BERT, Transformers, AdaNet

Corinna Cortes, Google Research NY

Outline

- BERT, Bidirectional Encoder Representations from Transformers
    - **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, October 2018
- Transformers
    - **Attention Is All You Need**, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, NIPS 2017
- AdaNet
    - **AdaNet: Adaptive Structural Learning of Artificial Neural Networks**, Corinna Cortes, Xavi Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, Scott Yang, ICML 2017

BERT, Bidirectional Encoder Representations from Transformers

What: State-of-the-Art architecture for 11 NLP tasks

“GLUE benchmark to 80.4% (7.6% absolute improvement), MultiNLI accuracy to 86.7 (5.6% absolute improvement) and the SQuAD v1.1 question answering Test F1 to 93.2 (1.5% absolute improvement), outperforming human performance by 2.0%.”

How: Transformer-like architecture, multi faceted costfuntion, fine-tuning of pre-trained model

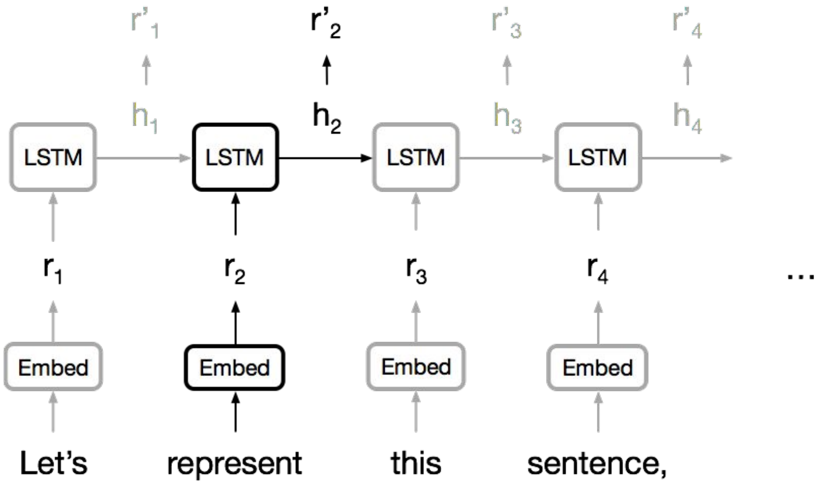
Contextual Word Representation

How to represent words **in a sentence**?

- The same word can have different meanings in different context.
    - He was in a play on Broadway.
    - Do you want to come out and play?
    - She didn't play a role in the accident.

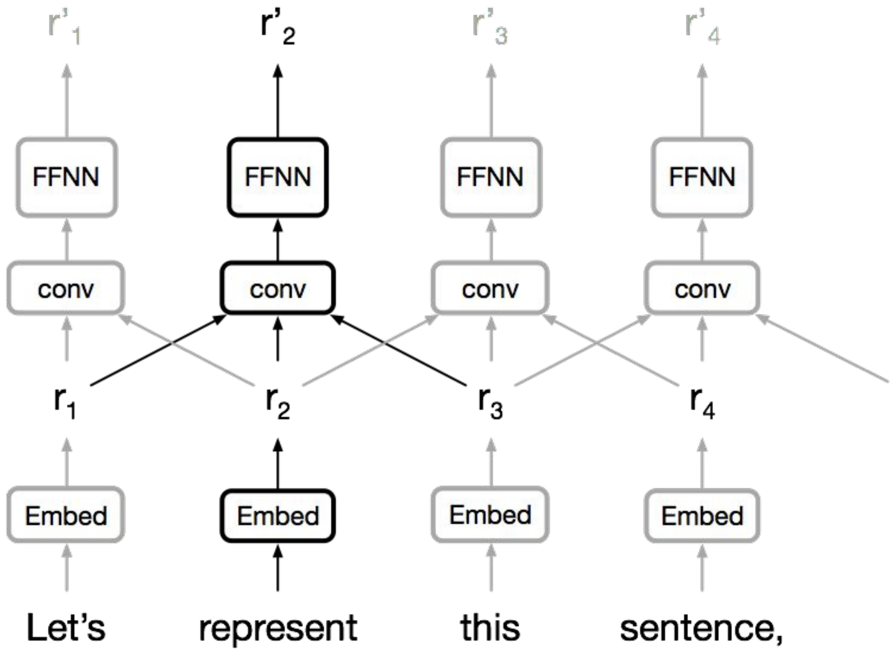
BERT, motivations

- RNN, LSTMs are common NLP models for structured and sequence prediction
- They are **uni-directional**
  - The training procedure for LM
    - Load a sequence of words
    - Use “history” representation to predict “future” words
- They are **not great for fine-tuning** for down-stream task
  - Many down-stream tasks require bidirectional context. Cannot mix the “history” and “future”
- **Cannot be parallelized**



Convolutional Neural Networks

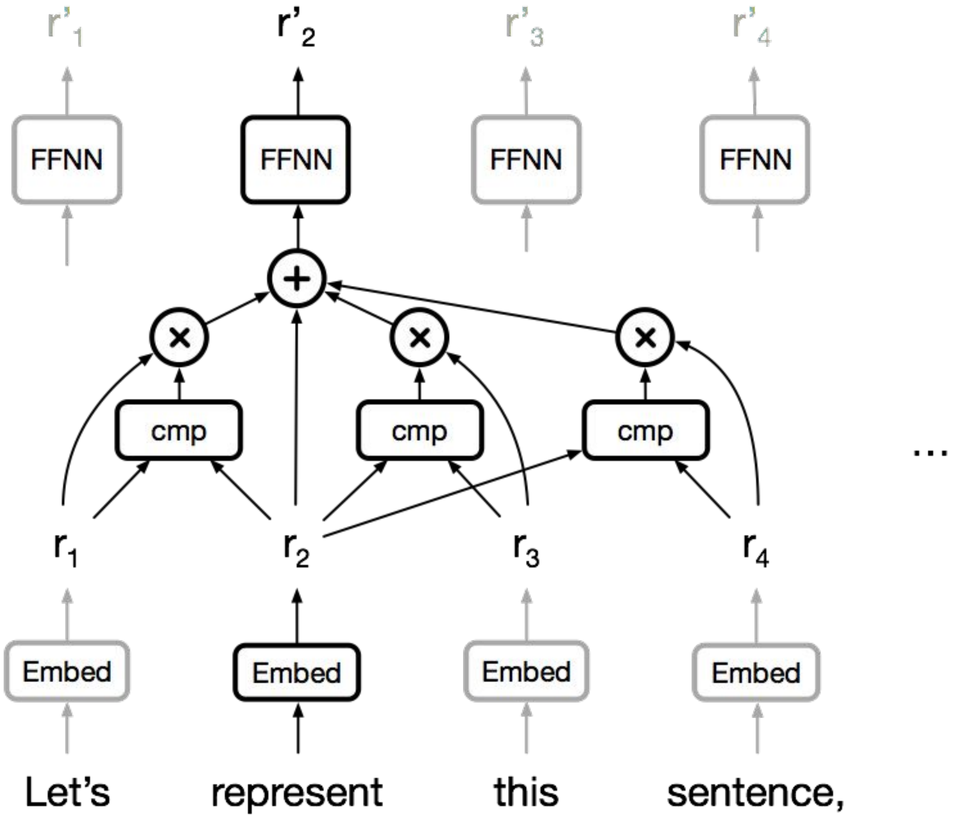
- Trivial to parallelize (per layer)
- Fit intuition that most dependencies are local
- ‘Path length’ between positions constant or logarithmic
- Long-distance dependencies require many layers



Self-attention

- Constant ‘path length’ between any two positions
- Global perceptive field
- Trivial to parallelize (per layer)

Self-attention



Multi-faceted loss function

Masked Language Model, Cloze task, (Taylor, 1953)

Randomly select 15% of words, replace the input word:

- 80% of the time: Replace the word with the [MASK] token, e.g.,
  my dog is hairy → my dog is [MASK]
- 10% of the time: Replace the word with a random word, e.g.,
  my dog is hairy → my dog is apple
- 10% of the time: Keep the word unchanged, e.g.,
  my dog is hairy → my dog is hairy. The purpose of this is to bias the representation towards the actual observed word.

Multi-class classification task on word-pieces

Multi-faceted loss function

Next Sentence Prediction, NSP

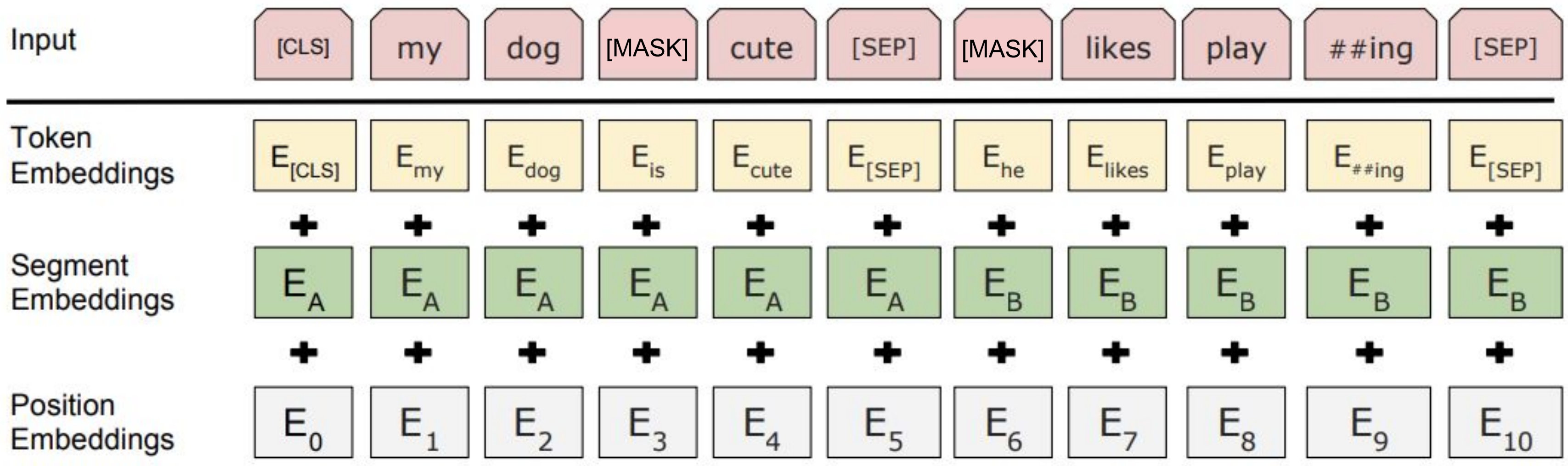
Paired sentences:

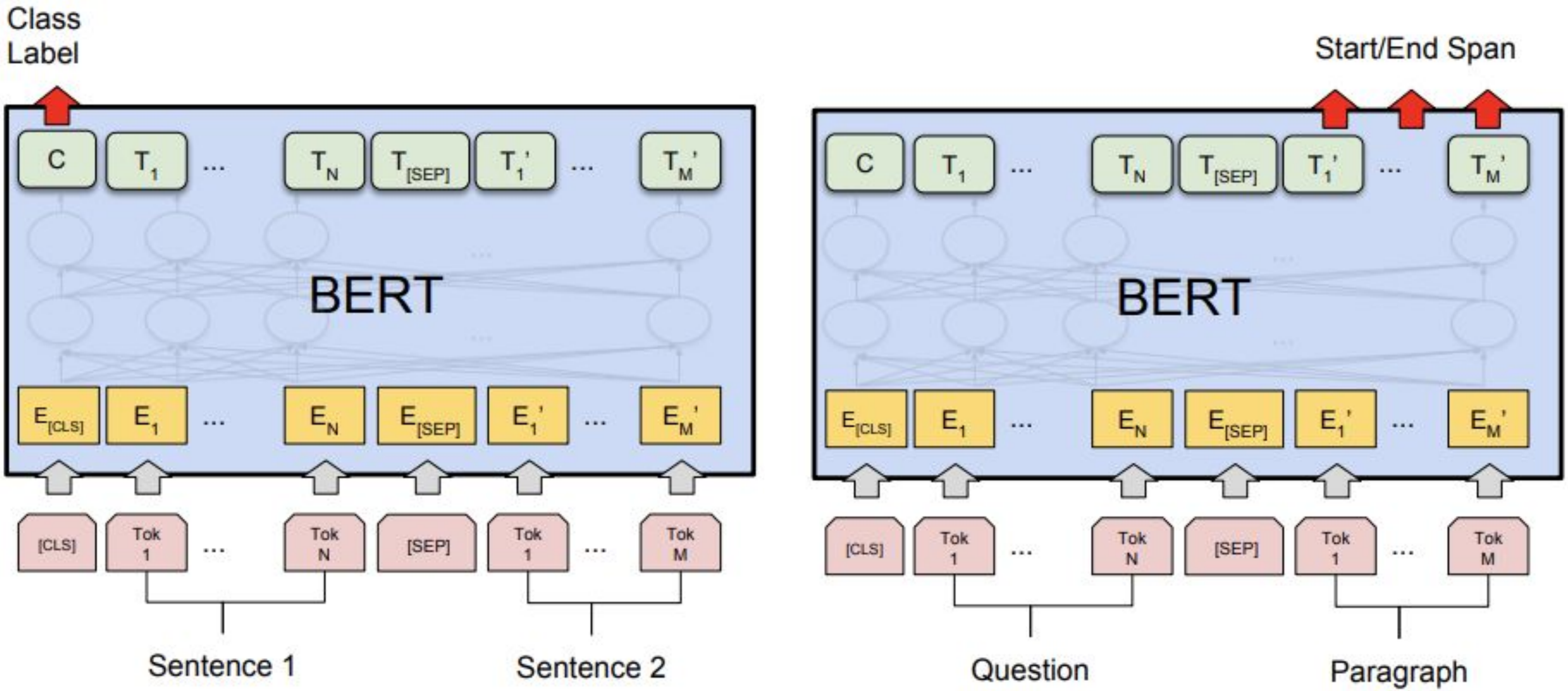
- 50% of the time: true next sentence
- 50% of the time: false next sentence

Binary classification task on IsNext

The two losses are added with equal weight.

Input representation





(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(c) Question Answering Tasks: SQuAD v1.1

Ablation studies

Yes, NSP helps

Yes, big models are good

Yes, large number of training steps is good

Yes, training is slow, but not that slow

Try BERT yourself

Open-source code available at

http://goo.gl/language/bert

Transformers

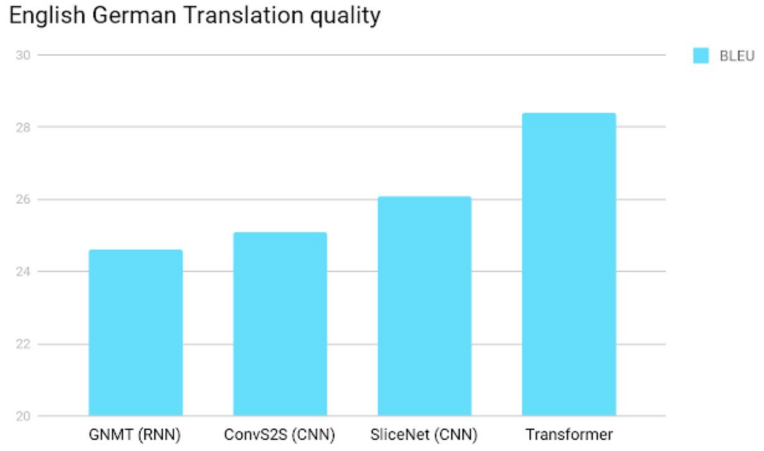
What:

State-of-the-Art on sequence prediction Machine Translation

- “Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU.”

How:

- Multi-headed attention models



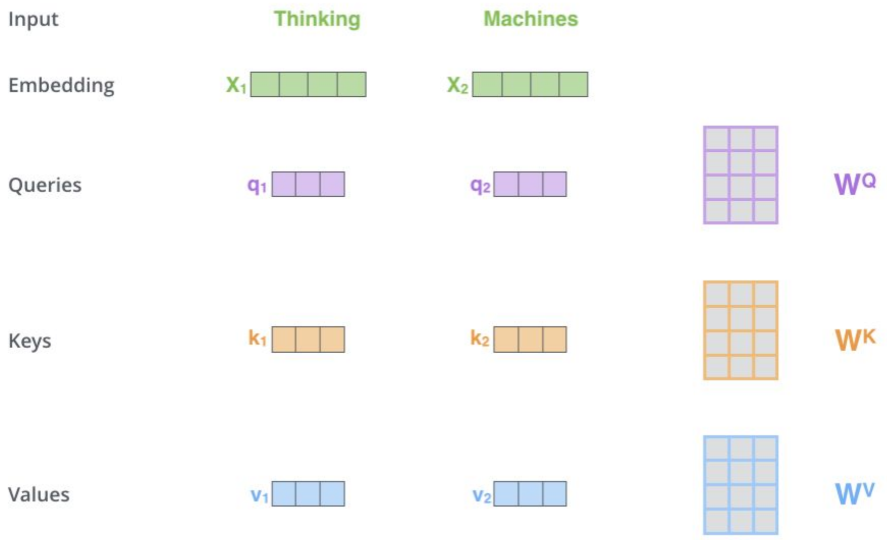
BLEU scores (higher is better) of single models on the standard WMT newstest2014 English to German translation benchmark.

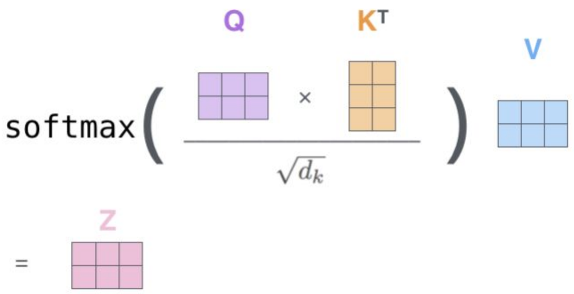
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Self-attention

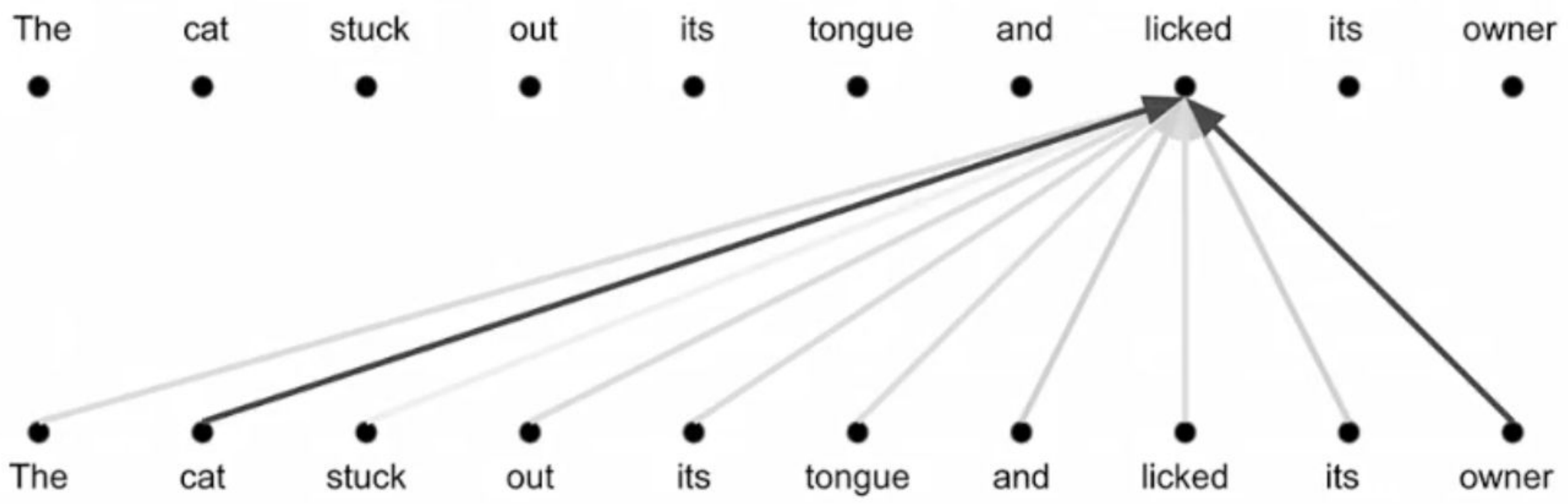
http://jalammar.github.io/illustrated-transformer/

Queries, Keys, and Values



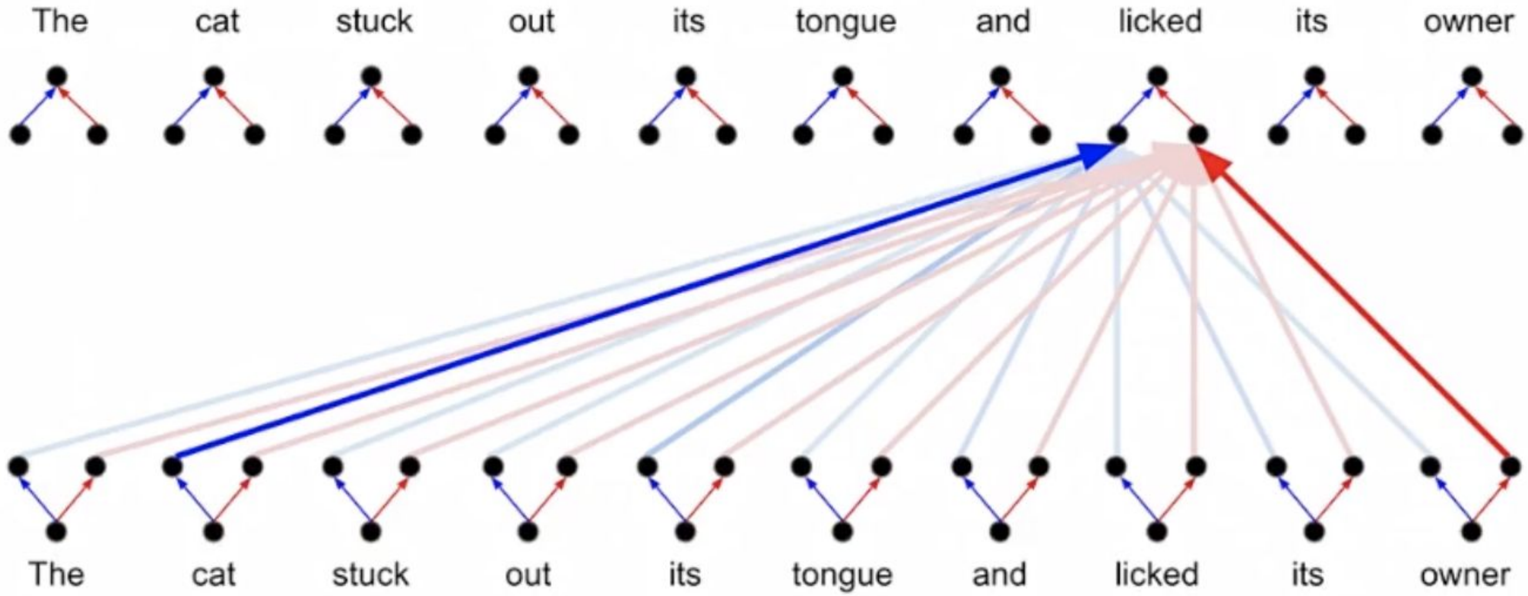


Attention: a weighted average

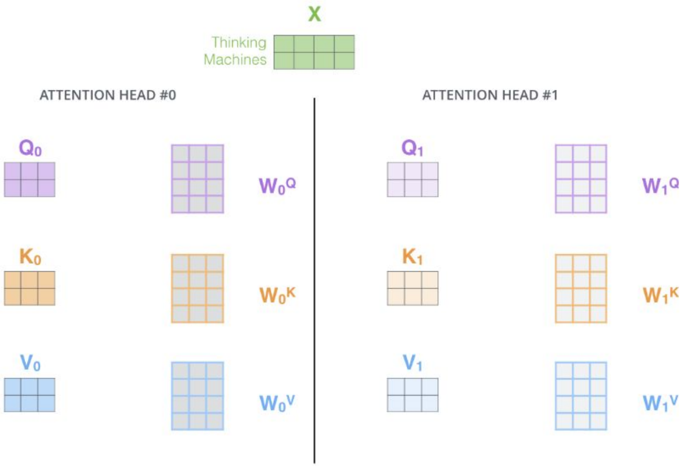


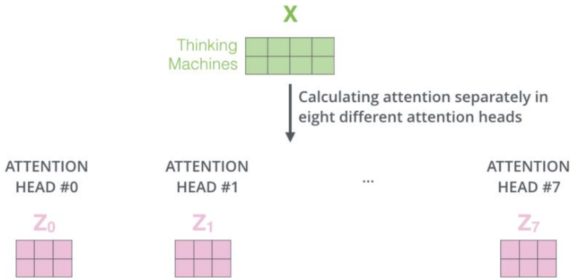
Multi-head Attention

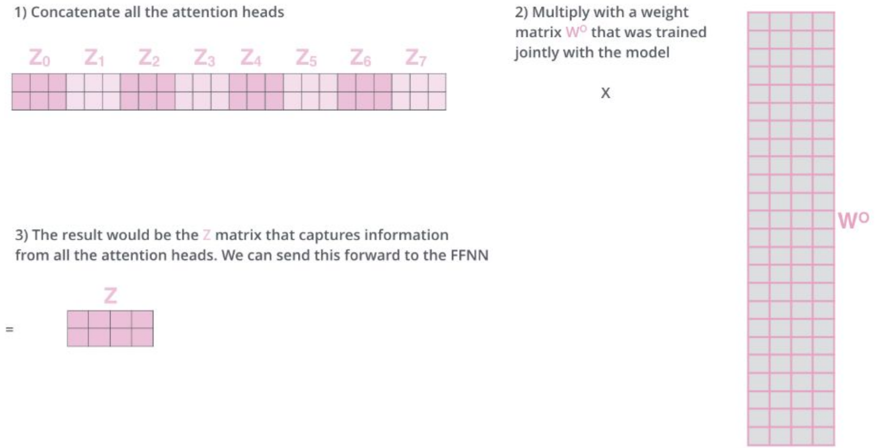
Parallel attention layers with different linear transformations on input and output.



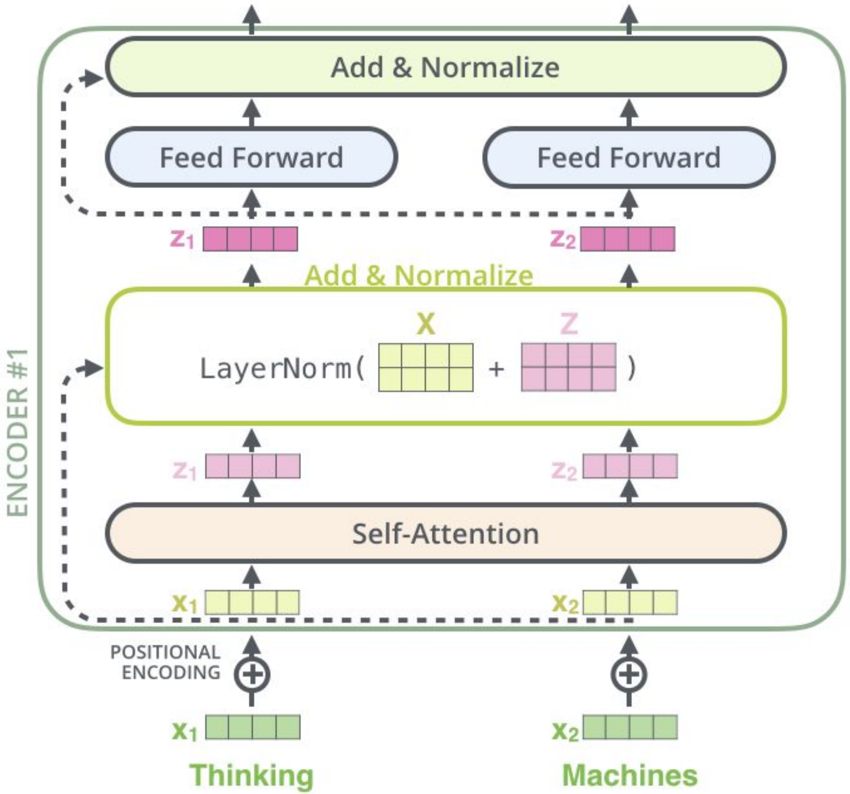
Multi-headed self-attention

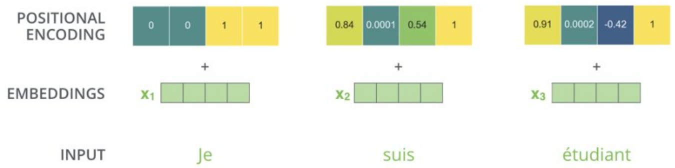




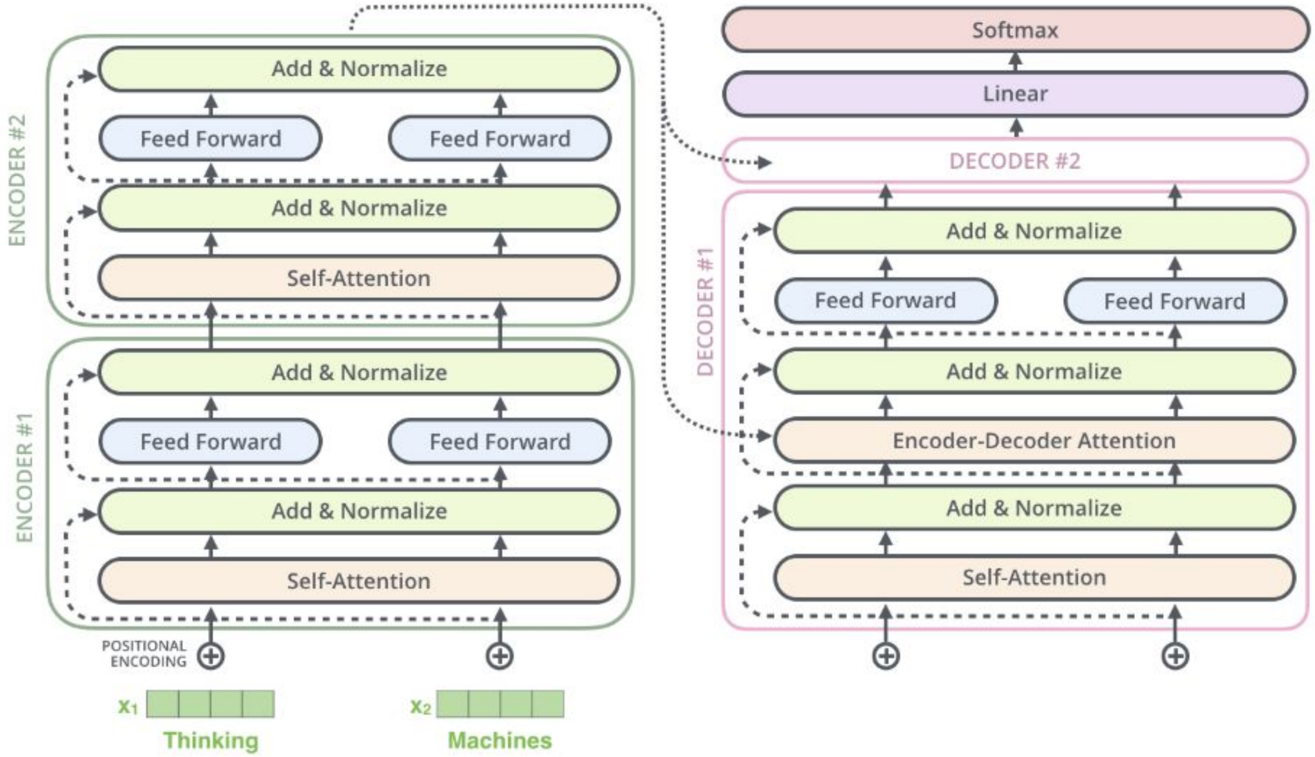


Self-attention, stacking





Encoding and decoding



Animated

Try Transformers yourself

Open-source code available at

https://github.com/tensorflow/tensor2tensor/

http://nlp.seas.harvard.edu/2018/04/03/attention.html

…

AdaNet

What: AdaNet is an adaptive algorithm for learning a neural architecture as an **ensemble** of **subnetworks**.

How: theoretically founded complexity terms to guide the construction.

A new approach: adaptive and iterative learning

AdaNet objective

Optimize mixture weights ***w*** to balance trade-off between **empirical error** and **network complexity**.

$$\mathrm{Loss}(\mathbf{w}) = \mathrm{Error}\left(\sum_{j=1}^{N} w_j h_j\right) + \sum_{j=1}^{N} |w_j| \mathrm{Complexity}(h_j)$$

AdaNet learning guarantees

1. The **generalization error** of the ensemble **is bounded** by optimizing the AdaNet objective [Cortes et. al, ‘17].
2. We are **directly minimizing** the bound on the generalization error.

Complexity

Some options:

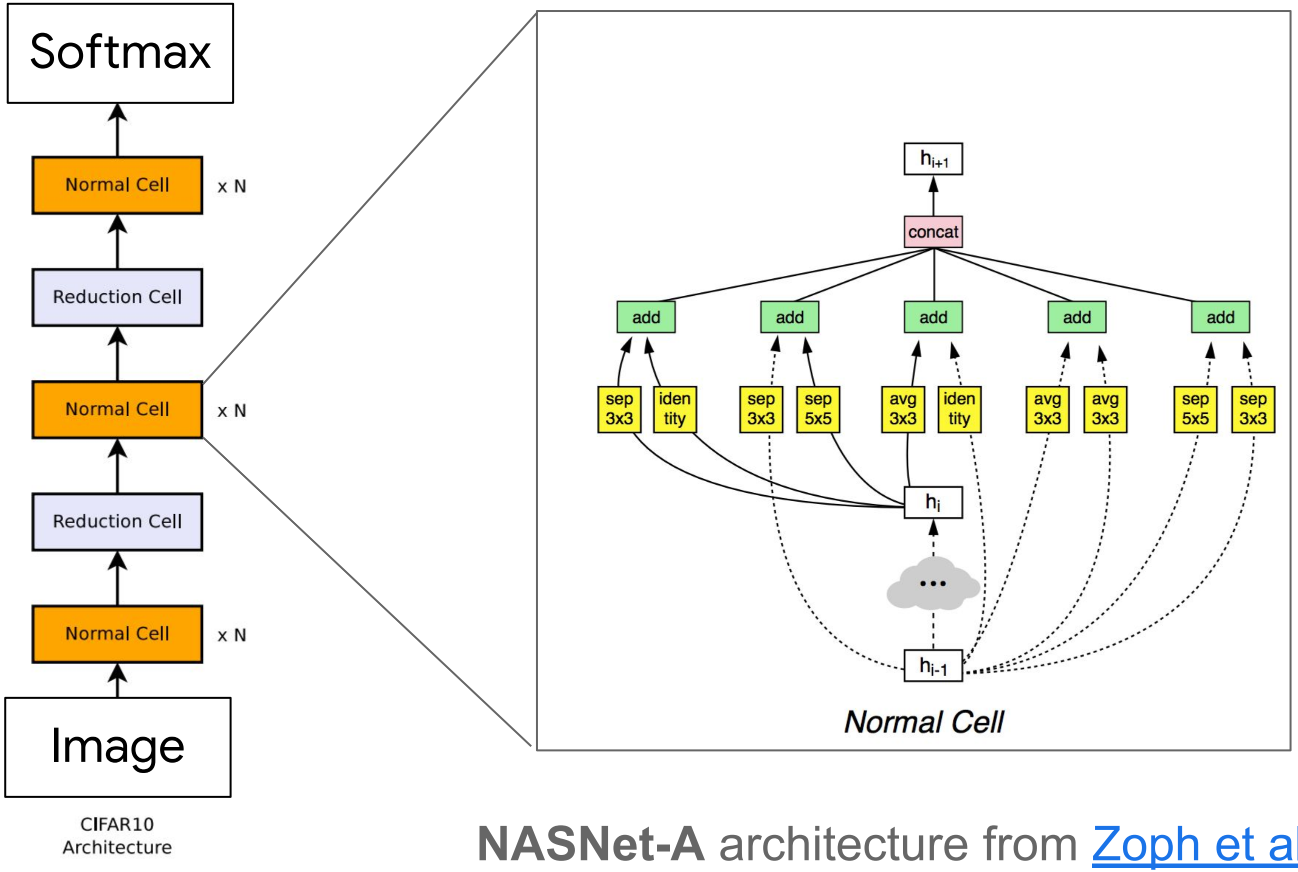
- Rademacher complexity upper-bound.
- Variance of subnetwork outputs.
- Norm of input-output Jacobian [Novak et al, ‘18].

AdaNet.CNN

- AdaNet extended to **convolutional** subnetworks.
- What kind of CNN building block to use?
    - Simple convolutions.
    - Strong prior.

NASNet-A Architecture

NASNet-A Architecture



NASNet-A architecture from [Zoph et al., '17]()

Classification error on CIFAR-10 and CIFAR-100.

Model	CIFAR-10	Params	CIFAR-100	Params
NASNet-A (6 @ 768)	2.65%*	3.3M	18.1%	3.4M
NASNet-A (7 @ 2304)	2.40%*	27.6M	15.95%	34.6M

Results marked with (*) from Zoph et al., '17.

Google

Classification error on CIFAR-10 and CIFAR-100.

Model	CIFAR-10	Params	CIFAR-100	Params
NASNet-A (6 @ 768)	2.65%*	3.3M	18.1%	3.4M
NASNet-A (7 @ 2304)	2.40%*	27.6M	15.95%	34.6M

Results marked with (*) from [Zoph et al., '17](#).

AdaNet x NASNet

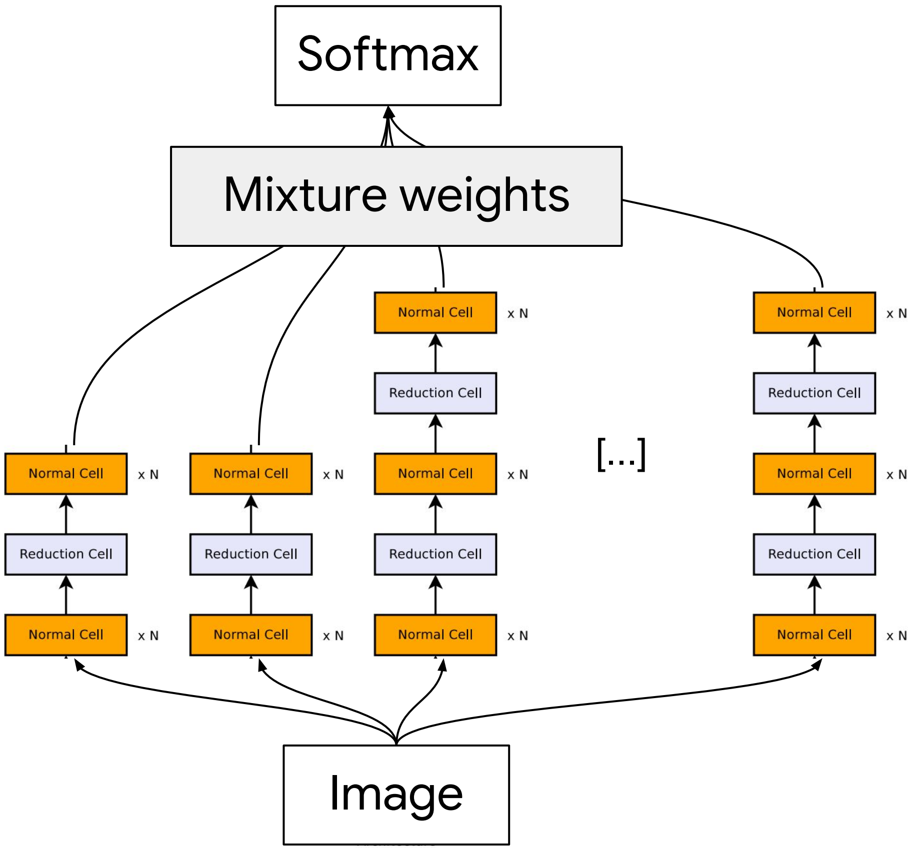
Complementary AutoML

- AdaNet can benefit from other ML algorithms.
- For example, it can learn to grow a NASNet subnetwork and provide **learning guarantees**.

Google

AdaNet + NASNet

NASNet-A
(6 @ 768)



Google

Classification error on CIFAR-10 and CIFAR-100.

Model	CIFAR-10	Params	CIFAR-100	Params
NASNet-A (6 @ 768)	2.65%*	3.3M	18.1%	3.4M
NASNet-A (7 @ 2304)	2.40%*	27.6M	15.95%	34.6M
AdaNet	**2.30%**	**26.4M**	**14.37%**	**30.7M**

4%-10% reduction in error!

Does this extend to other datasets?

AdaNet is easy to use

Before

```python
import tensorflow as tf

estimator = tf.estimator.Estimator(model_fn=my_model_fn)

tf.estimator.parameterized_train_and_evaluate(estimator)
```

Google

After go/try-adanet

```
import adanet
import tensorflow as tf

estimator = adanet.Estimator(MySubnetworkGenerator(my_model_fn))


tf.estimator.parameterized_train_and_evaluate(estimator)
```

For everyone!

https://github.com/**tensorflow/adanet**

Combining multiple TensorFlow Hub modules into one ensemble network with AdaNet