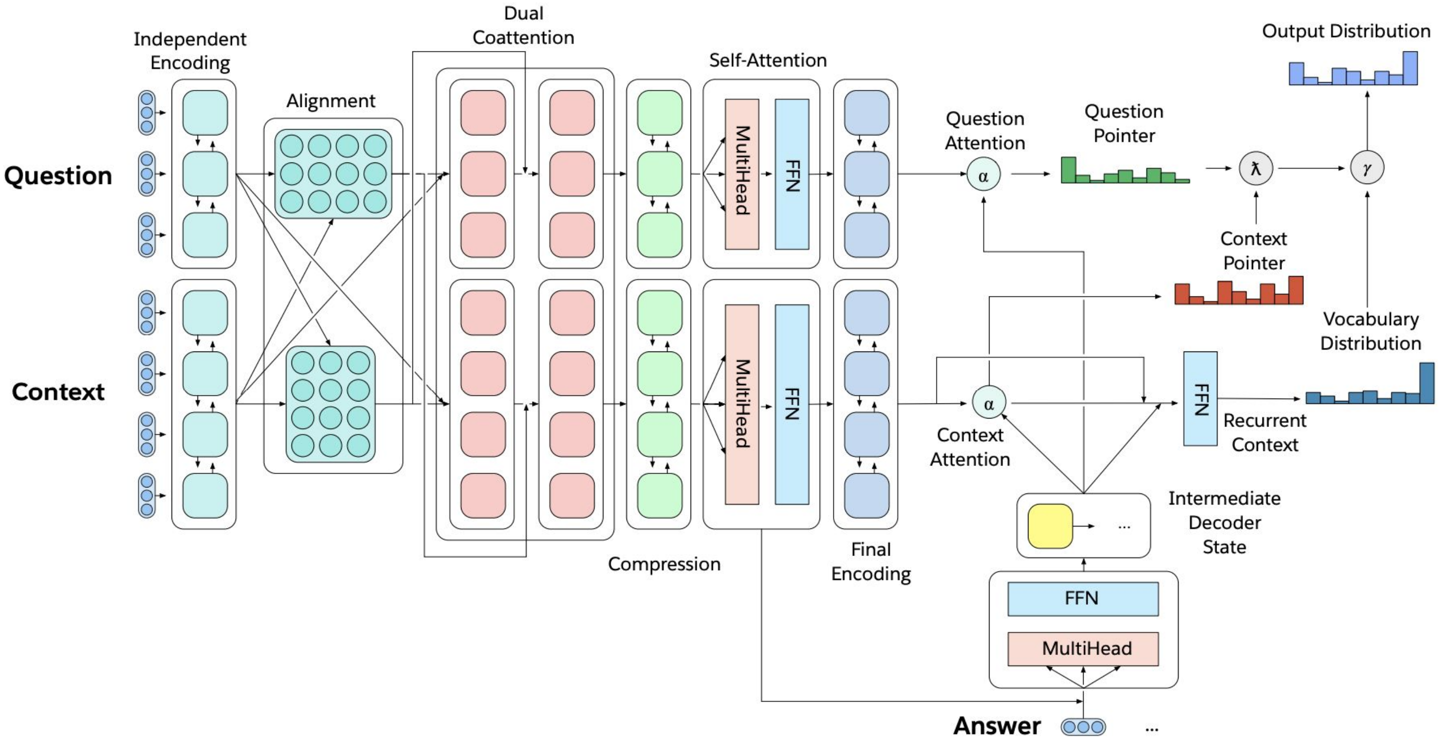
Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher
Salesforce Research
{bmccann,nkeskar,cxiong,rsocher}@salesforce.com

Motivation 1

➢ Multitask Question Answering Network (MQAN)

Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

Motivation: The 10 task model



Baseline

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

Motivation 2

Domain Control for Neural Machine Translation

Catherine Kobus and **Josep Crego** and **Jean Senellart**
firstname.lastname@systrangroup.com
SYSTRAN International / 5 rue Feydeau, 75002 Paris, France

Proceedings of Recent Advances in Natural Language Processing, pages 372–378,
Varna, Bulgaria, Sep 4–6 2017.

Motivation 2

- ➢ NMT systems are typically trained on domain specific data (*in-domain*)
  e.g., parliament proceedings **or** TED talks

- ➢ Models break when tested on *out-of-domain* data
  e.g., trained on TED talks ➡ tested on medical data

- ➢ Perform domain adaptation to teach the model diverse representations
  ➡ helps with out-of-domain data

Motivation 2

➢ Add domain-tags to the data:

Src: `Headache may be experienced`
Tgt: `Des céphalées peuvent survenir`

↓

Src: `Headache may be experienced` **`@MED@`**
Tgt: `Des céphalées peuvent survenir`

Motivation 2

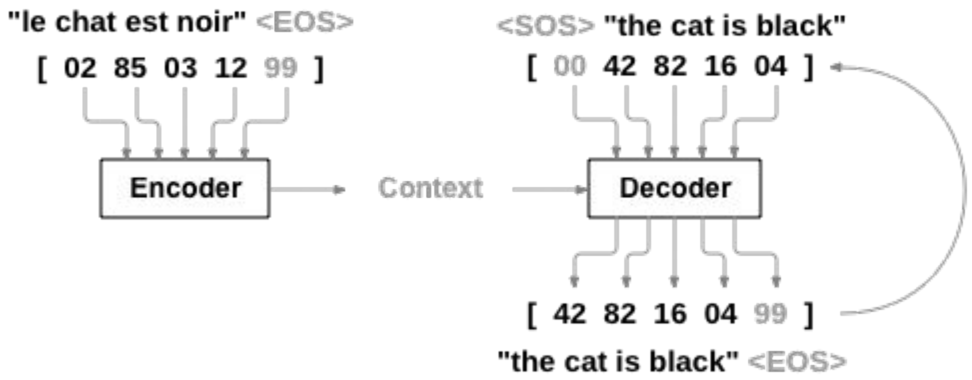
➢ Results:

Domain	Single	Join
Constraint	None	
IT	52.73	53.81
Literature	20.25	29.81
Medical	33.97	41.83
News	29.70	33.83
Parliamentary	37.34	37.53
Tourism	37.05	37.46

Thoughts

- Are complex models the only solution for complex problems?
- We need simpler baselines!
- Use domain adaptation for multitask learning.
- Approach 1: Add task tags instead of domain tags:
  `@nmt@` **`@sum@`** **`@dialog@`**
- Approach 2: Extend word embeddings to include *task (domain) embeddings*.

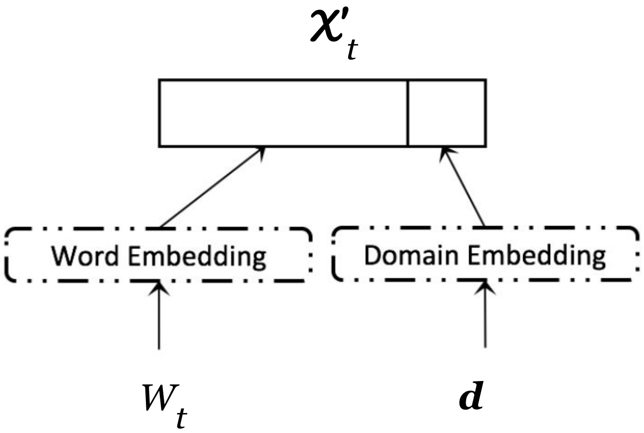
The Encoder-Decoder Architecture



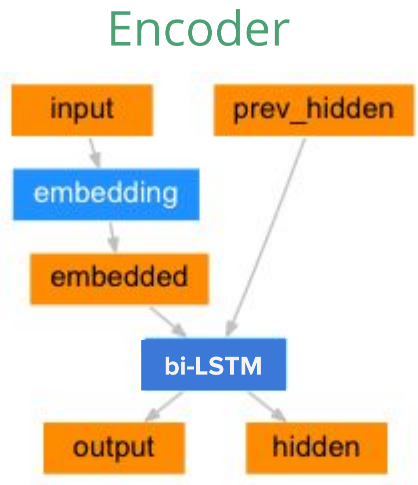
Kyunghyun Cho et al., *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*, EMNLP, 2014.

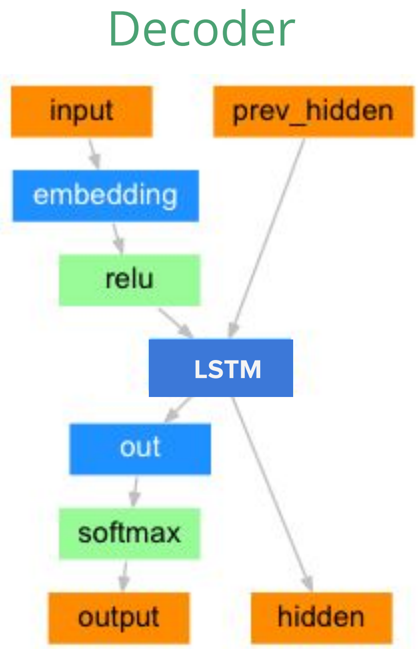
Encoder-Decoder

- **Encoder**
  - **Input sequence** $\mathbf{x} = (x_1, \cdots, x_{T_x})$
  - **Hidden state** $h_t = f(x_t, h_{t-1})$
  - **Encoded context** $c = q(\{h_1, \cdots, h_{T_x}\})$
- **Decoder**
  - **Probability** $p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c)$



The Encoder-Decoder Architecture





Sean Robertson, *Translation With A Sequence To Sequence Network And Attention*, PyTorch Tutorials.

Encoder-Decoder with Attention

- **Attention Decoder**
  - **Probability** $$p(y_i|y_1, \ldots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

    Hidden state for time i $$s_i = f(s_{i-1}, y_{i-1}, c_i)$$
  - **Context vector as weighted sum of hidden state** $$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$
  - **Weights** $$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$ **where** $$e_{ij} = a(s_{i-1}, h_j)$$

Dzmitry Bahdanau et al. *Neural machine translation by jointly learning to align and translate*, ICLR 2014.

Encoder-Decoder with Attention

- **Attention Decoder**
  - **Probability** $p(y_i | y_1, \ldots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$

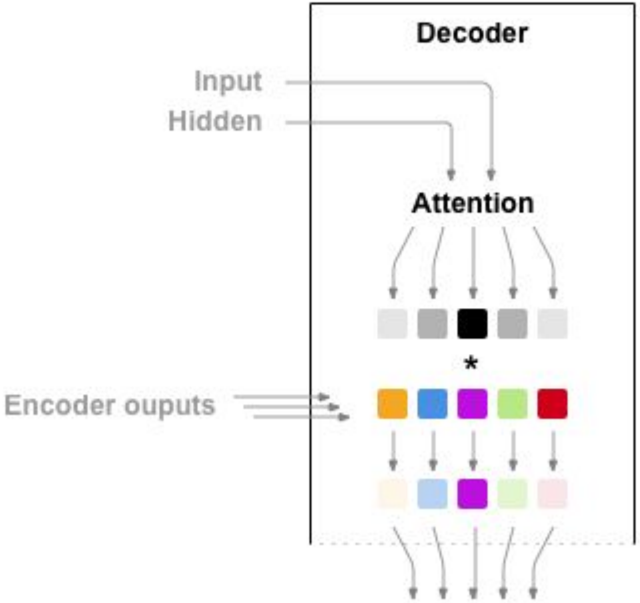
    Hidden state for time i $s_i = f(s_{i-1}, y_{i-1}, c_i)$
  - **Context vector as weighted sum of hidden state** $c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$
  - **Weights** $c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$ **where** $e_{ij} = a(s_{i-1}, h_j)$

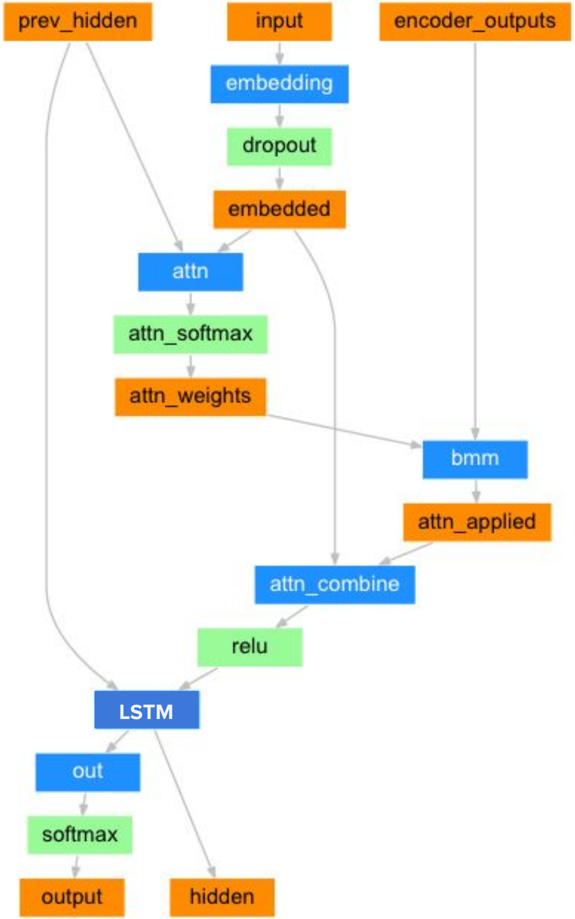
hidden state from encoder

Dzmitry Bahdanau et al. *Neural machine translation by jointly learning to align and translate*, ICLR 2014.

Encoder-Decoder with Attention

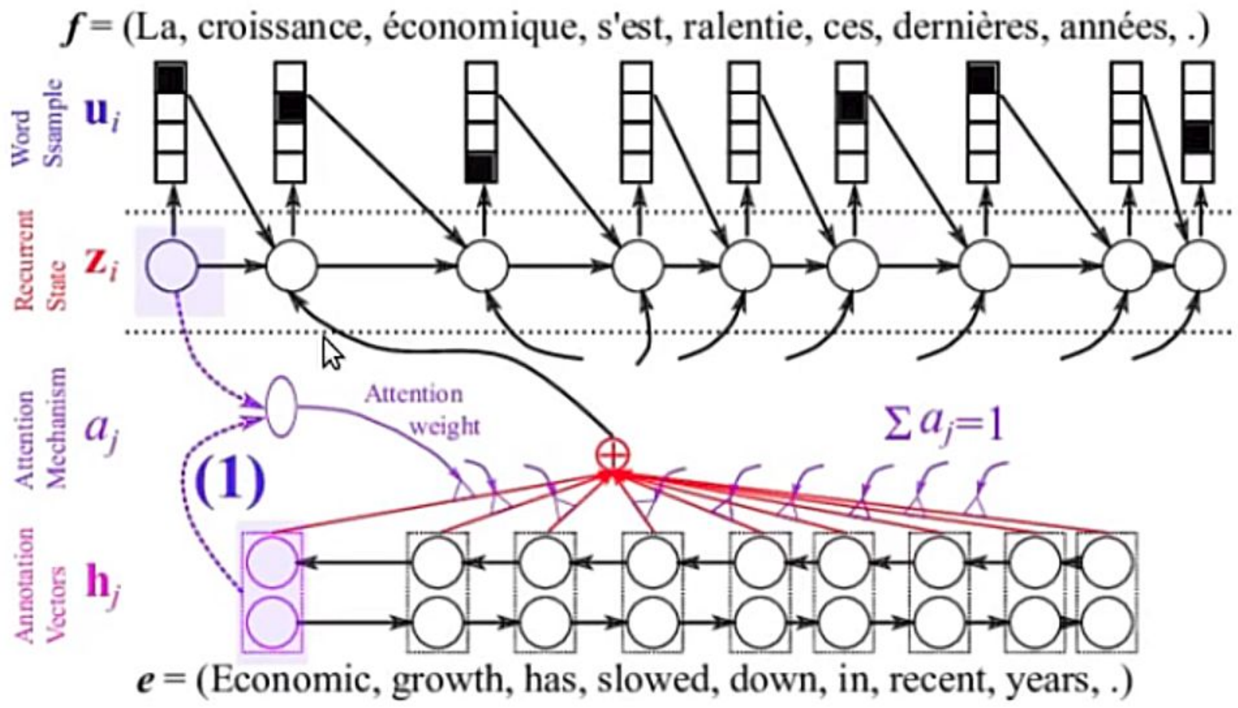
Attention Decoder





Sean Robertson, *Translation With A Sequence To Sequence Network And Attention*, PyTorch Tutorials.

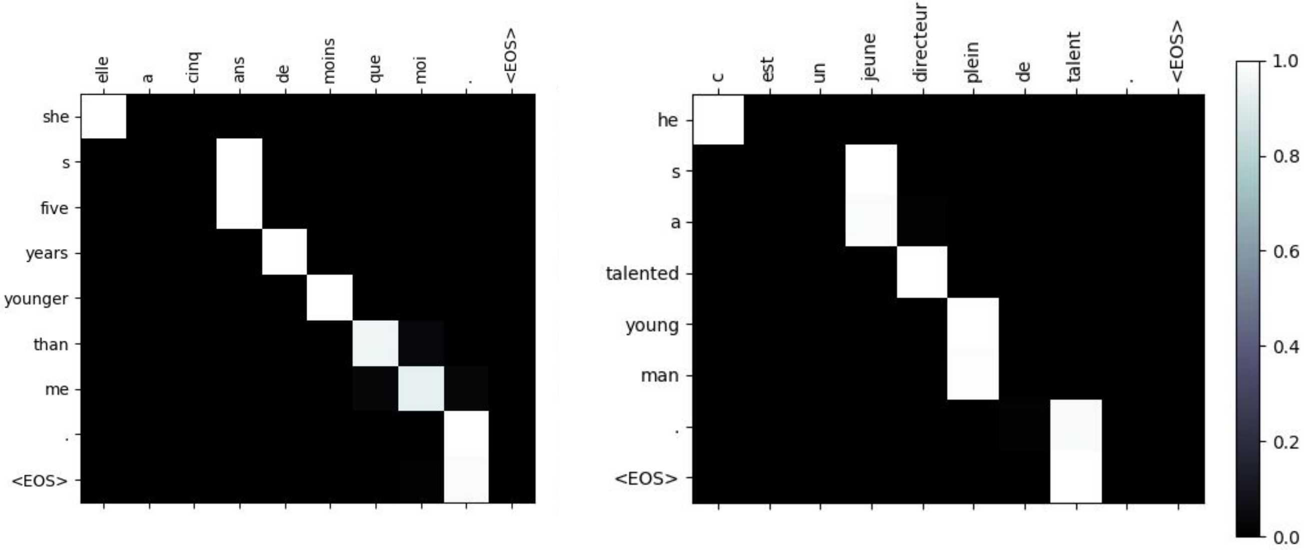
Encoder-Decoder with Attention



Guillaume Chevalier, *Attention Mechanisms in Recurrent Neural Networks (RNNs)*, YouTube.

Encoder-Decoder with Attention

[En-Fr] Visualizing the alignment model which scores how well the inputs around position j and the output at position i match:



Tasks and Datasets

Neural Machine Translation

- Use neural network to learn a model for translating from one language to another.
    - Can be trained directly on source and target text end-to-end.

Seminal Works:

- Cho et al. (EMNLP 2014) - *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*
    - A RNN encoder-decoder model to learn “a semantically and syntactically meaningful representation of linguistic phrases.”
    - A new hidden unit - GRU (Gated Recurrent Unit)
- Bahdanau et al. (ICLR 2015) - *Neural Machine Translation by Jointly Learning to Align and Translate*
    - Attention decoder.

Neural Machine Translation: Dataset

- German to English translation **[De-En]**
- International Workshop on Spoken Language Translation (IWSLT) 2014 dataset
  - 63.9k sentence pairs for training, 930 for validation, and 1660 for testing

text	#sent.	German $\|W\|$	German $\|V\|$	English $\|W\|$	English $\|V\|$
parallel	63.9k	1.16M	63.1k	1.22M	35.5k
dev2010	930	19.1k	4.2k	20.2k	3.4k
tst2010	1660	30.3k	5.2k	32.0k	3.9k

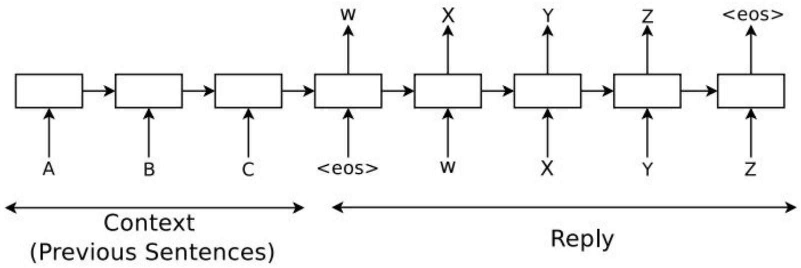
IWSLT 2014 Evaluation Campaign, *http://workshop2014.iwslt.org/59.php*

Dialogue Systems

Model and generate realistic (human-like) conversations.

Seminal Work:

- Vinyals et al. (ICML 2015) - *A Neural Conversational Model*
    - RNN-based sequence-to-sequence architecture.
    - Evaluated on 2 datasets - OpenSubtitles dataset (open) and IT Helpdesk Troubleshooting dataset (private)



Dialogue Systems

Movie subtitles in English from OpenSubtitles.

446,612 documents, 3.2G tokens

Examples:

- “Yeah , I want to tell him I 'm okay .”
- “I can 't make you believe it .”
- “We are going to the hospital .”

Jörg Tiedemann et al., *News from OPUS - A collection of multilingual parallel corpora with tools and interfaces.*, NeurIPS 2009

Text Summarization

Generate a headline or a short summary consisting of a few sentences to capture the salient ideas of an article or a passage.

Seminal Work:

- Nallapati et al. (CoNLL 2016) - *Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond*
  - A bidirectional GRU-RNN encoder and a unidirectional GRU-RNN decoder with attention.
  - Large vocabulary trick.
  - A new dataset (DUC corpus) of 1124 document summary pairs.

Text Summarization: Dataset

CNN/Daily Mail dataset.

Online news articles paired with multi-sentence (average 3.75 sentences) summaries.

	CNN			Daily Mail		
	train	valid	test	train	valid	test
# months	95	1	1	56	1	1
# documents	90,266	1,220	1,093	196,961	12,148	10,397
# queries	380,298	3,924	3,198	879,450	64,835	53,182
Max # entities	527	187	396	371	232	245
Avg # entities	26.4	26.5	24.5	26.5	25.5	26.0
Avg # tokens	762	763	716	813	774	780
Vocab size	118,497			208,045		

Karl Moritz Hermann et al., *Teaching Machines to Read and Comprehend*, NeurIPS 2015

Text Summarization

Example:

Text: The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “TopGear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.”

Summary: Producer Oisin Tymon will not press charges against Jeremy Clarkson, his lawyer says.

Karl Moritz Hermann et al., *Teaching Machines to Read and Comprehend*, NeurIPS 2015

Image Captioning

Generate 1 sentence descriptions of images.

Seminal Work:

- Xu et al. (ICML 2015) - *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*
  - A CNN encoder and an LSTM decoder.
  - Soft attention and hard attention both explored.

Image Captioning

Microsoft COCO dataset

164k images, 4 captions per image

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Tsung-Yi Lin et al., *Microsoft COCO: Common Objects in Context*, ECCV 2014

Evaluation

Metrics: BLEU

- **B**ilingua**l** **E**valuation **U**nderstudy
  - Measures an overlap of system generated text (summary, translation, etc.) against a set of reference texts.
- The BLEU ***n***-gram precision for a test corpus ***C*** and all hypothesis sentences ***S*** in ***C*** is

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

- The combines BLEU score is given as

$$\text{BLEU} = \underbrace{\text{BP}}_{\text{brevity penalty}} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n \right).$$

Kishore Papineni et al., *BLEU: a method for automatic evaluation of machine translation*, ACL 2002.

Metrics: ROUGE

- **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation
    - Measures an overlap of system generated text (summary, translation, etc.) against a set of reference texts.

- **ROUGE-N:** Measures unigram, bigram, trigram and higher order n-gram overlap.
- **ROUGE-L:** measures longest matching sequence of words using LCS (longest common subsequence).
    - Does not require consecutive matches.
    - Do not need to specify a pre-defined n-gram length.

Chin-Yew Lin, *ROUGE: A package for automatic evaluation of summaries*, Text Summarization Branches Out 2004.

Metrics: Perplexity

- A measure of how "perplexed" (surprised) a model is by the test data.
  - A lower perplexity score corresponds to a higher probability of the test data under the model.
- Defined as the inverse probability of the test set, normalized by the number of words.

$$PP(S) = P(w_1, \ldots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, \ldots, w_N)}}$$

Results

T1: Neural Machine Translation

```
** Hvar - Flirten, kokettieren, verführen - keine einfachen Aufgaben für
unsere Mädchen.

>> Hvar - flirting, flirting, seducing - no easy tasks for our girls.



** Dennoch liefern die neun "Schöne Münchnerin"-Kandidatinnen beim Shooting
mit People-Fotograf Tuan ab und trotzen Wind, Gischt und Regen wie echte
Profis.

>> However, the nine "Beautiful Munich" contestants in the shoot with People
photographer Tuan deliver and brave the wind, gush and rain like real
professionals.
```

T1: Neural Machine Translation

```
** "Ich hätte mich gefreut, wenn Mesut Özil weiter für Deutschland gespielt hätte.

>> "I would have been delighted if Mesut Özil had continued to play for Germany.


** Linz: Gartenhütte in Schrebergartensiedlung brannte

>> Linz: Garden hut in Schrebergartensiedlung burned
```

T2: Abstractive Summarization

1. cambodian leader hun sen on friday rejected opposition parties' demands for talks outside the country , accusing them of trying to "internationalize" the political crisis .

>> A. cambodian leader hun hen rejects opposition demands for talks in beijing.

>> B. cambodian government rejects opposition's call for talks abroad

T2: Abstractive Summarization

```
2. king norodom sihanouk on tuesday praised agreements by
cambodia's top two political parties, previously bitter
rivals, to form a coalition government led by strongman hun
sen .

>> A. king praises Cambodian coalition government of top two
political parties

>> B. two rival parties form coalition government at summit
convened by sihanouk.
```

T2: Abstractive Summarization

3. worried that party colleagues still face arrest for their politics , opposition leader sam rainsy sought further clarification friday of security guarantees promised by strongman hun sen. sam rainsy wrote in a letter to king norodom sihanouk that he was eager to attend the first session of the new national assembly on nov. 25, but complained that hun sen 's assurances were not strong enough to ease concerns his party members may be arrested upon their return to cambodia .

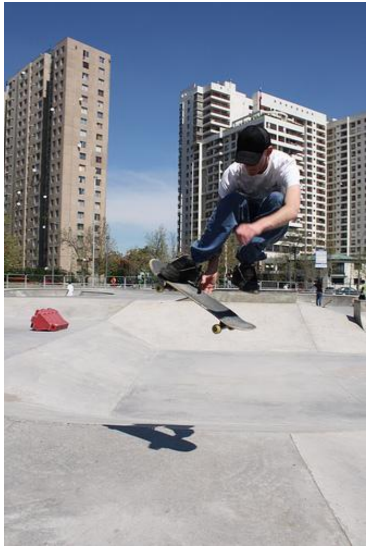
>> A. cambodian opposition leader doubts safety guaranty of prime minister

>> B. rainsy seeks stronger assurance of safety and freedom from prosecution.

T3: Dialog Generation

```
** Hello?
→ Hi?

** How are you?
→ I'm fine.

** What's your name?
→ It's hard to describe.

** How so?
→ I'm not sure.

** Benedict Cumberbatch?
→ Watson is married!

** Benadryl Cabbagepatch
→ I'm not sure.

** What color is the sky?
→ It's blue.

** What is your job?
→ It's not that I'm a f*cking werewolf!

** Are you a robot or human?
→ No.

** So i had the doctors test sammy's response to conditioning.
→ So he took the pills.

** What are you going to do ?
→ I'm going to get you a new car.

** They fear your power your intellect.
→ I'm not afraid of your power.
```

T4: Image Captioning



a man is doing a trick on a skateboard .

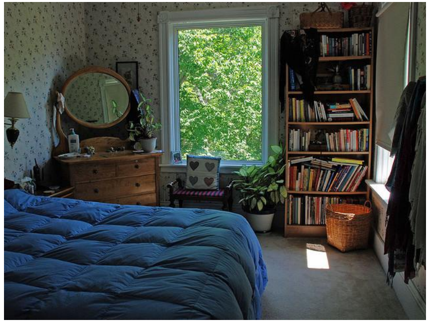
a man and woman are playing a game with a wii controller .

a kitchen with a stove and a stove top oven .

a large clock on a pole with a clock on it .

a group of people standing around a tennis court .

a bus that is driving down a street .



a bedroom with a bed , a bed , and a bed .

Results: Multitask Learning

Task	MQAN		Our Model		
	Single	Multi	Single	Multi-Domain Tag	Multi-Domain Embedding
NMT	25.0	14.2	**29.7**	28.1	28.9
Sum	19.0	25.7	26.4	29.3	**30.86**
Dialog	**85.0**	84.0	73.2	74.0	76.2

Results: Image Captioning

Metric	**BLEU-1**	**BLEU-4**
BRNN (Karpathy & Li, 2014)	64.2	20.3
Soft Attention (Xu et al. 2015)	**70.7**	**24.3**
Ours	60.6	15.9

Conclusion

Conclusion

- We present a strong baseline for multitask learning using seq2seq.
- This incorporates task (domain) information into the network.
- Allows to perform domain-adapted translations using a unique network that covers multiple tasks (domains).

Conclusion

- The encoder-decoder architecture is versatile.
- Attention is an indispensable part of this network architecture.
- A complex task need not necessarily have a complex solution.
- Domain adaptation methods can be effectively used in multitask settings, thus, helping the model *generalize* better.

Thank You.