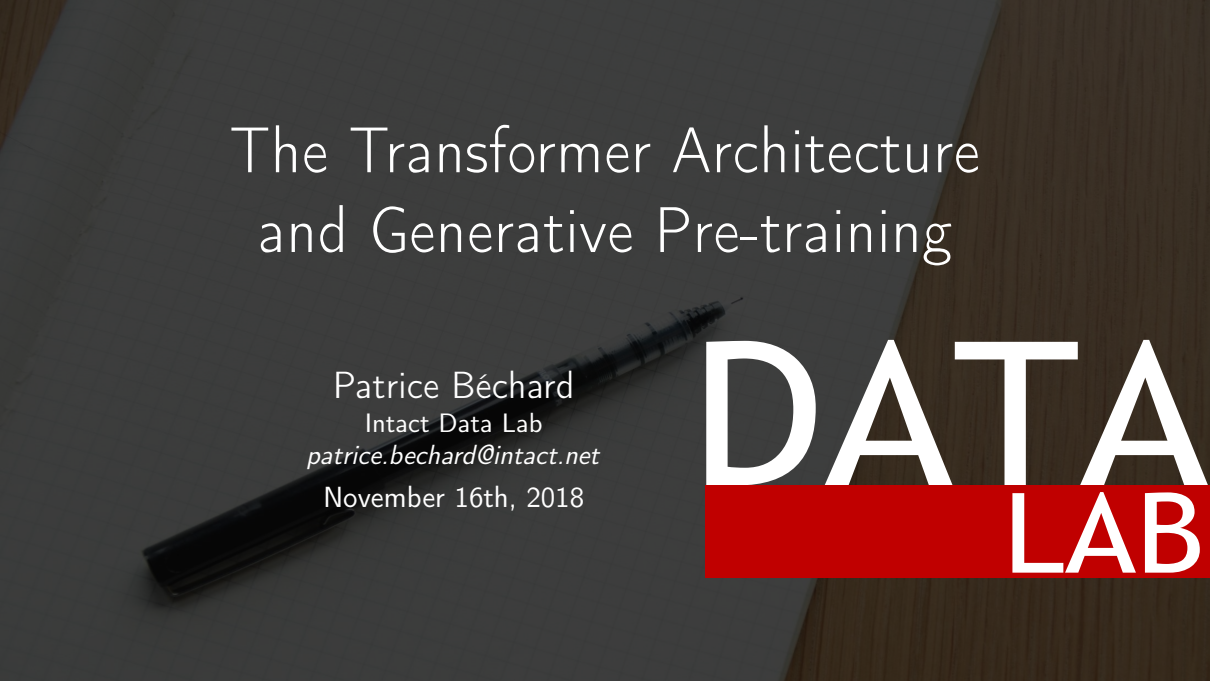
The Transformer Architecture and Generative Pre-training

Patrice Béchard
Intact Data Lab
patrice.bechard@intact.net

November 16th, 2018



Why are these recent advances important ?

- Classic Neural Machine Translation (NMT) approaches are difficult to parallelize and take a long time to train.
- Annotated datasets are expensive to build (time and resources).
- We have an *infinite* amount of raw textual data.

Plan

Neural Machine Translation

The Transformer Architecture

Word Embeddings

Generative Pre-training

Going Further

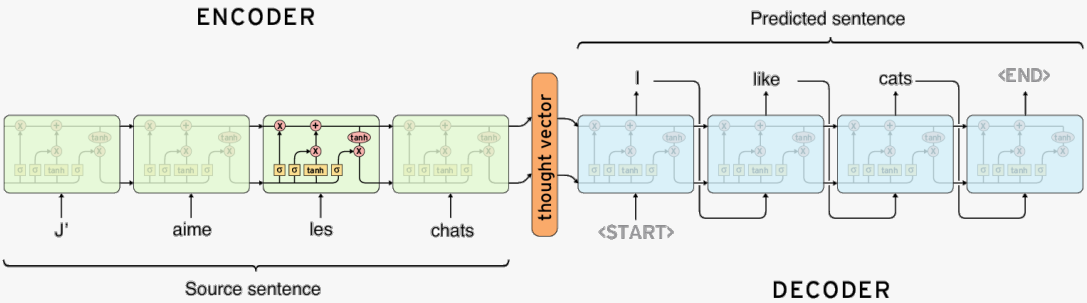
Neural Machine Translation

Neural Machine Translation

The task

Neural Machine Translation

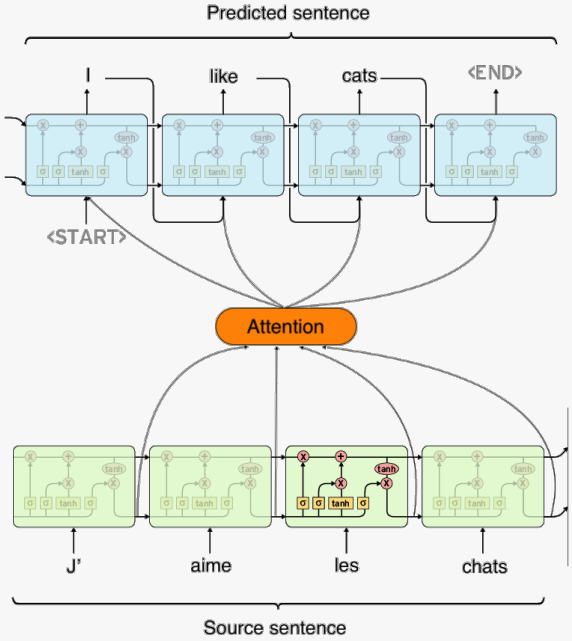
The Sequence to Sequence (seq2seq) model



Neural Machine Translation

The Attention Mechanism

- Shortcut between word from source sentence and target sentence



Neural Machine Translation

Limitations

The Transformer Architecture

The Transformer Architecture

Word Embeddings

Idea

Word Embeddings

Mikolov's Skip-gram and CBOW

Word Embeddings

Other approaches (GLOVE, ELMo)

Generative Pre-training

Idea

Generative Pre-training

Two Approaches to Transfer Learning

Generative Pre-training

OpenAI-GPT

Generative Pre-training

BERT

Generative Pre-training

BERT

Going Further

References I