A Survey on Advances in Neural Machine Translation

2016 Fall Digital Speech Processing Final Project

臺灣大學醫學系一年級 謝德威
學號B05401009

Alexander Shieh
National Taiwan University School of Medicine
teweishieh@gmail.com

Abstract

This survey covers some significant works on the recent advances in neural machine translation, a rapidly developing field surpassing conventional statistical machine translation results. We start by introducing the state-of-the-art machine translation system published by Google research in Oct. 2016. Then, we inspect its details by scraping each component of the proposed system and tracing back earlier breakthroughs that contributed to these components, which, in the end combined to produce amazing results that broaden the capabilities of neural machine translation.

Contents

Introduction

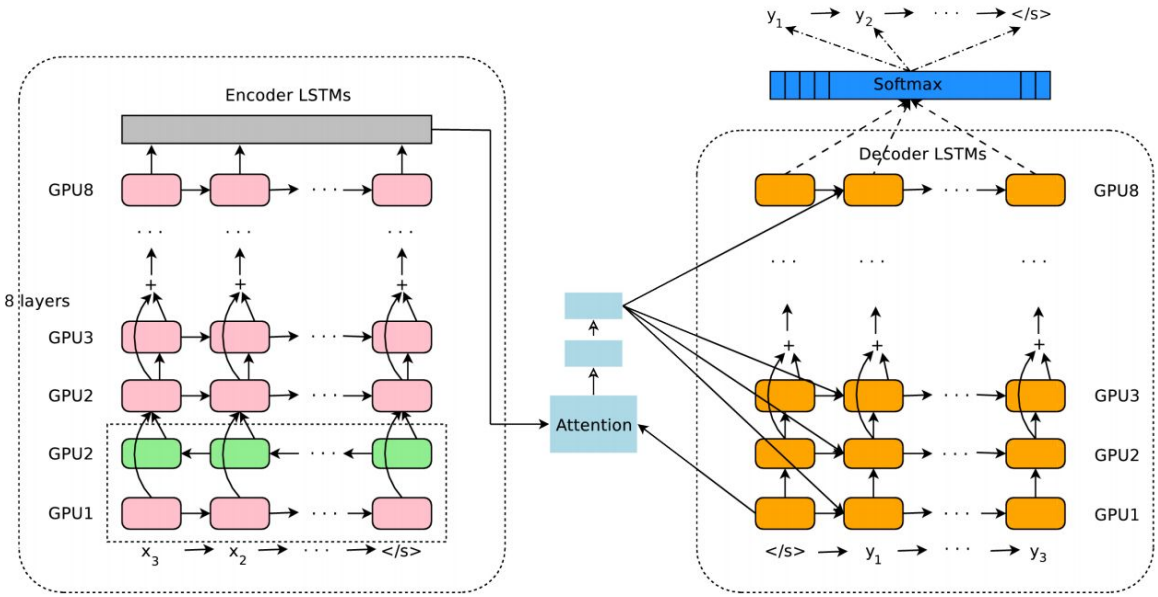
Machine translation is regarded as one of the most challenged and difficult task in natural language processing literature. In statistical machine translation settings, the problem is treated as calculating the probabilities of the target language model and the translation model of the language pair separately. Furthermore, the translation model has to deal with alignment problems and so on. However, in neural machine translation (NMT) , it is made possible to use a general, end-to-end model to solve this task.

More interestingly, recent breakthroughs in neural machine translation allows a single model to translate between different language pairs, instead of building separate models for each pair. Then, according to the latest results published by Google in Nov. 2016 showed the underlying strength and versatility of such approaches to achieve zero-shot learning across different languages using a single model, which demonstrates a surprising example of multitask and transfer learning.

We will start by reviewing the latest work done by Google, and trace back to the earliest works on modern neural machine translation using recurrent neural networks in 2014. Then, we will show in-depth the techniques developed in the following years that gained large success in tackling alignment and sentence encoding challenges, accompanied by surveying unsolved problems in these works. In the end, we will show some research on extensions of machine translation, such as real-time translation, and propose some possible future research topics related to neural machine translation.

Sequence to Sequence Learning with Recurrent Neural Networks

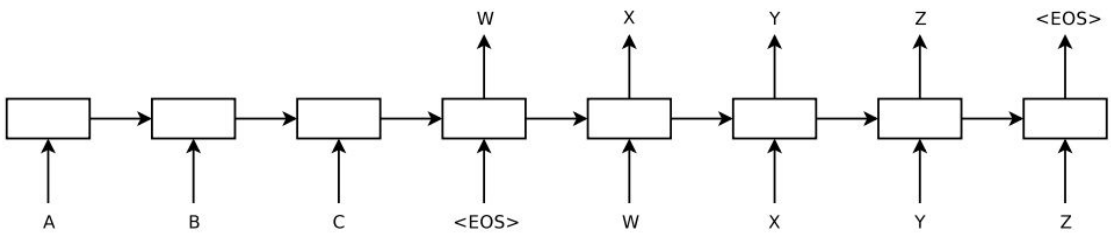
First, we will introduce the architecture of Google’s neural machine translation system (GNMT), then examine how each component was developed as it is. An overview of the model suggest it is constructed by three major components, namely the encoder, a stacked 8-layer LSTM network, the decoder network, another stacked 8-layer LSTM network, and the attention network, a one layer feedforward network connecting the two. Their functions and design insights will be illustrated in the following sections.



Above: The architecture proposed by Google Research.

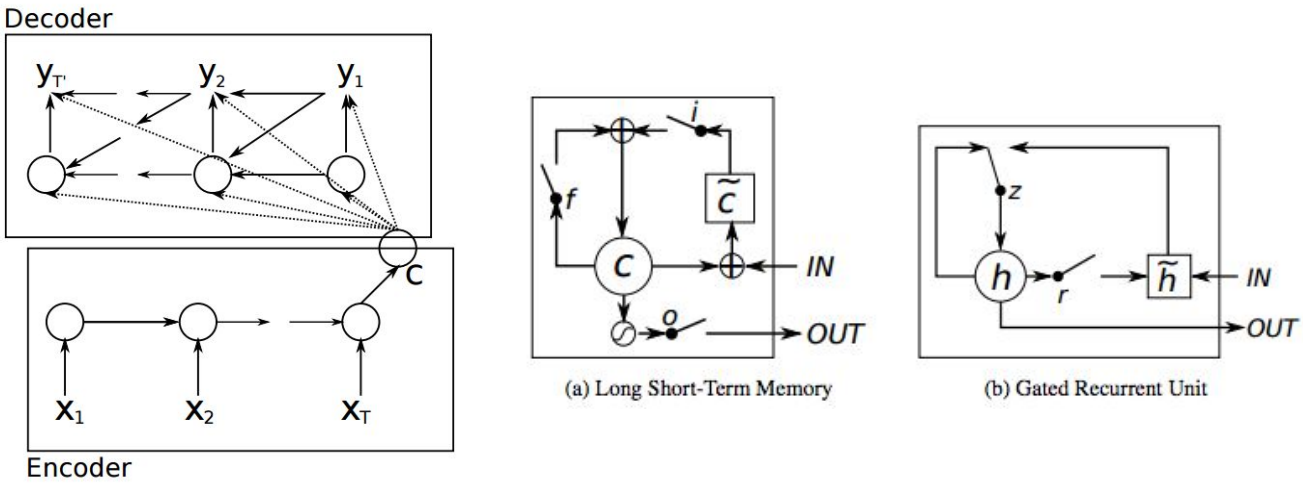
Neural machine translation was largely build upon Recurrent Neural Networks (RNNs), which was used as a general model in sequence to sequence learning. One important breakthrough in RNN research was the Long Short-Term Memory (LSTM) architecture proposed by Hochreiter and Schmidhuber in 1997[1]. LSTM successfully solved the problem of exploding and vanishing gradient when training conventional networks, thus capable of learning to bridge long intervals.

In 2014, Sutskever et al.[2] published its results of using a 4-layer deep LSTM model with 1000 cells in each layer and 1000 dimensional word embeddings to learn English to French (Using WMT'14 Dataset) translation. This simple straightforward approach achieved a BLEU score of 34.81, surpassing past baseline of 33.30. The BLEU (bilingual evaluation understudy) score is a common metrics to measure translation quality, and is highly correlated to human judgements. The paper also reveals that LSTM learned much better if the given source sentences are reversed while the target sentence are not.



Above: The simple approach proposed by Sutskever et al.

Meanwhile, Cho et al.[3] from Universite de Montreal come up with a similar construction that achieved comparable results. In addition, it also proposed a much simpler recurrent unit, called Gated Recurrent Unit (GRU). This construction was later found to generate comparable results to LSTM with significant less complexity and sometimes faster training time.



Left: The RNN encoder / decoder model; Right: Comparison of LSTM and GRU.

References

1. S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. Neural Computation, 1997.
2. I. Sutskever, O. Vinyals and Q. Le. Sequence to Sequence Learning with Neural Networks. NIPS, 2014.
3. K. Cho, D. Bahdanau, F. Bougares, H, Schwenk and Y. Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. EMNLP, 2014.
4. J. Chung, C. Gulcehre, K. Cho and Y. Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. NIPS Deep Learning Workshop, 2014.

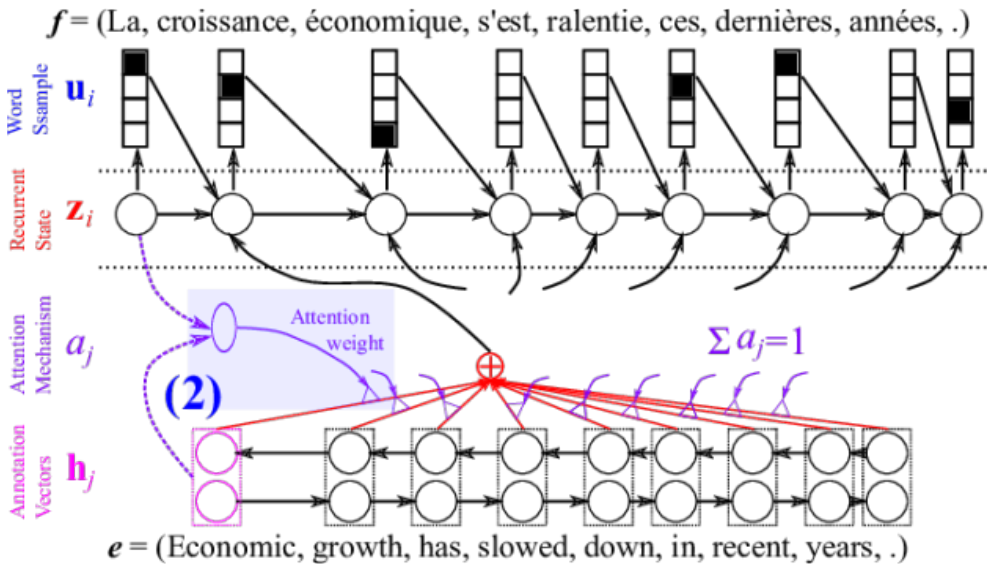
Bidirectional RNNs and Attention Mechanism

However, these simple approaches have serious limitations on the length of the sentence given for translation. More specifically, its performance would peak at sentences with length about 20 characters, and then drops as the sentence length

extended.[1] Therefore, Cho et al. gave a novel approach that combined the encoding technique of Bidirectional RNNs and the Attention Mechanism in 2015.

The Bidirectional RNNs obtained significant results in speech recognition in 2013, generating results slightly better than DNN and GMM-HMM baselines.[2] The technique was to read the source sentence alongside with its reverse, then concatenate them to become the input for the decoder network. With this improvement, the decoder network can benefit from a more complete description of the whole sentence.[3]

As for the Attention Mechanism, it was used to enhance the encoder and the decoder's ability to align and focus on generating its current output. Implemented by giving a weight for each bidirectional state of the sentence, and represent the input sentence as a weighted sum of these states. The weight was calculated by a feedforward network using the bidirectional state and the last output of the decoder as inputs.



Above: The graphical summarization of Attention Mechanism with Bidirectional Recurrent Neural Network.[4]

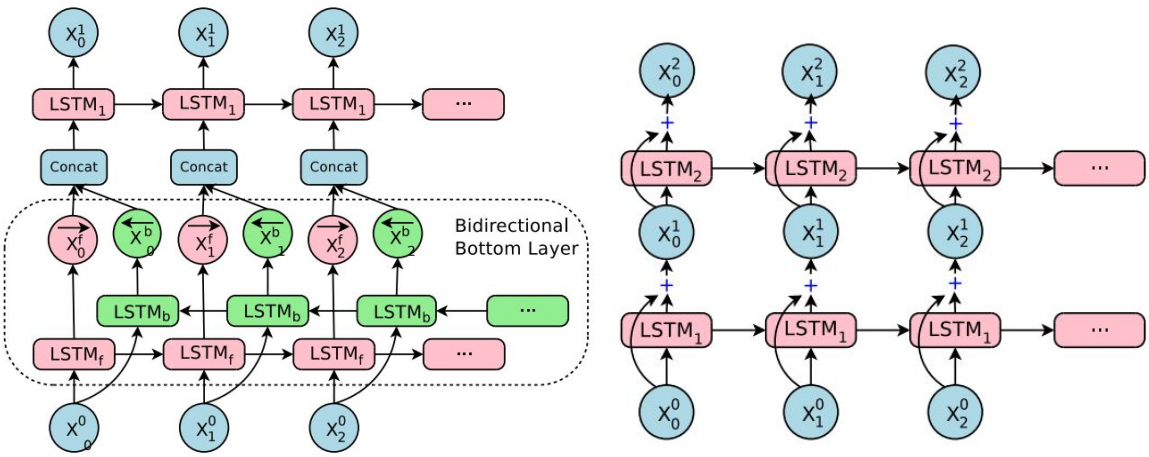
This approach also solves the fact that simple representation of the sentence as a single vector is counter intuitive to the fundamentals of information theory, that is, a longer sentence should carry more information and thus have a longer encoded length. Moreover, if the encoder network was replaced by a Convolutional Neural Network (CNN), this construction can be used to generate caption for images and videos as depicted in Vinyals et al.[5] and Cho et al.[6] later in 2015.

References

1. K. Cho, B. Merrienboer, D. Bahdanau and Y. Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. SSST-8.
2. A. Graves, N. Jaitly and A. Mohamed. Hybrid Speech Recognition with Deep Bidirectional LSTM. ASRU, 2013.
3. D. Bahdanau, K. Cho and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.
4. K. Cho. Introduction to Neural Machine Translation with GPUs (part 3).
5. O. Vinyals, A. Toshev, S. Bengio and D. Erhan. Show and Tell: A Neural Image Caption Generator. arXiv:1411.4555v2 [cs.CV] 20 Apr 2015.
6. K. Cho, A. Courville and Y. Bengio. Describing Multimedia Content using Attention-based Encoder--Decoder Networks. IEEE Transactions on Multimedia, 2015.

Stacked RNN and Residual Connections

Now we head back to Google's translation system GNMT. Its construction was largely the same as the one Cho et al. proposed in the previous section, with modifications to its LSTM layers. Each layer's output will be merged by its previous layer's output to become the input for the next layer. By doing so allows the network to be expanded to more layers with feasible training speed and accuracy.



Above: Detailed construction for the encoder and stacked LSTM with residual connections in Google's system.

Moreover, similar to previous works, the output produced by the decoder, which was in a conditional probability format, was sent to beam search to generate the final

translated sentence. However, Google's team introduced a more sophisticated scoring based on empirical data for the beam search.

Solution to Fixed Vocabulary - Wordpiece Model

One of the most common challenges to natural language processing applications is the out of vocabulary (OOV) problem. Most language and encoding models are restricted to fixed size vocabulary, and must incorporate methods generalize to more words and lower the potential hazard of high error rate on production data. Some common solutions includes backoff methods and using a subword unit model. The same problem also applies to neural machine translation, which starts by encoding one-hot vocabulary vectors and convert it to continuous space word representations (similar to the construction of word2vec).

In recent neural machine translation research, Sennrich et al. from University of Edinburgh first described a Byte Pair Encoding approach to aggregate frequent adjacent character pairs into single n-grams subwords.[1] Similar approaches have be used in speech recognition and voice search as well. A more detailed implementation was described in Schuster and Nakajima, applied in Japanese and Korean voice search.[2]

The algorithm first initialize a language model with an inventory of basic subword units (i.e. characters or smallest word fragments) on the training set. Next, the create a new unit by combining two subword units and add the unit that maximizes the likelihood of the language model to the inventory, until the expected vocabulary size is matched. The Wordpiece model used in Google's system has a subword vocabulary size of 32,000.[3]

References

1. R.Sennrich, B. Haddow, A. Birch. Neural Machine Translation of Rare Words with Subword Units. ACL, 2016.
2. M. Schuster and K. Nakajima. Japanese and Korean voice search. ICASSP, 2012.
3. Y. Wu, M. Schuster, Z. Chen, Q. Le, M. Norouzi and the Google Brain team. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144v2 [cs.CL] 8 Oct 2016

Summarizing Google's Neural Machine Translation System

Now we can fully understand the underlying techniques of Google's Neural Machine Translation System. The bidirectional encoder, stacked RNN with residual connections, the Attention Mechanism and Wordpiece model contributed substantially to the success of high quality translation in production data. Furthermore, Google also added human evaluation to their system in addition to BLEU scores and achieved an average quality increase of 60% compared to conventional phrase based systems in several languages.

It is remarkable that machine translation has gone this far and this close to human strength in such a complicated task. We can observe this in the following round-trip translation (RTT) result. Though not a good way to evaluate a machine translation systems, the RTT demonstrated by Google Translate is pretty accurate.

The English to Chinese to English result of Google Translate
The survey covers a number of important works in the area of neural machine translation, a rapidly evolving field that transcends the latest advances in conventional statistical machine translation results. We start by introducing the most advanced machine translation system that Google Research released in October 2016. We then examined the details of each component of the proposed system by crawling it and traced back to earlier breakthroughs that contributed to these components, which, in the end, combined to produce astonishing results that broadened the ability of neural machine translation.
The English to Chinese to English to Chinese to English to Chinese result of Google Translate
該調查涵蓋了神經機器翻譯領域的一些重要作品，這是一個快速發展的領域，超越了傳統統計機器翻譯的最新進展。我們首先介紹Google Research於2016年10月發布的最先進的機器翻譯系統。然後，我們通過爬行和追溯到對這些組件做出貢獻的早期突破，審查了提出的系統的每個組件的詳細信息。該組合產生令人驚訝的結果並且拓寬了神經機器翻譯的能力。

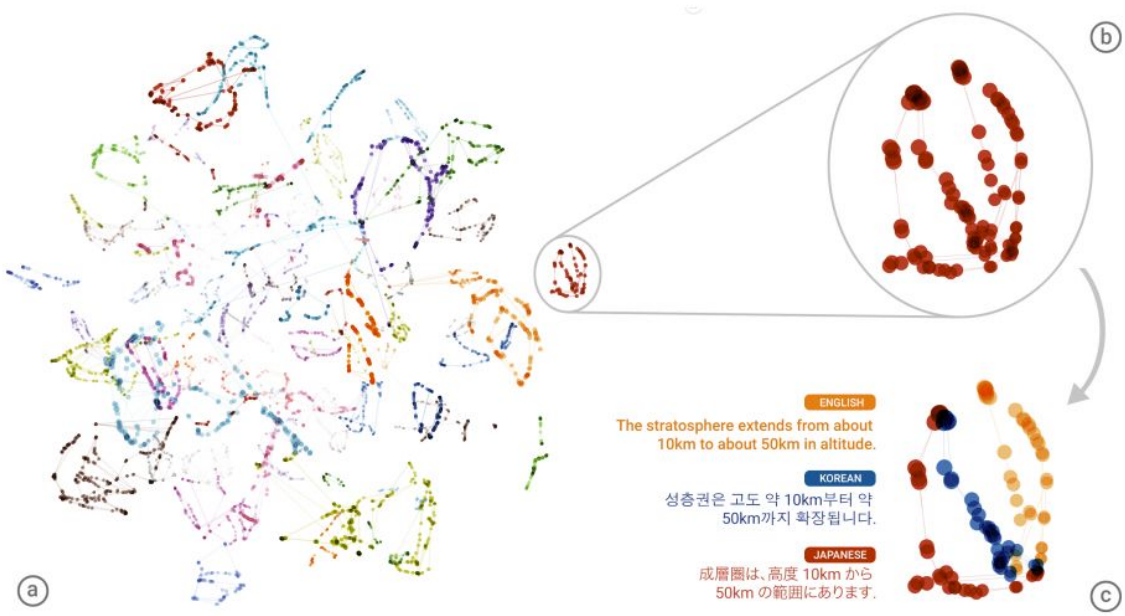
Multitask / Transfer Learning in Multi-lingual Machine Translation

Following the impressive result of GNMT, the Google Brain team continued to investigate the prowess of this model. In Nov. 2016 they published another article, describing the capability of training one to many, many to one and even many to many languages in one single model.[1] Their approach was simply adding a token in front of the source sentence indicating the desired target language, without even mentioning the source language. Surprisingly, Google's system can achieve with little or no loss in its translation quality (BLEU scores, in this case). Moreover,

languages with less data available can benefit from multi-lingual training by observing more indirect samples.

Another interesting discovery was the model's flexibility that allows zero-shot learning. For example, if trained with English↔Portuguese and English↔Spanish, the model can inference reasonable Portuguese→Spanish translations. This is called zero-shot because no direct knowledge or explicit data was given to the model. Thought the result was not as good as bridged (i.e. Portuguese translate to English then Spanish) NMT models, it can be enhanced with incremental training, which utilizes significantly less data than models with only one language pair, to achieve the same level of quality.

The last important fact discovered was the similarities in the representation of sentences in different languages. They extracted the attention vector (the representation of a sentence at a certain translation state in the attention network) of sentences with similar meanings in different languages. It turns out that these sentences tends to form into a cluster, which also is potentially an indication of better translation scores. This can also be generalized into the notion that the system itself would learn a intermediate language so that it could translate into multiple language pairs.



Above: Sentence representations visualized by Google. [1]

References

1. M. Johnson, M. Schuster, Q. Le, M. Krikun, Y. Wu and the Google Brain team. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv:1611.04558v1 [cs.CL] 14 Nov 2016

Reflections and Future Developments

In this survey we covered several milestones in NMT and presented the elegant system developed by Google. Still, here are some interesting topics closely related to NMT but not covered in this survey, including multi-lingual speech recognition[1] and real-time translation via reinforcement learning[2]. Overall, the advances in NMT provided as various useful tools to apply in not only natural language processing and understanding, but also other machine learning applications.

We would also like to address some critical reviews on the research done so far. First of all, is it possible to directly use sentence information (such as its vector representation similarities) to optimize translation quality if the clustering of attention vectors are indeed directly related to translation quality among different languages.

In addition, are there other ways of representing a sentence other than using bidirectional neural networks and attention mechanism, and process the entire sentence at once. If such method exists, can it be used in tasks such as abstractive summarization? On the other hand, if chances are that the system can correct its previous translation when given more context, can it optimize its own translation iteratively, like a professional human translator would do in practice. Might this create then modify approach be helpful in bridging the final gap with human translation quality?

More broadly speaking, there are still lots of unsolved mystery related to the powerfulness of recurrent neural networks[3]. For example, there are times that researchers are surprised by how powerful a model can be and not expecting it to work on certain difficult tasks. This leads to the question that can we measure a neural network's capability of learning, or capacity of learning given its architecture beforehand? Also, there will be continuing debate on creating a single end-to-end model with simple intuitions or develop sophisticated models with detailed calculations.

Finally, with the superb computation power, enormous amount of data and ever so complicated models, are we just storing all these nonlinear mappings of inputs and outputs into more and more parameters, or really on the course of discovering a universal learning algorithm. From another point of view, is the supervised learning scope too narrow to create such an algorithm, and should we focus more on developing an algorithm that can automatically generalize and extend itself?

References

1. A.Graves and N. Jaitly. Towards End-To-End Speech Recognition with Recurrent Neural Networks. ICML, 2014.
2. J. Gu, G. Neubig, K. Cho and V. Li. Learning to Translate in Real-time with Neural Machine Translation. arXiv:1610.00388v3 [cs.CL] 10 Jan 2017
3. A. Karpathy. The Unreasonable Effectiveness of Recurrent Neural Networks.
4. Neural Machine Translation (ACL 2016 Tutorial)
5. Y. Bengio, A. Courville and P.I Vincent. Representation Learning: A Review and New Perspectives. arXiv:1206.5538v3 [cs.LG] 23 Apr 2014