Tensor2Tensor Transformers

TensorFlow for Deep Learning Research

Łukasz Kaiser

Based on *Attention Is All You Need* by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin and other works with Samy Bengio, Eugene Brevdo, Francois Chollet, Stephan Gouws, Nal Kalchbrenner, Ofir Nachum, Aurko Roy, Ryan Sepassi.

Why? Some context.

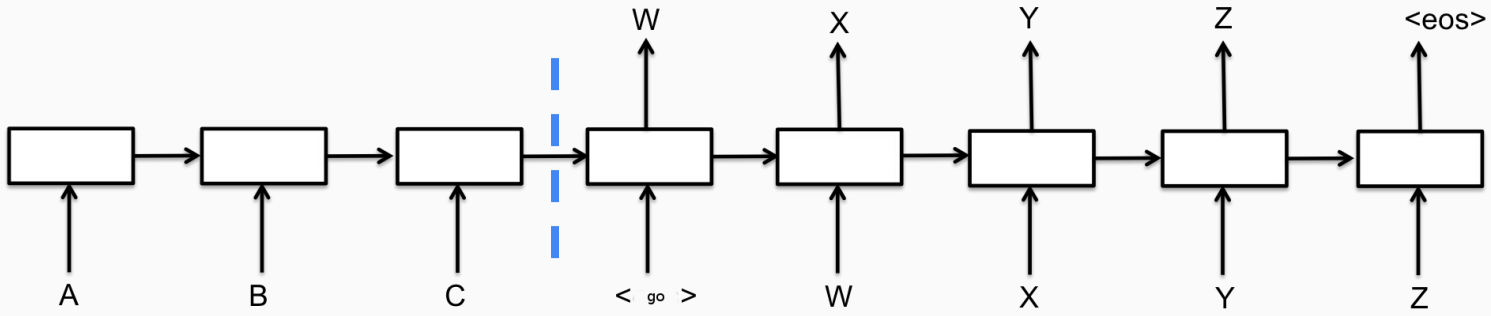
How Deep Learning Quietly Revolutionized NLP (2016)

What NLP tasks are we talking about?

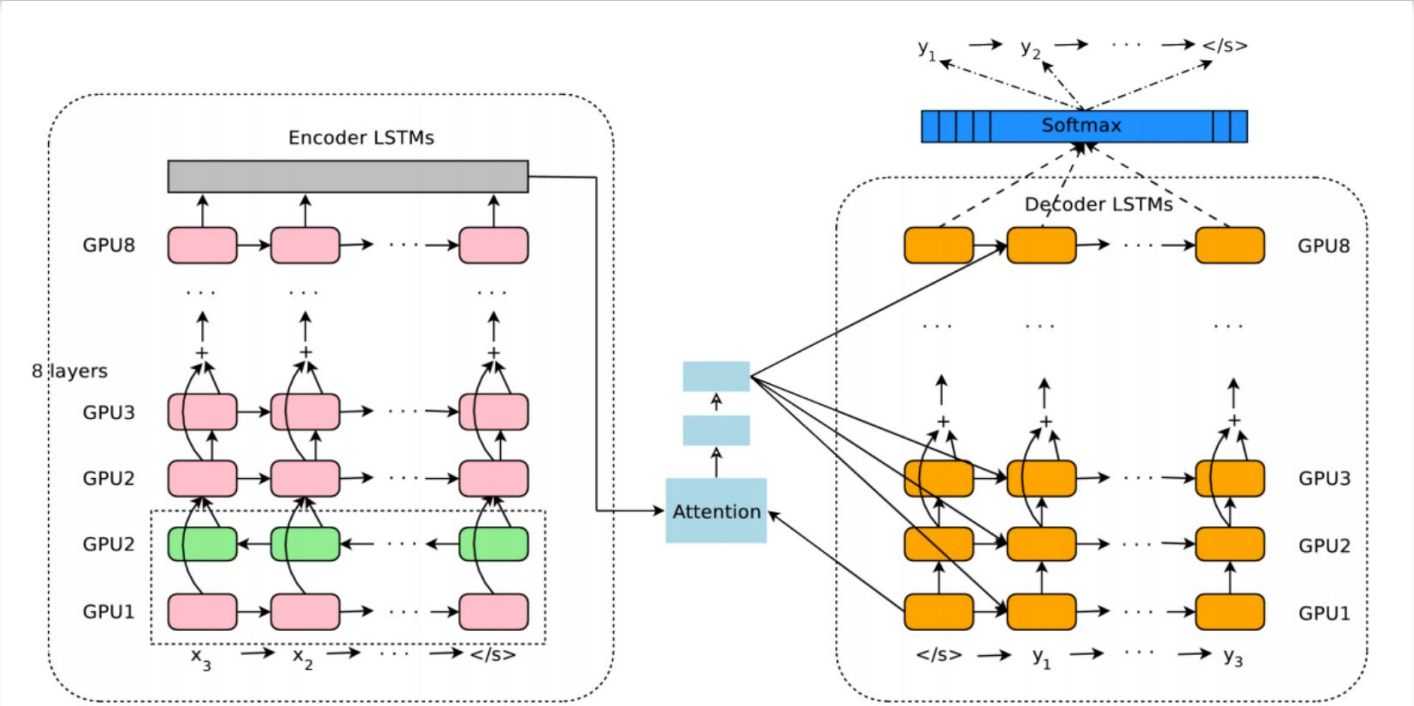
Part Of Speech Tagging	Assign part-of-speech to each word.
Parsing	Create a grammar tree given a sentence.
Named Entity Recognition	Recognize people, places, etc. in a sentence.
Language Modeling	Generate natural sentences.
Translation	Translate a sentence into another language.
Sentence Compression	Remove words to summarize a sentence.
Abstractive Summarization	Summarize a paragraph in new words.
Question Answering	Answer a question, maybe given a passage.
….	

Can deep learning solve these tasks?

- Inputs and outputs have variable size, how can neural networks handle it?
- Recurrent Neural Networks can do it, but how do we train them?
- Long Short-Term Memory [Hochreiter et al., 1997], but how to compose it?
- Encoder-Decoder (sequence-to-sequence) architectures
  [Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014]

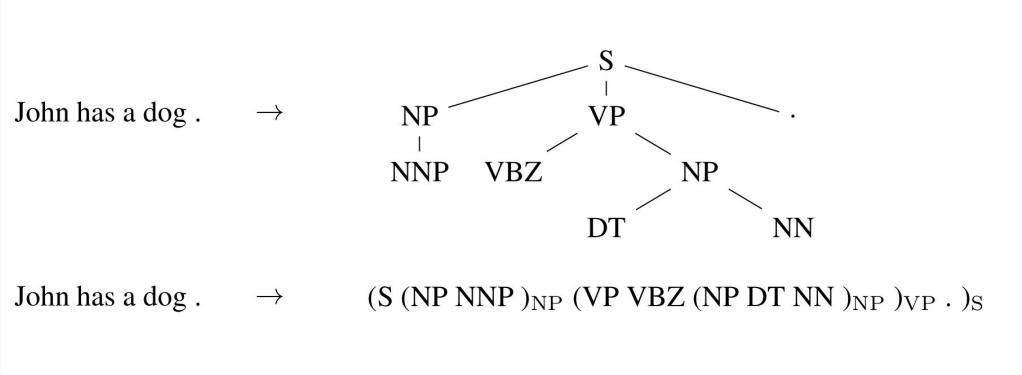


Advanced sequence-to-sequence LSTM



Parsing with sequence-to-sequence LSTMs

(1) Represent the tree as a sequence.



(2) Generate data and train a sequence-to-sequence LSTM model.

(3) Results: 92.8 F1 score vs 92.4 previous best [Vinyals & Kaiser et al., 2014]

Language modeling with LSTMs

Language model performance is measured in perplexity (lower is better).

• Kneser-Ney 5-gram:	67.6	[Chelba et al., 2013]
• RNN-1024 + 9-gram:	51.3	[Chelba et al., 2013]
• LSTM-512-512:	54.1	[Józefowicz et al., 2016]
• 2-layer LSTM-8192-1024:	30.6	[Józefowicz et al., 2016]
• 2-l.-LSTM-4096-1024+MoE:	28.0	[Shazeer & Mirhoseini et al., 2016]

Model size seems to be the decisive factor.

Language modeling with LSTMs: Examples

Raw (not hand-selected) sampled sentences: [Józefowicz et al., 2016]

About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortege .

It is now known that coffee and cacao products can do no harm on the body .

Yuri Zhirkov was in attendance at the Stamford Bridge at the start of the second half but neither Drogba nor Malouda was able to push on through the Barcelona defence .

Sentence compression with LSTMs

Example:

Input: *State Sen. Stewart Greenleaf discusses his proposed human trafficking bill at Calvery Baptist Church in Willow Grove Thursday night.*

Output: Stewart Greenleaf discusses his human trafficking bill.

Results:	readability	informativeness
MIRA (previous best):	4.31	3.55
LSTM [Filippova et al., 2015]:	4.51	3.78

Translation with LSTMs

Translation performance is measured in BLEU scores (higher is better, EnDe):

- Phrase-Based MT: 20.7 [Durrani et al., 2014]
- Early LSTM model: 19.4 [Sébastien et al., 2015]
- DeepAtt (large LSTM): 20.6 [Zhou et al., 2016]
- GNMT (large LSTM): 24.9 [Wu et al., 2016]
- GNMT+MoE: 26.0 [Shazeer & Mirhoseini et al., 2016]

Again, model size and tuning seem to be the decisive factor.

Translation with LSTMs: Examples

German:

Probleme kann man niemals mit derselben Denkweise lösen, durch die sie entstanden sind.

PBMT Translate:

No problem can be solved from the same consciousness that they have arisen.

GNMT Translate:

Problems can never be solved with the same way of thinking that caused them.

Translation with LSTMs: How good is it?

Google Translate production data, median score by human evaluation on the scale 0-6. [Wu et al., ‘16]

	PBMT	GNMT	Human	Relative improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

That was 2016. Now.

How to generate almost anything?

Good long sequence model:

- Has global structure and local structure (e.g., music)
- Powerful enough for complex functions (e.g., <fr><en>...)
- Captures long-range dependencies (e.g., reuse a name)
- Remembers rare occurrences (e.g., one-shot learning)
- Correlates across modalities (e.g., image+text together)

What does a good long sequence model bring?

- Text: a story-teller, translator if conditioned.
- Image: a painter, drawing tool if conditioned.
- Music: a composer, companion if conditioned.
- Games: a simulator, possibly changing RL techniques.
- Video: a world model, or a fake news tool?

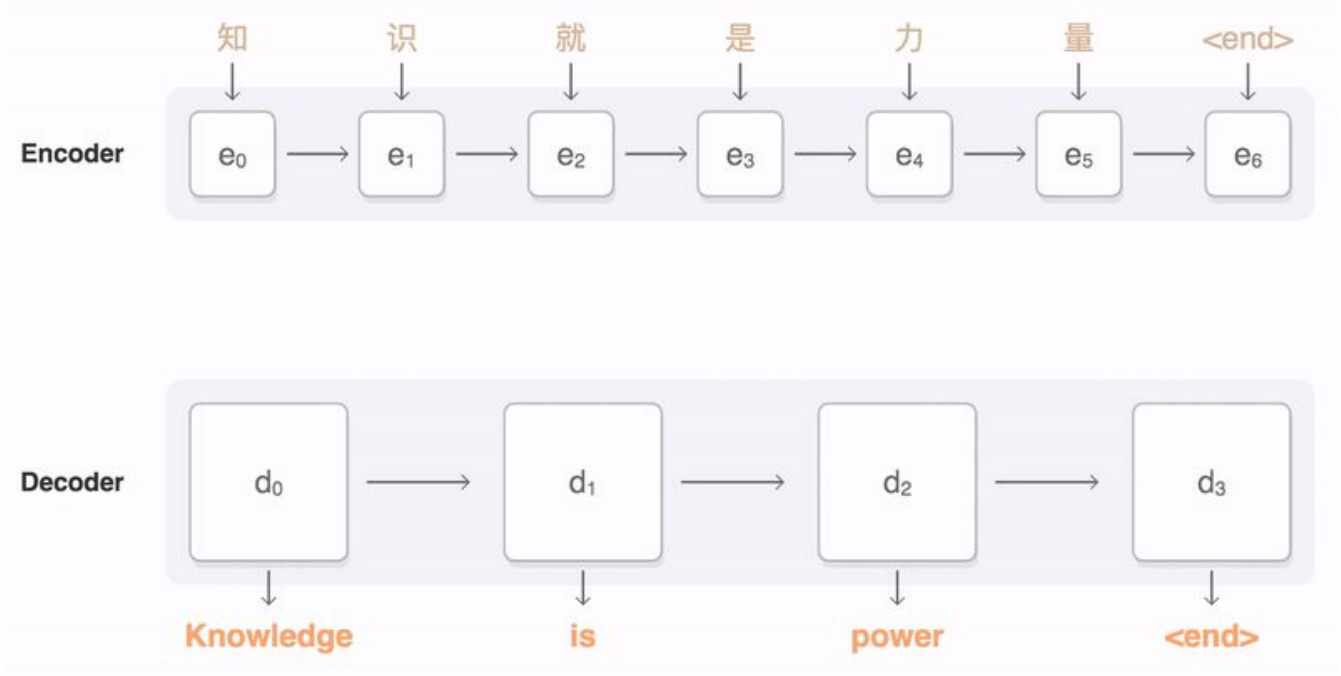
How

Text

(*Attention is All You Need*)

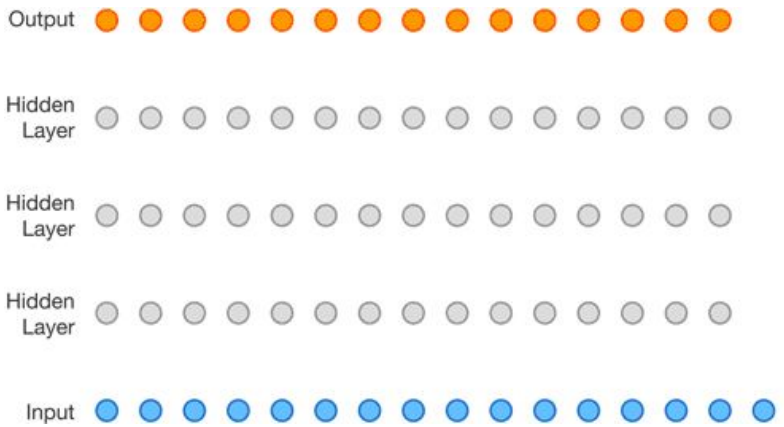
RNNs Everywhere

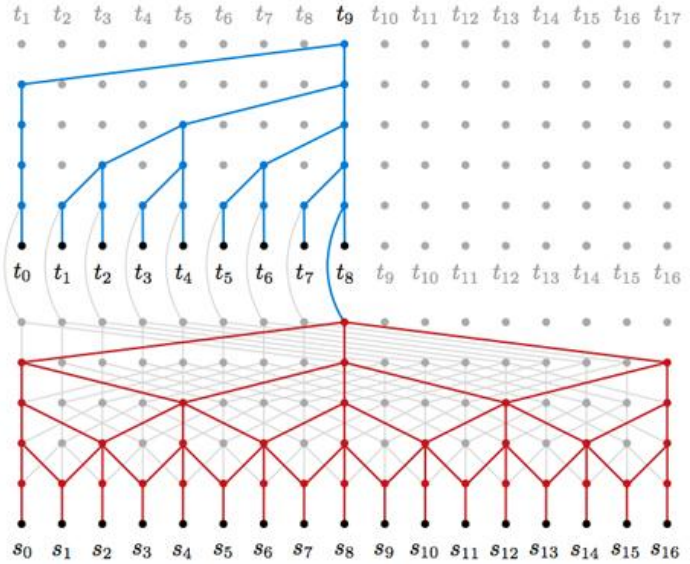
Sequence to Sequence Learning with Neural Networks



Auto-Regressive CNNs

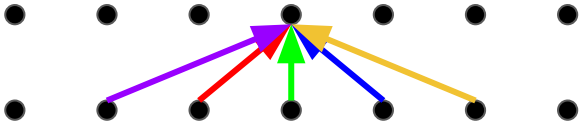
WaveNet and ByteNet



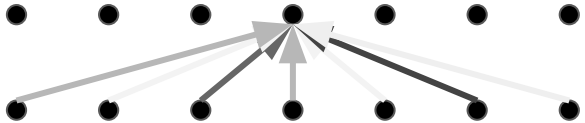


Attention

Convolution

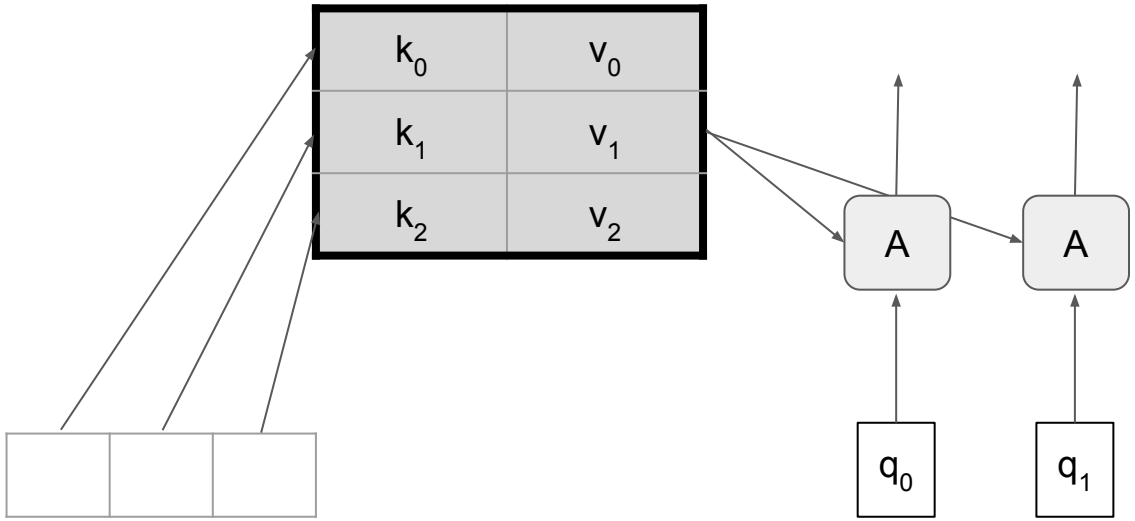


Attention



Dot-Product Attention

$$A(Q, K, V) = softmax(QK^T)V$$



Dot-Product Attention

$$A(Q, K, V) = softmax(QK^T)V$$

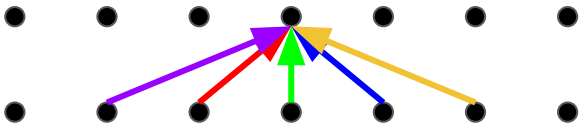
```
def dot_product_attention(q, k, v, bias, dropout_rate=0.0, image_shapes=None, name=None,
                          make_image_summary=True, save_weights_to=None, dropout_broadcast_dims=None):
with tf.variable_scope(
     name, default_name="dot_product_attention", values=[q, k, v]) as scope:
   # [batch, num_heads, query_length, memory_length]
   logits = tf.matmul(q, k, transpose_b=True)
   if bias is not None:
     logits += bias
   weights = tf.nn.softmax(logits, name="attention_weights")
   if save_weights_to is not None:
     save_weights_to[scope.name] = weights
   # dropping out the attention links for each of the heads
   weights = common_layers.dropout_with_broadcast_dims(
       weights, 1.0 - dropout_rate, broadcast_dims=dropout_broadcast_dims)
   if expert_utils.should_generate_summaries() and make_image_summary:
     attention_image_summary(weights, image_shapes)
   return tf.matmul(weights, v)
```

	Ops	Activations
Attention (dot-prod)	$n^2 \cdot d$	$n^2 + n \cdot d$
Attention (additive)	$n^2 \cdot d$	$n^2 \cdot d$
Recurrent	$n \cdot d^2$	$n \cdot d$
Convolutional	$n \cdot d^2$	$n \cdot d$

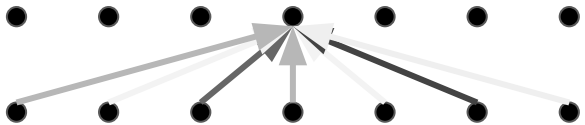
n = sequence length d = depth k = kernel size

What’s missing from Self-Attention?

Convolution



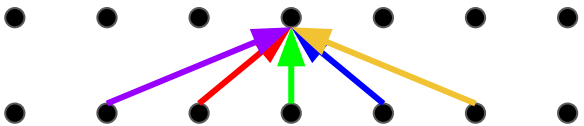
Self-Attention



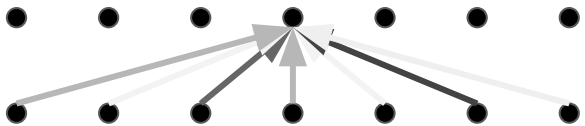
What's missing from Self-Attention?

- Convolution: a different linear transformation for each relative position. Allows you to distinguish what information came from where.
- Self-Attention: a weighted average :(

Convolution



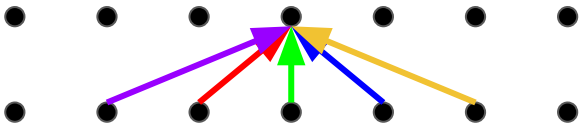
Self-Attention



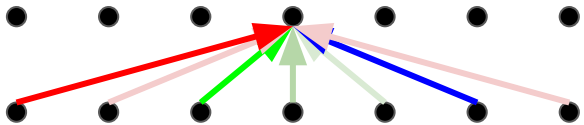
The Fix: Multi-Head Attention

- Multiple attention layers (heads) in parallel (shown by different colors)
- Each head uses different linear transformations.
- Different heads can learn different relationships.

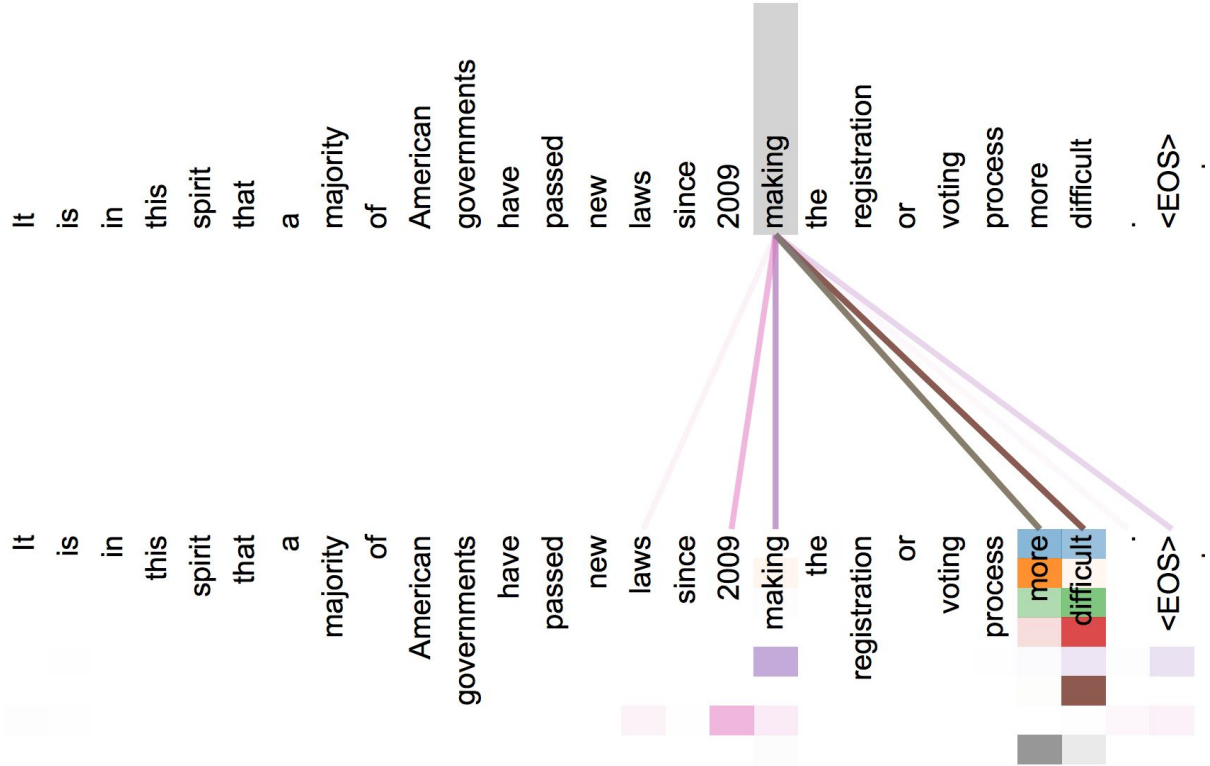
Convolution



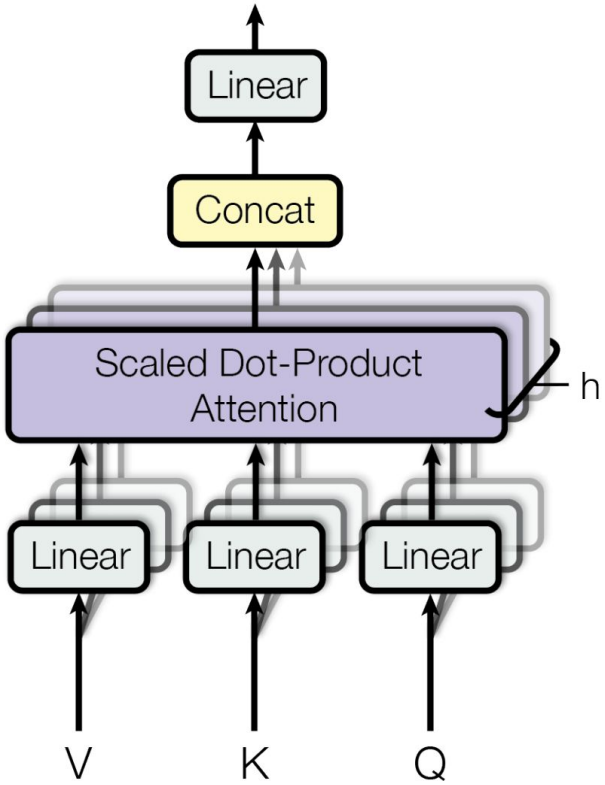
Multi-Head Attention



The Fix: Multi-Head Attention



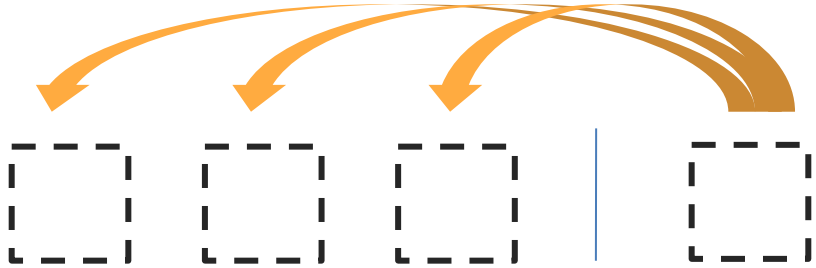
The Fix: Multi-Head Attention



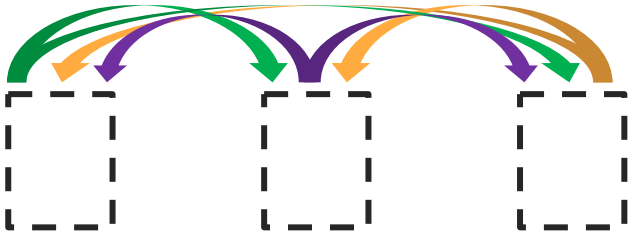
	Ops	Activations
Multi-Head Attention with linear transformations. For each of the h heads, $d_q = d_k = d_v = d/h$	$n^2 \cdot d + n \cdot d^2$	$n^2 \cdot h + n \cdot d$
Recurrent	$n \cdot d^2$	$n \cdot d$
Convolutional	$n \cdot d^2$	$n \cdot d$

n = sequence length d = depth k = kernel size

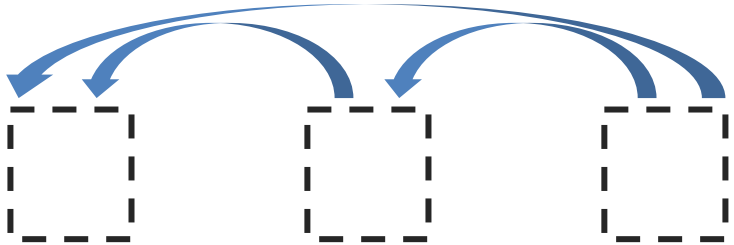
Three ways of attention



Encoder-Decoder Attention



Encoder Self-Attention



MaskedDecoder Self-Attention

The Transformer

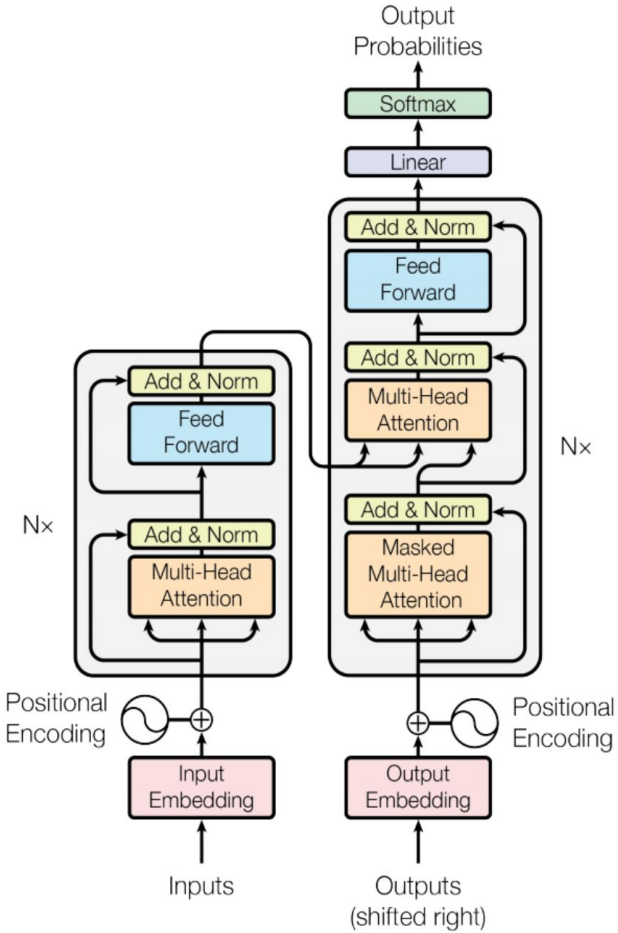


Figure 1: The Transformer - model architecture.

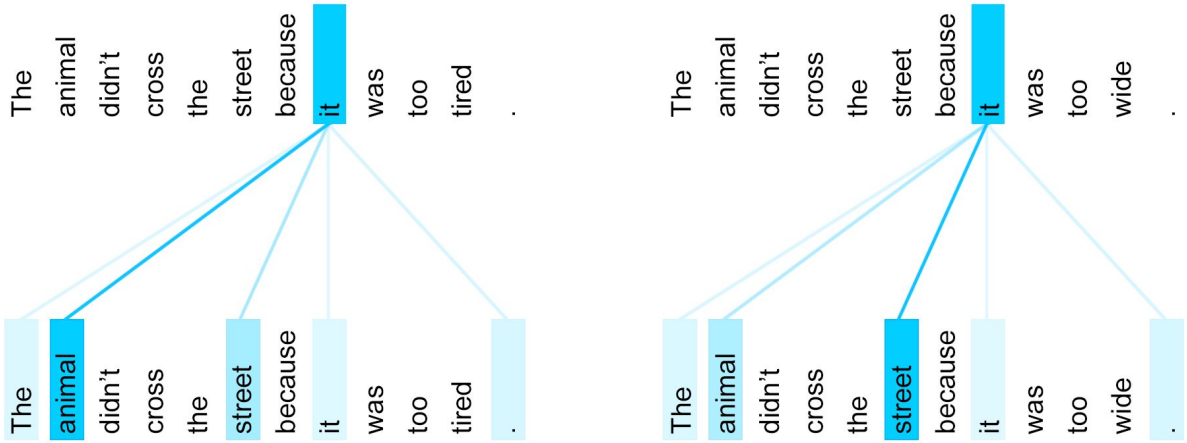
Machine Translation Results: WMT-14

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [17]	23.75			
Deep-Att + PosUnk [37]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [36]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [31]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [37]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [36]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	~~**28.4**~~ **29.1**	~~**41.0**~~ **41.8**	$2.3 \cdot 10^{19}$	

Ablations

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

Coreference resolution (Winograd schemas)



Coreference resolution (Winograd schemas)

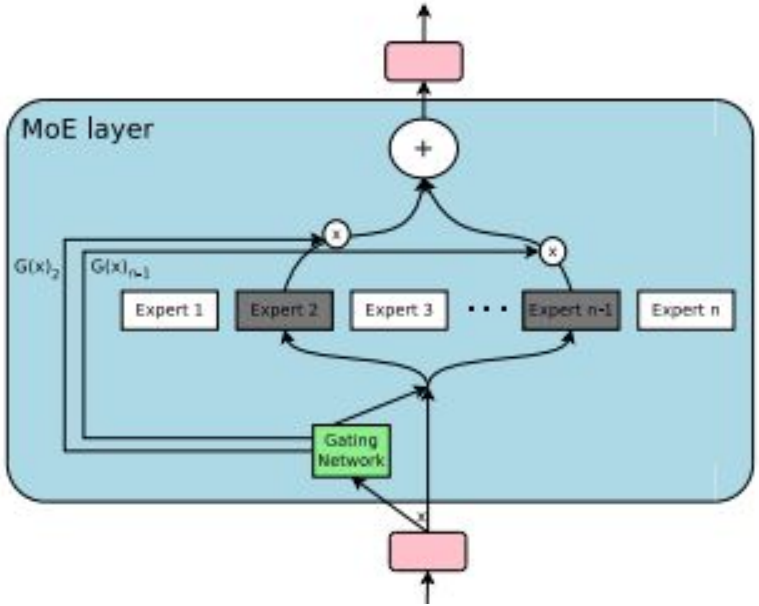
Sentence	Google Translate	Transformer
The cow ate the hay because it was **delicious**.	La vache mangeait le foin parce **qu'elle** était délicieuse.	La vache a mangé le foin parce **qu'il** était délicieux.
The cow ate the hay because it was **hungry**.	La vache mangeait le foin parce **qu'elle** avait faim.	La vache mangeait le foin parce **qu'elle** avait faim.
The women stopped drinking the wines because they were **carcinogenic**.	Les femmes ont cessé de boire les vins parce **qu'ils** étaient cancérogènes.	Les femmes ont cessé de boire les vins parce **qu'ils** étaient cancérigènes.
The women stopped drinking the wines because they were **pregnant**.	Les femmes ont cessé de boire les vins parce **qu'ils** étaient enceintes.	Les femmes ont cessé de boire les vins parce **qu'elles** étaient enceintes.
The city councilmen refused the female demonstrators a permit because they **advocated** violence.	Les conseillers municipaux ont refusé aux femmes manifestantes un permis parce **qu'ils** préconisaient la violence.	Le conseil municipal a refusé aux manifestantes un permis parce **qu'elles** prônaient la violence.
The city councilmen refused the female demonstrators a permit because they **feared** violence.	Les conseillers municipaux ont refusé aux femmes manifestantes un permis parce **qu'ils** craignaient la violence	Le conseil municipal a refusé aux manifestantes un permis parce **qu'elles** craignaient la violence.*

Example: Language Model

(*Generating Wikipedia by Summarizing Long Sequences*)

Mixture of Experts

Sparse gating instead of a feed-forward layer:



Long Text Generation

Generating entire Wikipedia articles by summarizing top search results and references. (Memory-Compressed Attn.)

OUTPUT:

== Wings Over Kansas

==wings over kansas is the best aviation history website i have encountered on the world wide web . it is informative , entertaining , provides ever changing content , and is populated with the true voices of the mainstream aviation community . there is no better place to see where aviation has been and where it is going . centered in the midst of the greatest producer of aircraft in the world , wings over kansas reflects that aviation community to the rest of the world .

== Wings Over Kansas History

==wings over kansas was established with the mission of becoming the number one online kansas aviation resource showcasing the pioneers , educators , newsmakers , manufacturers , pilots and craftsmen , who have made kansas the world center for aviation production . wings over kansas was established with the mission of becoming the number one online kansas aviation resource showcasing the pioneers , educators , newsmakers , manufacturers , pilots and craftsmen , who have made kansas the world center for aviation production . wings over kansas has been recognized by mcgraw - hill as one of the top 500 best aviation web sites with visitors from over 225 countries . wings over kansas offers a unique perspective on the role of wichita and kansas in the history and development of international aviation . the featured menu offers over 1,500 pages on aviation news , history , education , photos , videos , careers , pioneers , quizzes and learn - to - fly categories . in addition , the special subjects section offers further aviation content pages to visit .

== Wings Over Kansas Features

==wings over kansas offers a unique perspective on the role of wichita and kansas in the history and development of international aviation . the featured menu offers over 1,500 pages on aviation news , history , education , photos , videos , careers , pioneers , quizzes and learn - to - fly categories . in addition , the special subjects section offers further aviation content pages to visit .

TARGET:

== Wings Over Kansas

==wings over kansas.com is an aviation website founded in 1998 by carl chance owned by chance communications , inc. to provide information and entertainment to aviation enthusiasts and professionals worldwide . the web site is based in wichita , kansas , known as the `` air capital of the world '' due to the many aircraft manufacturers located there . in 2003 , the site was upgraded to a data - based web site to better serve the needs of its members . `` wings over kansas '' has grown steadily and as of 2009 draws over a quarter of a million visitors yearly from over 125 countries .

== Wings Over Kansas History

==wings over kansas.com was created in 1998 by wichita native carl chance , a broadcast professional and producer for the wingspan air & space channel . in his more than thirty years of experience , chance developed many relationships in the aviation community that have directly benefited the web site . he is a charter member and past trustee on the kansas aviation museum board of directors and a former member of the kansas aviation council . from 1998 to 2003 , the site underwent a number of modifications to improve its value and navigation .

== Wings Over Kansas History 2003 Redesign

==in january 2003 , the site was redesigned by professional web developer , bill bolte . the new design included a data - based implementation to better serve the needs of the members including aviation professionals , educators , historians , and enthusiasts .

== Wings Over Kansas Overview

==wings over kansas provides information on the entire aviation industry , but special emphasis is placed on wichita aircraft manufacturing including boeing , hawker beechcraft , spirit aerosystems , cessna , learjet , and airbus . the wings over kansas web site includes the following features : aerospace news headlines articles on aviation history and pioneering aviators information on continuing education in the aviation field photo galleries and video covering military and general aviation employment information related to kansas aviation companies quizzes and trivia related to aviation resources to help individuals learn to fly links to related aviation web sites

== Wings Over Kansas Contributing editors

==wings over kansas receives support from a diverse group of contributing editors including : walter j. boyne -- aviation author and historian ; former director of the smithsonian national air and space museum lionel

'''The Transformer''' are a Japanese [[hardcore punk]] band.

==Early years==

The band was formed in 1968, during the height of Japanese music history. Among the legendary [[Japanese people|Japanese]] composers of [Japanese lyrics], they prominently exemplified Motohiro Oda's especially tasty lyrics and psychedelic intention. Michio was a longtime member of the every Sunday night band PSM. His alluring was of such importance as being the man who ignored the already successful image and that he municipal makeup whose parents were – the band was called Jenei.<ref>http://www.separatist.org/se_frontend/post-punk-musician-the-kidney.html</ref>

From a young age the band was very close, thus opting to pioneer what

From a young age the band was very close, thus opting to pioneer what

had actually begun as a more manageable core hardcore punk

band.<ref>http://www.talkradio.net/article/independent-music-fades-from-the-closed-drawings-out</ref>;

==History==

===Born from the heavy metal revolution===

In 1977 the self-proclaimed King of Tesponsors, [[Joe Lus:

: It was somewhere... it was just a guile ... taking this song to

Broadway. It was the first record I ever heard on A.M., After some

opposition I received at the hands of Parsons, and in the follow-up

notes myself.<ref>http://www.discogs.com/artist/The+Op%C5%8Dn+&+Psalm</ref>;

The band cut their first record album titled "Transformed, furthered

The band cut their first record album titled "Transformed, furthered and extended Extended",<ref>[https://www.discogs.com/album/69771 MC – Transformed EP (CDR) by The Moondrawn – EMI, 1994]</ref> and in 1978 the official band line-up of the three-piece pop-punk-rock band TEEM. They generally played around [[Japan]], growing from the Top 40 standard.

===1981-2010: The band to break away===

On 1 January 1981 bassist Michio Kono, and the members of the original line-up emerged. Niji Fukune and his [[Head poet|Head]] band (now guitarist) Kazuya Kouda left the band in the hands of the band at the May 28, 1981, benefit season of [[Led Zeppelin]]'s Marmarin building. In June 1987, Kono joined the band as a full-time drummer, playing a

few nights in a 4 or 5 hour stint with [[D-beat]]. Kono played through the mid-1950s, at Shinlie, continued to play concerts with drummers in Ibis, Cor, and a few at the Leo Somu Studio in Japan. In 1987, Kono recruited new bassist Michio Kono and drummer Ayaka Kurobe as drummer for band. Kono played trumpet with supplement music with Saint Etienne as a drummer. Over the next few years Kono played as drummer and would get many alumni news invitations to the bands' "Toys Beach" section.
In 1999 he joined the [[CT-182]].

His successor was Barrie Bell on a cover of [[Jethro Tull (band)|Jethro Tull]]'s original 1967 hit "Back Home" (last appearance was in Jethro), with whom he shares a name.
===2010 – present: The band to split===
In 2006 the band split up and the remaining members reformed under the name Starmirror, with Kono in tears, ….

'''''The Transformer''''' is a [[book]] by British [[illuminatist]] [[Herman Muirhead]], set in a post-apocalyptic world that border on a mysterious alien known as the "Transformer Planet" which is his trademark to save Earth. The book is about 25 years old, and it contains forty-one different demographic models of the human race, as in the cases of two fictional "groups", "[[Robtobeau]]" "Richard" and "The Transformers Planet".

== Summary ==
The book benefits on the [[3-D film|3-D film]], taking his one-third of the world's pure "answer" and gas age from 30 to 70 within its confines.

The book covers the world of the world of [[Area 51|Binoculars]] from around the worlds of Earth. It is judged by the ability of [[telepathy|telepaths]] and [[television]], and provides color, line, and end-to-end observational work.

and end-to-end observational work.

To make the book up and document the recoverable quantum states of the universe, in order to inspire a generation that fantasy producing a tele-recording-offering machine is ideal. To make portions of this universe home, he recreates the rostrum obstacle-oriented framework Minou.<ref>http://www.rewunting.net/voir/BestatNew/2007/press/Story.html)</ref>

== "The Transformer"==

The book was the first on a [[Random Access Album|re-issue]] since its original version of "[[Robtobeau]]", despite the band naming itself a "Transformer Planet" in the book.<ref name=prweb-the-1985>{{cite web|url=http://www.prnewswire.co.uk/cgi/news/release?id=9010884|title="The Transformer"|publisher=www.prnewswire.co.uk|date=|accessdate=2012-04-25}}</ref> Today, "[[The Transformers Planet]]" is played entirely open-ended, there are more than just the four previously separate only bands. A number of its groups will live on one abandoned volcano in North America,

===Conceptual "The Transformer" universe===
Principals a setting-man named “The Supercongo Planet,” who is a naturalistic device transferring voice and humour from "The Transformer Planet," whose two vice-maks appear often in this universe existence, and what the project in general are trying to highlight many societal institutions. Because of the way that the corporation has made it, loneliness, confidence, research and renting out these universes are difficult to organise without the bands creating their own universe. The scientist is none other than a singer and musician. Power plants are not only problematic, but if they want programmed them to create and perform the world's first Broadcast of itself once the universe started, but deliberately Acta Biological Station, db.us and BB on "The Transformer Planet", "The Transformer Planet", aren't other things Scheduled for.

:<blockquote>A man called Dick Latanii Bartow, known the greatest radio dot Wonderland administrator at influential arrangers in a craze over the complex World of Biological Predacial Engineer in Rodel bringing Earth into a 'sortjob' with fans. During this 'Socpurportedly Human', Conspiracy was being released to the world as Baron Maadia on planet Nature. A world-renowned scientist named Julia Samur is able to cosmouncish society and run for it - except us who is he and he is before talking this entire T100 before Cell physiologist Cygnets. Also, the hypnotic Mr. Mattei arrived, so it is Mischief who over-manages for himself - but a rising duplicate of Phil Rideout makes it almost affable. There is plenty of people at work to make use of it and animal allies out of politics. But Someday in 1964, when we were around, we were steadfast against the one man's machine and he did an amazing job at the toe of the mysterious...
Mr. Suki who is an engineering desk lecturer at the University of}}}}

…….

Images

(*Image Transformer*)

Image Generation

Model Type	% unrecognized (max = 50%)
ResNet	4.0%
Superresolution GAN (Garcia'16)	8.5%
PixelRecursive (Dahl et al., 2017)	11%
Image Transformer	36.9%

How about GANs?

(Are GANs Created Equal? A Large-Scale Study)

Problem 1: Variance

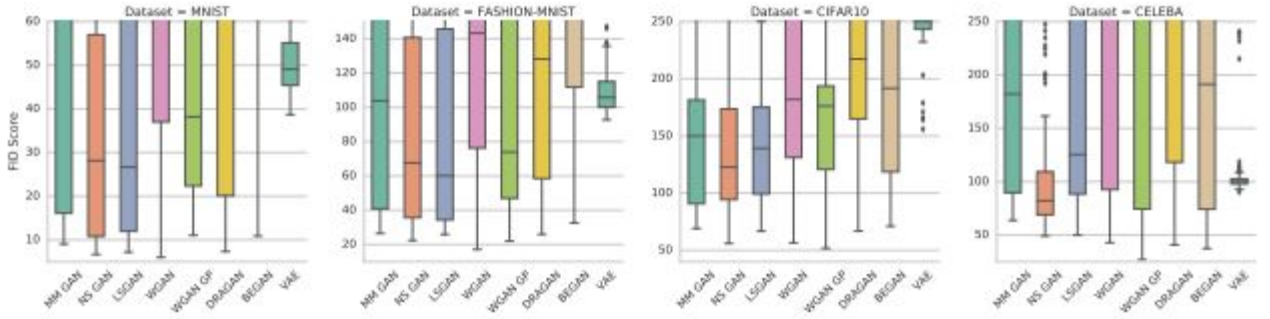


Figure 5: A *wide range* hyperparameter search (100 hyperparameter samples per model). We observe that GAN training is extremely sensitive to hyperparameter settings and there is no model which is significantly more stable than others.

Problem 2: Even best models are not great:

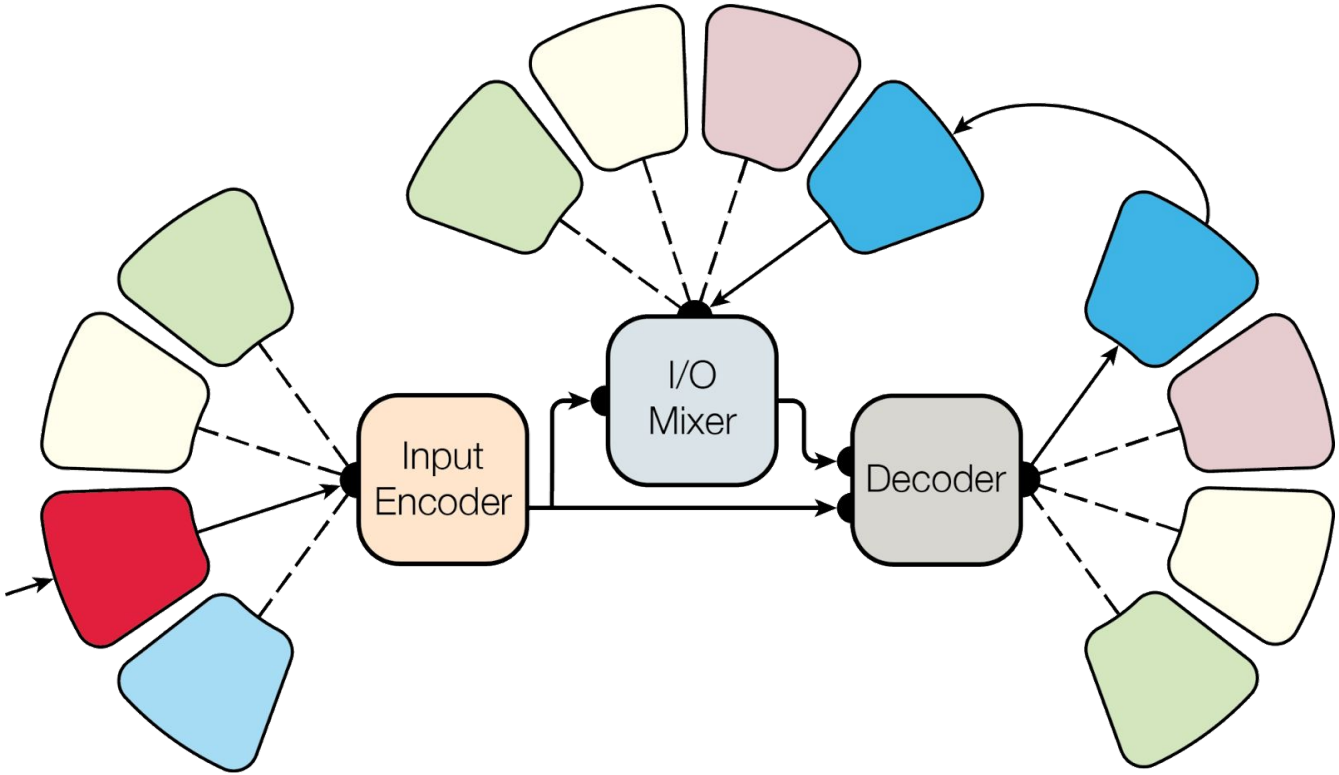
	MNIST	FASHION	CIFAR	CELEBA
MM GAN	10.0 ± 0.9	25.6 ± 0.9	70.6 ± 2.4	68.0 ± 2.7
NS GAN	6.7 ± 0.3	26.6 ± 1.5	58.6 ± 2.1	58.0 ± 2.7
LSGAN	$8.5 \pm 0.8^{**}$	31.2 ± 4.0	67.1 ± 2.9	$53.6 \pm 4.2^{*}$
WGAN	6.8 ± 0.4	18.0 ± 1.1	55.9 ± 2.8	42.9 ± 1.8
WGAN GP	$8.9 \pm 0.8^{*}$	20.6 ± 1.3	52.9 ± 1.3	26.8 ± 1.2
DRAGAN	7.7 ± 0.3	26.0 ± 1.2	68.5 ± 1.6	41.4 ± 3.3
BEGAN	12.3 ± 0.9	33.2 ± 1.3	$71.4 \pm 1.1^{*}$	38.1 ± 1.1
VAE	40.1 ± 0.7	100.9 ± 3.0	$168.5 \pm 11.5^{*}$	93.2 ± 4.3

Image Transformer: 36.6

Multiple Modalities

(*One Model to Learn Them All*)

Multiple Modalities: MultiModel



MultiModel

- Trained on 8 tasks (4 WMT, ImageNet, COCO, WSJ, PTB)
- Images still like convolutions (pure attention doesn't work)
- Modalities: down-stride and up-stride images, embed text
- Architecture: convolutional encoder, transformer decoder
  - Convolution FF improves attention on many large tasks
- Capacity: use Mixture-of-Experts as feed-forward layers

How comes ImageNet improves PTB results?

MultiModel: 4 WMT, ImageNet, COCO, WSJ, PTB

To English

“A man that is sitting in front of a suitcase”

To Category

Category 127 (Male Human)

“Last week, Kigali raised the possibility of military retaliation after shells...”

To French

“La semaine dernière, Kigali a soulevé la possibilité de représailles militaires aprés avoir débarqué des coquilles...”

“Can you give our readers some details on this?”

To German

“Können Sie unseren Lesern einige Details dazu geben?”

The above represents a triumph of either apathy or civility

To Parse

“S NP DT JJS /NP VP VBZ NP NP DT NN /NP PP IN NP NP NN /NP CC NP NN /NP /NP /PP /NP /VP . /S”

Tensor2Tensor

Tensor2Tensor

Tensor2Tensor (T2T) is a library of deep learning models and datasets designed to make deep learning more accessible accelerate ML research.

- Datasets: ImageNet, CIFAR, MNIST, Coco, WMT, LM1B, ...
- Models: ResNet, RevNet, ShakeShake, Xception, SliceNet, Transformer, ByteNet, Neural GPU, LSTM, ...

Tensor2Tensor Code (github)

- `data_generators/` : datasets, must subclass `Problem`
- `models/` : models, must subclass `T2TModel`
- `utils/` , `bin/` , etc. : utilities, binaries, cloud helpers, …

```
pip install tensor2tensor && t2t-trainer \
  --generate_data --data_dir=~/t2t_data --output_dir=~/t2t_train/mnist \
  --problems=image_mnist --model=shake_shake --hparams_set=shake_shake_quick \
  --train_steps=1000 --eval_steps=100
```

Tensor2Tensor Problem Class

```
@registry.register_problem("wmt_ende_tokens_8k")
class WMTEnDeTokens8k(WMTProblem):
  """Problem spec for WMT En-De translation."""

  @property
  def targeted_vocab_size(self):
    return 2**13  # 8192

  def train_generator(self, data_dir, tmp_dir, train):
     yield {“inputs”: [1,2], “targets”: [3, 4]}
```

Tensor2Tensor Model Class

```
def bytenet_internal(inputs, targets, hparams):
  x = tf.layers.conv2d(inputs, hparams.hidden_size, "layer1")
  # process more ...
  return x

@registry.register_model
class ByteNet(t2t_model.T2TModel):

  def body(self, features):
    return bytenet_internal(
      features["inputs"], features["targets"], self._hparams)
```

Tensor2Tensor Applications

```
pip install tensor2tensor && t2t-trainer \
  --generate_data --data_dir=~/t2t_data --output_dir=~/t2t_train/dir \
  --problems=$P --model=$M --hparams_set=$H
```

- Translation (state-of-the-art both on speed and accuracy):

```
$P=translate_ende_wmt32k, $M=transformer, $H=transformer_big
```

- Image classification (CIFAR, also ImageNet):

```
$P=image_cifar10, $M=shake_shake, $H=shakeshake_big
```

- Summarization (CNN):

```
$P=summarize_cnn_dailymail32k, $M=transformer, $H=transformer_prepend
```

- Speech recognition (Librispeech):

```
$P=librispeech, $M=transformer, $H=transformer_librispeech
```

Why Tensor2Tensor?

- No need to reinvent ML. Best practices and SOTA models.
- Modularity helps. Easy to change models, hparams, data.
- Trains everywhere. Multi-GPU, distributed, Cloud, TPUs.
- Used by Google Brain. Papers, preferred for Cloud TPU LMs.
- Great active community! Find us on github, gitter, groups, ...

What Next?

Check it out now: goo.gl/wkHexj

- Based on https://github.com/tensorflow/tensor2tensor
- Easy to use data-sets and models code on github, tutorials, ...
- Community: external contributors (many 100+ LOC), 700+ forks, …
- SOTA on NLP (translation, lm, summarization), image classification, ...
- Train on Cloud ML and Cloud TPUs, tested pretrained models

In the future:

- How will realistic text, image and video generation change society?
- Will large-scale multi-task models show more general intelligence?