DOTE 6635: Artificial Intelligence for Business Research

Transformers

Renyu (Philip) Zhang

1

Agenda

- Sequence-to-Sequence (Seq2seq) and Neural Machine Translation (NMT)
- Attention is All You Need
- Extensions, Adaptations, and Substitutes of Transformers

2

2

NLP Roadmap

Language models are few-shot learners
T Brown, B Mann, N Ryder… - Advances in neural …, 2020 - proceedings.neurips.cc
… up **language models** greatly improves task-agnostic, **few-shot** … GPT-3, an autoregressive **language model** with 175 billion … **language model**, and test its performance in the **few-shot** …
☆ Save 99 Cite Cited by 21815 Related articles All 27 versions »

Attention is all you need
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… the number of **attention** heads and the **attention** key and value dimensions, keeping the amount of computation constant, as described in Section 3.2.2. While single-head **attention** is 0.9 …
★ Save 99 Cite Cited by 36430 Related articles All 35 versions »

Bert: Pre-training of deep bidirectional transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv …, 2018 - arxiv.org
… We introduce **BERT** and its detailed implementation in this … For finetuning, the **BERT** model is first initialized with the pre-trained … , we present **BERT** fine-tuning results on 11 **NLP** tasks. …
☆ Save 99 Cite Cited by 34042 Related articles All 41 versions »

Probability & Vector Representations of NLP → Word2vec, RNN & Seq2seq → Transformer-based Models, BERT, GPT → Large Language Models (LLM)

Before 2010 | 2013-2017 | 2018-2022 | 2023-now

Efficient estimation of word representations in vector space
T Mikolov, K Chen, G Corrado, J Dean - arXiv preprint arXiv:1301.3781, 2013 - arxiv.org
… **vector** X = **vector**("biggest")−**vector**("big") + **vector**("small"). Then, we search in the **vector space** for the **word** … question (we discard the input question **words** during this search). When the …
☆ Save 99 Cite Cited by 27227 Related articles All 48 versions »

[PDF] **Glove: Global vectors** for word representation
J Pennington, R Socher … - Proceedings of the 2014 …, 2014 - aclanthology.org
… We use our insights to construct a new model for word representation which we call **GloVe**, for **Global Vectors**, because the **global** corpus statistics are captured directly by the model. …
☆ Save 99 Cite Cited by 26265 Related articles All 34 versions »

[PDF] Language models are unsupervised multitask learners
A Radford, J Wu, R Child, D Luan, D Amodei … - OpenAI blog, 2019 - persagen.com
Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on taskspecific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset-matching or exceeding
☆ Save 99 Cite Cited by 2834 Related articles All 16 versions »

Training language models to follow instructions with human feedback
L Ouyang, J Wu, X Jiang, D Almeida… - Advances in …, 2022 - proceedings.neurips.cc
Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through a language model API, we collect a dataset of …
☆ Save 99 Cite Cited by 4049 Related articles All 10 versions »

3

3

NOVEMBER 27 2024

Announcing the NeurIPS 2024 Test of Time Paper Awards

COMMUNICATIONS CHAIRS 2024 / 2021 Conference

By Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, Cheng Zhang

We are honored to announce the Test of Time Paper Awards for NeurIPS 2024. This award is intended to recognize papers published 10 years ago at NeurIPS 2014 that have significantly shaped the research field since then, standing the test of time.

This year, we are making an exception to award two Test of Time papers given the undeniable influence of these two papers on the entire field. The awarded papers are:

- **Generative Adversarial Nets**
  Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio
- **Sequence to Sequence Learning with Neural Networks**
  Ilya Sutskever, Oriol Vinyals, Quoc V. Le

Seq2Seq paper has passed the test of time!

Sequence to Sequence Learning with Neural Networks has been cited more than 27,000 times as of this blog post. With the current fast advances of large language models and foundation models in general, making a paradigm shift in AI and applications, the field has benefited from the foundation laid by this work. It is the cornerstone work that set the encoder-decoder architecture, inspiring later attention-based improvements leading to today's foundation model research.

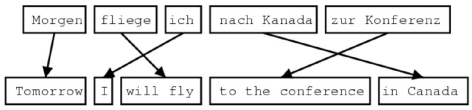
4

4

Neural Machine Translation (NMT)

Stanford CS224N: https://web.stanford.edu/class/cs224n/

- NMT is a way to do machine translation with a single end-to-end neural network: Sequence-to-sequence (seq2seq), which involves 2 RNNs.
- Machine translation is highly nontrivial and once was a huge research field in CS and NLP.



1519年600名西班牙人在墨西哥登陆，去征服几百万人口的阿兹特克帝国，初次交锋他们损兵三分之二。

In 1519, six hundred Spaniards landed in Mexico to conquer the Aztec Empire with a population of a few million. They lost two thirds of their soldiers in the first clash.

translate.google.com (2009): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of soldiers against their loss.
translate.google.com (2013): 1519 600 Spaniards landed in Mexico to conquer the Aztec empire, hundreds of millions of people, the initial confrontation loss of soldiers two-thirds.
translate.google.com (2015): 1519 600 Spaniards landed in Mexico, millions of people to conquer the Aztec empire, the first two-thirds of the loss of soldiers they clash.

5

5

Seq2Seq for NMT

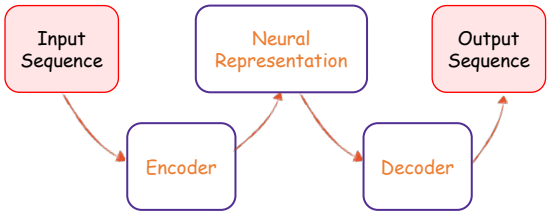
Stanford CS224N: https://web.stanford.edu/class/cs224n/

- Seq2seq is a Conditional Language Model:
  - Predicting the next word of the target sentence y conditioned on the source sentence x and prior texts.

$$P(y|x) = P(y_1|x)\,P(y_2|y_1,x)\,P(y_3|y_1,y_2,x)\ldots\underbrace{P(y_T|y_1,\ldots,y_{T-1},x)}$$

Probability of next target word, given target words so far and source sentence *x*

- Encoder-decoder architecture: Encoder takes input and produces a neural representation; Decoder produces output based on that neural representation.
  - Seq2seq: both input and output are sequences.
  - Summarization: Long text → short text
  - Dialogue: previous utterances → next utterance
  - Parsing: Input text → output parse as a sequence
  - Code generation: Natural language → Python code

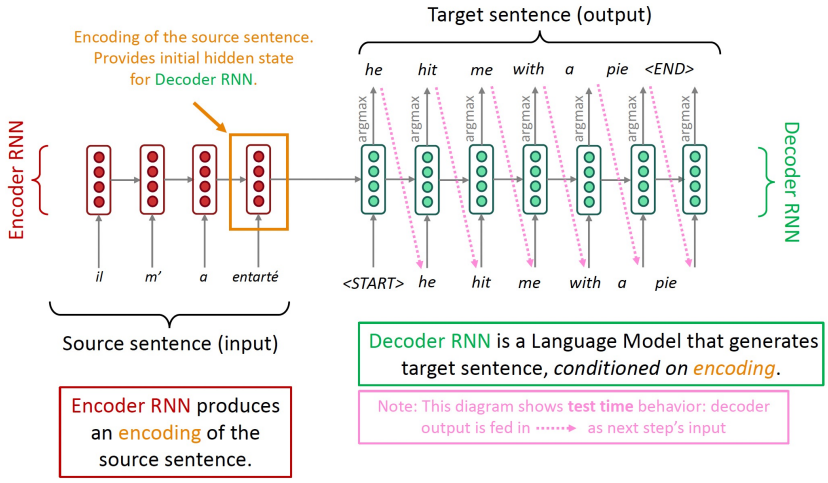


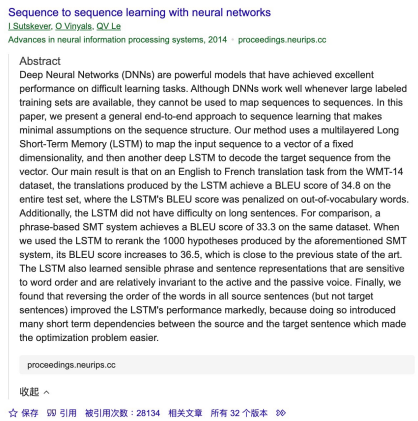
6

6

Seq2Seq Architecture

Stanford CS224N: https://web.stanford.edu/class/cs224n/
Scribed notes of last year: https://github.com/rphilipzhang/AI-PhD-S24/blob/main/Notes/Scribed_Notes-AI-PhD-S24.pdf




Sequence to sequence learning with neural networks
I Sutskever, O Vinyals, QV Le
Advances in neural information processing systems, 2014 · proceedings.neurips.cc

Abstract
Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous state of the art. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier.

proceedings.neurips.cc

收起 ^

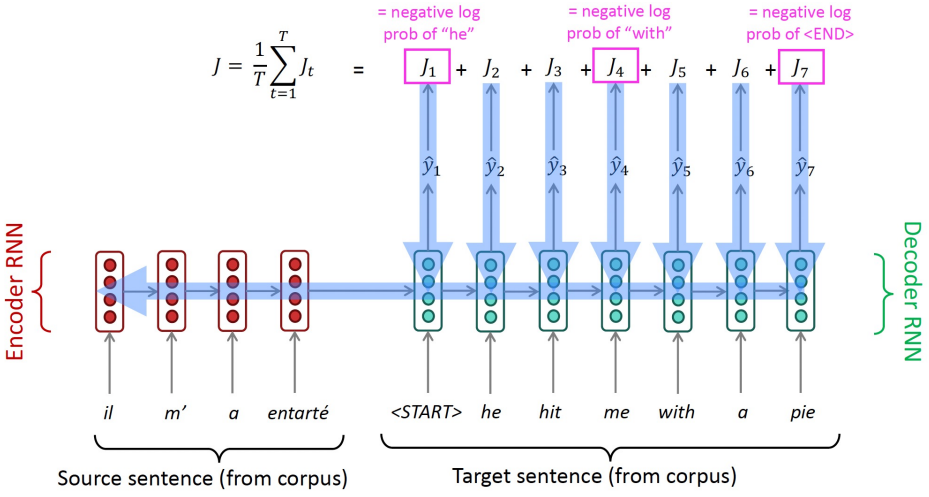
☆ 保存 99 引用 被引用次数：28134 相关文章 所有 32 个版本 »

7

7

Seq2Seq Training

Stanford CS224N: https://web.stanford.edu/class/cs224n/
Scribed notes of last year: https://github.com/rphilipzhang/AI-PhD-S24/blob/main/Notes/Scribed_Notes-AI-PhD-S24.pdf



Seq2seq is optimized as a **single system.** Backpropagation operates "*end-to-end*".

8

8

NMT: The First Major Success of DL-NLP

Stanford CS224N: https://web.stanford.edu/class/cs224n/

- NMT transformed from a mere research attempt in 2014 to the leading standard in 2016.
- 2014: The Seq2seq paper.
- 2016: Adopted by Google Translate.
- 2018: Adopted by everyone.

Microsoft SYSTRAN Google facebook
Baidu 百度 網易 NetEase www.163.com Tencent 腾讯 搜狗搜索

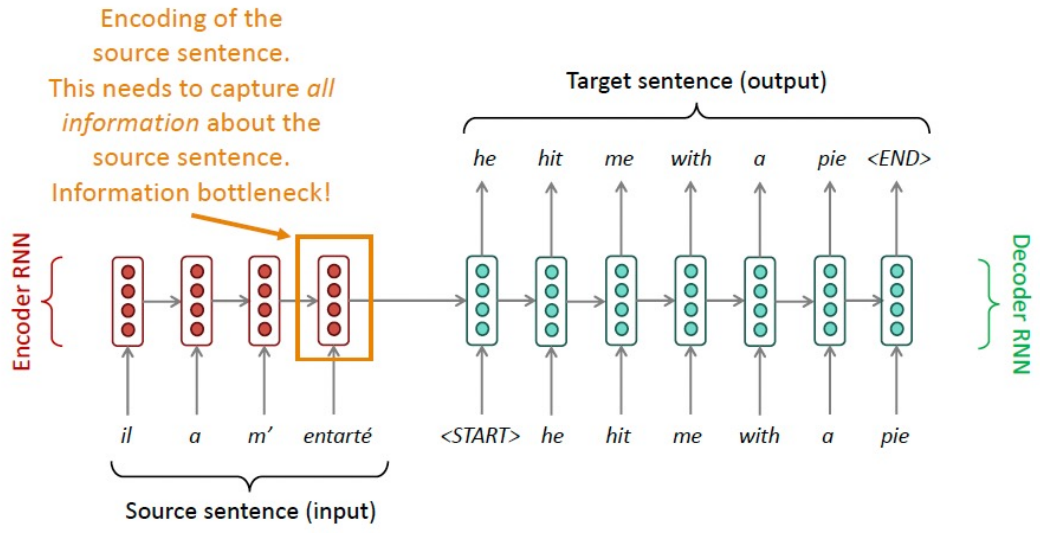
- The original statistical machine translation (SMT) system, built by hundreds of engineers over many years, soon outperformed by NMT trained by small groups of engineers in a few months.

9

9

Information Bottleneck in RNN

Stanford CS224N: https://web.stanford.edu/class/cs224n/



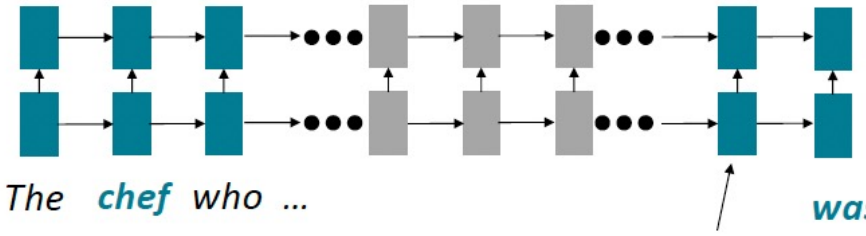
10

10

Issue with RNN: Linear Interaction Distance

Stanford CS224N: https://web.stanford.edu/class/cs224n/

- Human languages are intrinsically NOT linearly ordered.



Info of ***chef*** has gone through O(sequence length) many layers!

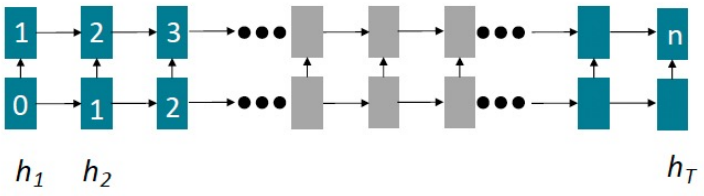
11

11

Issue with RNN: Non-parallelizability

Stanford CS224N: https://web.stanford.edu/class/cs224n/

- Forward and backward passes both have O(sequence length) unparallelizable operations.
- GPUs can perform independent small computations quickly in a large scale.
- Future hidden states cannot be computed (in full) before past RNN hidden states have been computed.
- Cannot scale with a very large dataset.



Numbers indicate min # of steps before a state can be computed

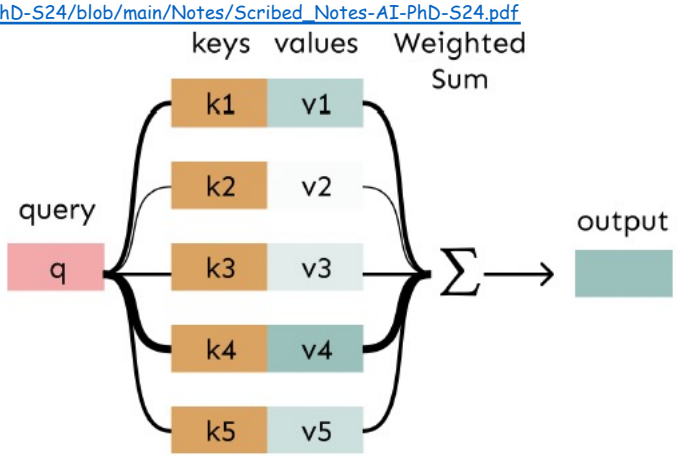
12

12

Attention as a Very General DL Technique

Scribed notes of last year: https://github.com/rphilipzhang/AI-PhD-S24/blob/main/Notes/Scribed_Notes-AI-PhD-S24.pdf
Stanford CS224N: https://web.stanford.edu/class/cs224n/

- **Attention**: Given a set of vector values and a vector of query, attention is a technique to compute a weighted sum of the values dependent on the query.
  - The weighted sum is a selective summary of the information contained in the values, where the query determines which values to focus on.
  - A fixed-size representation of an arbitrary set of representations (values), dependent on some other representation (query).



- In seq2seq + attention, each decoder hidden state (query) attends to all the encoder hidden states (values)..

Neural machine translation by **jointly learning** to **align** and **translate**
D Bahdanau, K Cho, Y Bengio - arXiv preprint arXiv:1409.0473, 2014 - arxiv.org
… 3 **LEARNING** TO **ALIGN** AND **TRANSLATE** In this section, we propose a novel architecture for **neural machine translation**. The new architecture consists of a bidirectional RNN as an …
☆ Save 99 Cite Cited by 37547 Related articles All 25 versions »

13

13

A Family of Attention Models

Stanford CS224N: https://web.stanford.edu/class/cs224n/

Name	Alignment score function	Citation
Content-base attention	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$	Graves2014
Additive(*)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$	Bahdanau2015
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ Note: This simplifies the softmax alignment to only depend on the target position.	Luong2015
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.	Luong2015
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$	Luong2015
Scaled Dot-Product(^)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	Vaswani2017

14

14

Agenda

- Sequence-to-Sequence (Seq2seq) and Neural Machine Translation (NMT)
- Attention is All You Need
- Extensions, Adaptations, and Substitutes of Transformers

15

15

Attention is All You Need

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

- Transformer: No RNN architecture, just attention mechanism.
- Self-attention: To generate y_t, we need to pay attention to $y_{<t}$.

$$w'_{ij} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{k}}$$

Query $\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i$ Key $\mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i$ Value $\mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i$

$$w'_{ij} = \mathbf{q}_i^\top \mathbf{k}_j$$

$$w_{ij} = \text{softmax}(w'_{ij})$$

$$\mathbf{y}_i = \sum_j w_{ij} \mathbf{v}_j \,.$$

Why does it work?

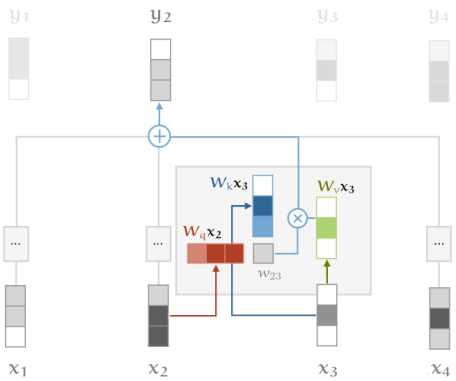


Illustration of the self-attention with key, query and value transformations.

Attention is all you need
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent …
… **We** implement this inside of scaled dot-product **attention** by masking out (setting to $-\infty$) …
☆ Save 99 Cite Cited by 150580 Related articles All 91 versions »

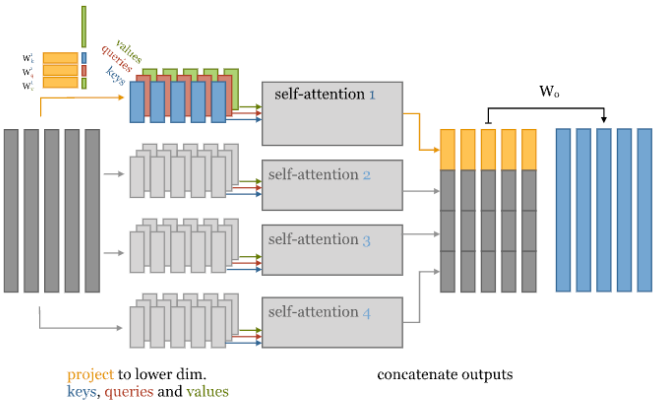
16

16

Multi-head Attention

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

- Multi-head attention is a way to speed up the training procedure.
- Instead of using a large matrix to compute all attentions, we can compute multiple attention matrices and concatenate the final vectors.
- Allows for parallel computing: Deploy attention mechanisms to multiple computing cores in parallel and sum them up at the end.
- Input dim = 256, 8 attention heads, each with 32 dimensions.



The basic idea of multi-head self-attention with 4 heads. To get our keys, queries and values, we project the input down to vector sequences of smaller dimension.

17

17

Position Encoding

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

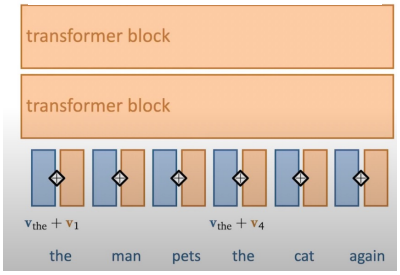
word embeddings:

$\mathbf{v}_{\text{the}}, \mathbf{v}_{\text{man}}, \mathbf{v}_{\text{pets}}, \mathbf{v}_{\text{cat}}, \mathbf{v}_{\text{again}}$

- Position embeddings: Position vectors which are learned.
- Position encoding: The function from position to vector.
- The final input of the model is the sum of word embeddings and position embeddings.

position embeddings:

$\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4, \mathbf{v}_5, \ldots$



18

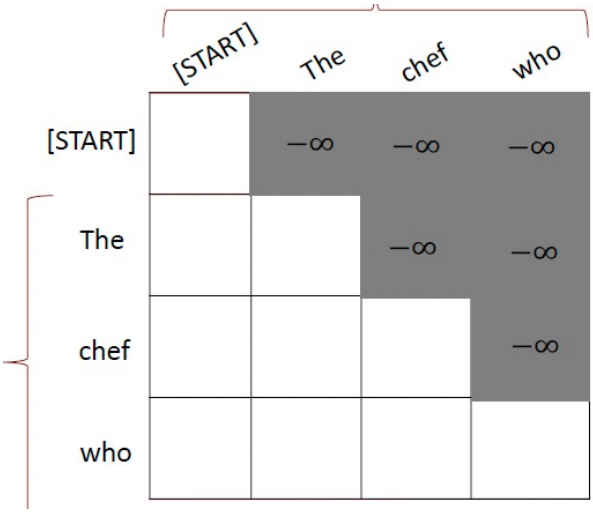
18

Auto-Regression

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

- Self-supervised learning for transformers.
- To use self-attention in decoders, we need to mask the future.
- Inefficient implementation: Change the set of keys and queries to include only past words.
- Parallelizable implementation: Mask out attention to future words by setting the weight to -inf.

$$w'_{ij} = \begin{cases} q_i^T k_j, j \le i \\ -\infty, j > i \end{cases}$$



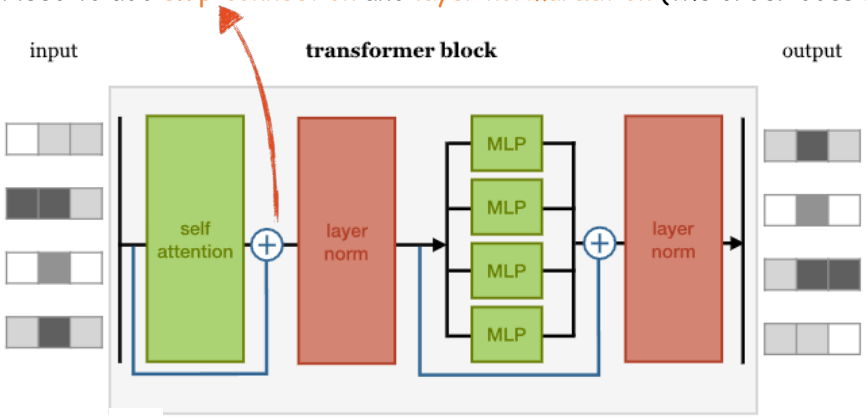
19

19

Transformer

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

- Transformer = Multi-head self-attention + MLP + position encoding + autoregression
- Need to add skip-connection and layer normalization (the order does not matter).



20

20

Layer Normalization

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

- Layer normalization: A trick to help models train faster.
- Cut down on uninformative variation in hidden values by normalizing to unit mean and standard deviation within each layer: Normalized gradients.

Layer normalization
JL Ba, JR Kiros, GE Hinton - arXiv preprint arXiv:1607.06450, 2016 - arxiv.org
… , we transpose batch **normalization** into **layer normalization** by computing the mean and variance used for **normalization** from all of the summed inputs to the neurons in a **layer** on a …
☆ Save 99 Cite Cited by 10350 Related articles All 6 versions »

- Let $x \in \mathbb{R}^d$ be an individual (word) vector in the model.
- Let $\mu = \sum_{j=1}^{d} x_j$; this is the mean; $\mu \in \mathbb{R}$.
- Let $\sigma = \sqrt{\frac{1}{d}\sum_{j=1}^{d}(x_j - \mu)^2}$; this is the standard deviation; $\sigma \in \mathbb{R}$.
- Let $\gamma \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$ be learned "gain" and "bias" parameters. (Can omit!)
- Then layer normalization computes:

$$\text{output} = \frac{x - \mu}{\sqrt{\sigma + \epsilon}} * \gamma + \beta$$

Normalize by scalar mean and variance

Modulate by learned elementwise gain and bias

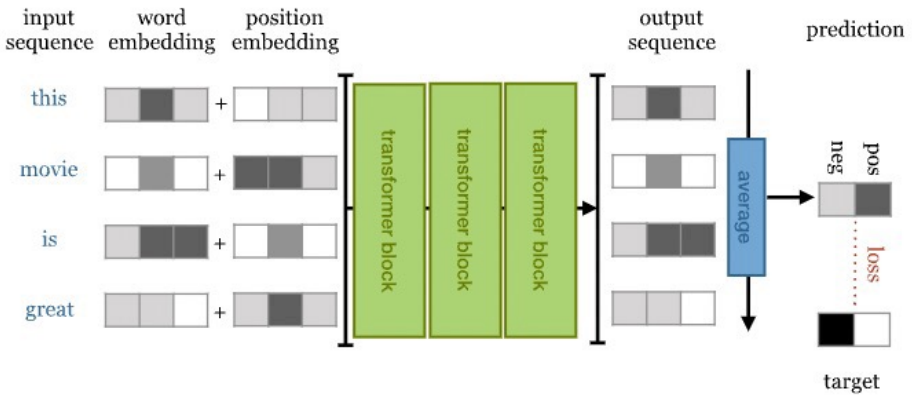
21

21

Classification Transformer

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

- Directly train a classifier on top of a transformer.



22

22

Attention is All You Need

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

- Input: Sequence in language one and Sequence in language two.
- Architecture: Encoder + Decoder
- 8 heads, 512 embedding dimensions, 2048 sentence length
- Trained on 8 GPUs for 5 days.

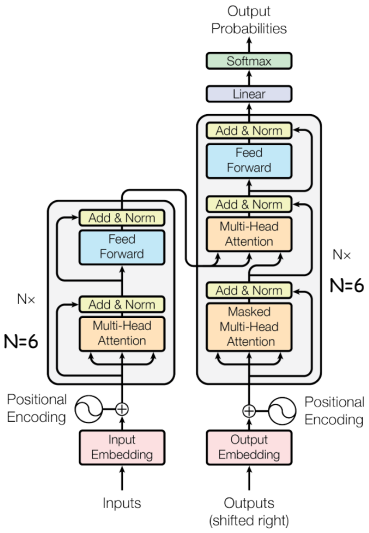


Figure 1: The Transformer - model architecture.

23

23

Attention is All You Need

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

Machine Translation

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	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 [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	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**	**41.8**	$2.3 \cdot 10^{19}$	

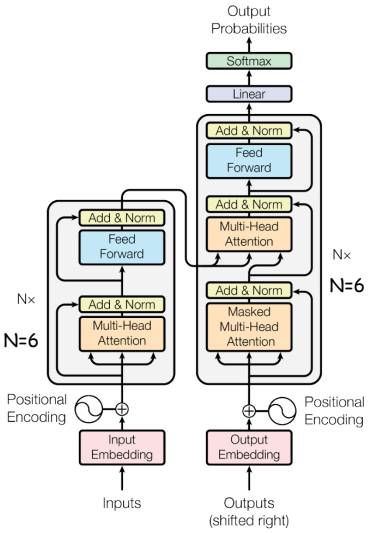


Figure 1: The Transformer - model architecture.

24

24

Attention is All You Need

Stanford CS224N: https://web.stanford.edu/class/cs224n/
https://peterbloem.nl/blog/transformers

Document Generation

Model	Test perplexity	ROUGE-L
seq2seq-attention, L = 500	5.04952	12.7
Transformer-ED, L = 500	2.46645	34.2
Transformer-D, L = 4000	2.22216	33.6
Transformer-DMCA, no MoE-layer, L = 11000	2.05159	36.2
Transformer-DMCA, MoE-128, L = 11000	1.92871	37.9
Transformer-DMCA, MoE-256, L = 7500	1.90325	38.8

The old standard

Transformers all the way down.

The parallelizability of transformer enables large-scale pre-training!

Figure 1: The Transformer - model architecture.

25

25

Application of Language Model: Detecting FTD

Psychiatry Research 304 (2021) 114135

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Detecting formal thought disorder by deep contextualized word representations

Justyna Sarzynska-Wawer [*,a], Aleksander Wawer [1,b], Aleksandra Pawlak [2,c], Julia Szymanowska [2,c], Izabela Stefaniak [d], Michal Jarkiewicz [3,d], Lukasz Okruszek [a]

[a] Institute of Psychology, Polish Academy of Sciences, Jaracza 1, 00-378 Warszawa, Poland
[b] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warszawa, Poland
[c] University of Social Sciences and Humanities, Chodakowska 19/31, 03-815 Warszawa, Poland
[d] Institute of Psychiatry and Neurology, Sobieskiego 9, 02-957 Warszawa, Poland

ARTICLE INFO

Keywords:
Schizophrenia
Language
Natural language processing
Deep learning

ABSTRACT

Computational linguistics has enabled the introduction of objective tools that measure some of the symptoms of schizophrenia, including the coherence of speech associated with formal thought disorder (FTD). Our goal was to investigate whether neural network based utterance embeddings are more accurate in detecting FTD than models based on individual indicators. The present research used a comprehensive Embeddings from Language Models (ELMo) approach to represent interviews with patients suffering from schizophrenia (N=35) and with healthy people (N=35). We compared its results to the approach described by Bedi et al. (2015), referred to here as the coherence model. Evaluations were also performed by a clinician using the Scale for the Assessment of Thought, Language and Communication (TLC). Using all six TLC questions the ELMo obtained an accuracy of 80% in distinguishing patients from healthy people. Previously used coherence models were less accurate at 70%. The classifying clinician was accurate 74% of the time. Our analysis shows that both ELMo and TLC are sensitive to the symptoms of disorganization in patients. In this study methods using text representations from language models were more accurate than those based solely on the assessment of FTD, and can be used as measures of disordered language that complement human clinical ratings.

- How to detect formal thought disorder (FTD)?
- Embeddings from LSTM language models (ELMo) can more accurately detect/predict FTD than individual indicators (benchmark: coherence model, which can somehow be viewed as traditional dictionary-based NLP method).
- Accuracy (N=70, 35 healthy and 35 patients):
  - ELMo: 80%
  - Coherence models: <70%
  - Clinician: 74%

[HTML] Detecting formal thought disorder by deep contextualized word representations

J Sarzynska-Wawer, A Wawer, A Pawlak… - Psychiatry …, 2021 - Elsevier

Computational linguistics has enabled the introduction of objective tools that measure some of the symptoms of schizophrenia, including the coherence of speech associated with formal thought disorder (FTD). Our goal was to investigate whether neural network based utterance embeddings are more accurate in detecting FTD than models based on individual indicators. The present research used a comprehensive Embeddings from Language Models (ELMo) approach to represent interviews with patients suffering from schizophrenia …

☆ Save 99 Cite Cited by 14561 Related articles All 24 versions Web of Science: 74 »

26

26

Application of Transformer: Polarized Framing of Immigration

PNAS RESEARCH ARTICLE | COMPUTER SCIENCES OPEN ACCESS

Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration

Dallas Card [a,b,1], Serina Chang [a], Chris Becker [c], Julia Mendelsohn [b], Rob Voigt [d,e], Leah Boustan [f,g], Ran Abramitzky [c,g], and Dan Jurafsky [a,b]

Edited by Joseph Ferrie, Northwestern University, Evanston, IL; received November 10, 2021; accepted June 15, 2022 by Editorial Board Member Kenneth W. Wachter

July 29, 2022 | 119 (31) e2120510119 | https://doi.org/10.1073/pnas.2120510119

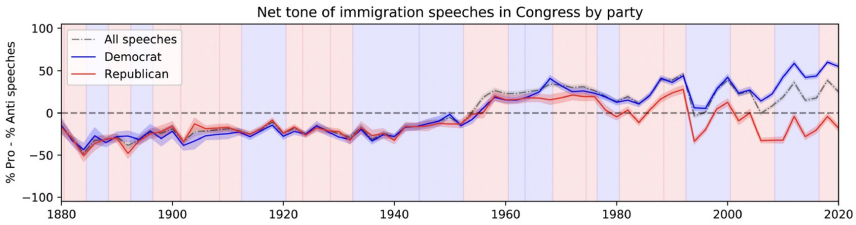
Significance

In the first comprehensive quantitative analysis of the past 140 y of US congressional and presidential speech about immigration, we identify a dramatic rise in proimmigration attitudes beginning in the 1940s, followed by a steady decline among Republicans (relative to Democrats) over the past 50 y. We also reveal divergent usage of positive (e.g., families) and negative (e.g., crime) frames—over time, by party, and between frequently mentioned European and non-European groups. Finally, to capture more suggestive language, we introduce a method for measuring implicit dehumanizing metaphors long associated with immigration (animals, cargo, etc.) and show that such metaphorical language has been significantly more common in speeches by Republicans than Democrats in recent decades.

We classify and analyze 200,000 US congressional speeches and 5,000 presidential communications related to immigration from 1880 to the present. Despite the salience of antiimmigration rhetoric today, we find that political speech about immigration is now much more positive on average than in the past, with the shift largely taking place between World War II and the passage of the Immigration and Nationality Act in 1965. However, since the late 1970s, political parties have become increasingly polarized in their expressed attitudes toward immigration, such that Republican speeches today are as negative as the average congressional speech was in the 1920s, an era of strict immigration quotas. Using an approach based on contextual embeddings of text, we find that modern Republicans are significantly more likely to use language that is suggestive of metaphors long associated with immigration, such as "animals" and "cargo," and make greater use of frames like "crime" and "legality." The tone of speeches also differs strongly based on which nationalities are mentioned, with a striking similarity between how Mexican immigrants are framed today and how Chinese immigrants were framed during the era of Chinese exclusion in the late 19th century. Overall, despite more favorable attitudes toward immigrants and the formal elimination of race-based restrictions, nationality is still a major factor in how immigrants are spoken of in Congress.

immigration | metaphor | dehumanization | framing | Congress

- Finetune RoBERTa to classify congressional speeches as proimmigration, antiimmigration or neutral.
- Quantitative analysis of 140 years of US congressional and presidential speech about immigration



Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration
D Card, S Chang, C Becker, J Mendelsohn, R Voigt, L Boustan, R Abramitzky, D Jurafsky
Proceedings of the National Academy of Sciences, 2022 · National Acad Sciences

We classify and analyze 200,000 US congressional speeches and 5,000 presidential communications related to immigration from 1880 to the present. Despite the salience of antiimmigration rhetoric today, we find that political speech about immigration is now much more positive on average than in the past, with the shift largely taking place between World War II and the passage of the Immigration and Nationality Act in 1965. However, since the late 1970s, political parties have become increasingly polarized in their

SHOW MORE

☆ Save 99 Cite Cited by 101 Related articles All 10 versions »

Application of Transformer: Remote Work

Remote Work across Jobs, Companies, and Space

Stephen Hansen, Peter John Lambert, Nicholas Bloom, Steven J. Davis, Raffaella Sadun & Bledi Taska

WORKING PAPER 31007 DOI 10.3386/w31007 ISSUE DATE March 2023

The pandemic catalyzed an enduring shift to remote work. To measure and characterize this shift, we examine more than 250 million job vacancy postings across five English-speaking countries. Our measurements rely on a state-of-the-art language-processing framework that we fit, test, and refine using 30,000 human classifications. We achieve 99% accuracy in flagging job postings that advertise hybrid or fully remote work, greatly outperforming dictionary methods and also outperforming other machine-learning methods. From 2019 to early 2023, the share of postings that say new employees can work remotely one or more days per week rose more than three-fold in the U.S and by a factor of five or more in Australia, Canada, New Zealand and the U.K. These developments are highly non-uniform across and within cities, industries, occupations, and companies. Even when zooming in on employers in the same industry competing for talent in the same occupations, we find large differences in the share of job postings that explicitly offer remote work.

- Use DistilBERT pre-trained on 1M text chunks of job vacancy postings to measure the work-from-homeness of the 250 M jobs (Work from Home Algorithmic Measure), achieving 99% accuracy that outperforms dictionary-based methods.
- The number of WFM jobs has risen significantly since 2019 and it differs w.r.t. different industries.

Remote work across jobs, companies, and space
S Hansen, PJ Lambert, N Bloom, SJ Davis, R Sadun… - 2023 - nber.org
The pandemic catalyzed an enduring shift to remote work. To measure and characterize this shift, we examine more than 250 million job vacancy postings across five English-speaking countries. Our measurements rely on a state-of-the-art language-processing framework that we fit, test, and refine using 30,000 human classifications. We achieve 99% accuracy in flagging job postings that advertise hybrid or fully remote work, greatly outperforming dictionary methods and also outperforming other machine-learning methods. From 2019 to …
☆ Save 99 Cite Cited by 36 Related articles All 20 versions »

Agenda

- Sequence-to-Sequence (Seq2seq) and Neural Machine Translation (NMT)
- Attention is All You Need
- Extensions, Adaptations, and Substitutes of Transformers

29

29

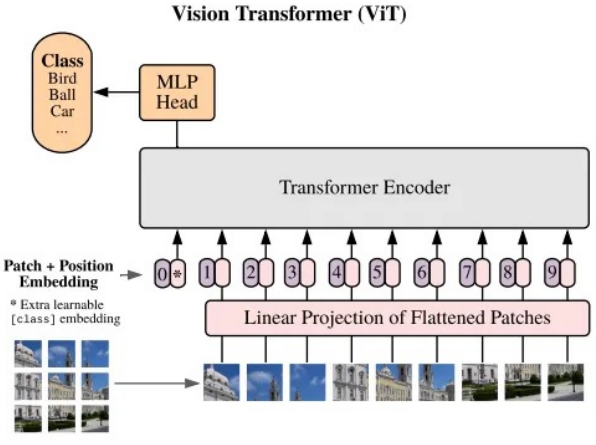
ViT (2020)

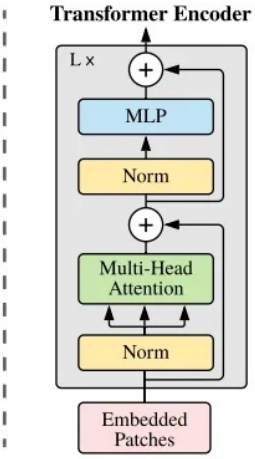
An image is worth 16x16 words: Transformers for image recognition at scale
A Dosovitskiy, L Beyer, A Kolesnikov… - arXiv preprint arXiv …, 2020 - arxiv.org
… To do so, we split an image into patches and provide the sequence of … Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image …
☆ 保存 99 引用 被引用次数：32273 相关文章 所有 17 个版本 »

- Transformer encoders can also effectively process image.
- Decompose images into patches of 16 by 16 pixels. Patch = token.

Transformer lacks the inductive biases of CNN: Translation invariance and locality.

It may be smart to combine CNN and transformers.





Reference: http://cs231n.stanford.edu/slides/2023/lecture_9.pdf

30

30

ViT vs. ResNet

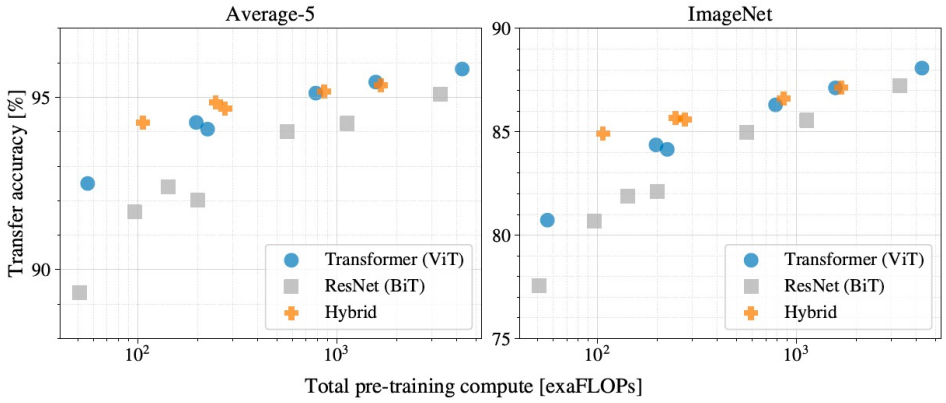


Figure 5: Performance versus pre-training compute for different architectures: Vision Transformers, ResNets, and hybrids. Vision Transformers generally outperform ResNets with the same computational budget. Hybrids improve upon pure Transformers for smaller model sizes, but the gap vanishes for larger models.

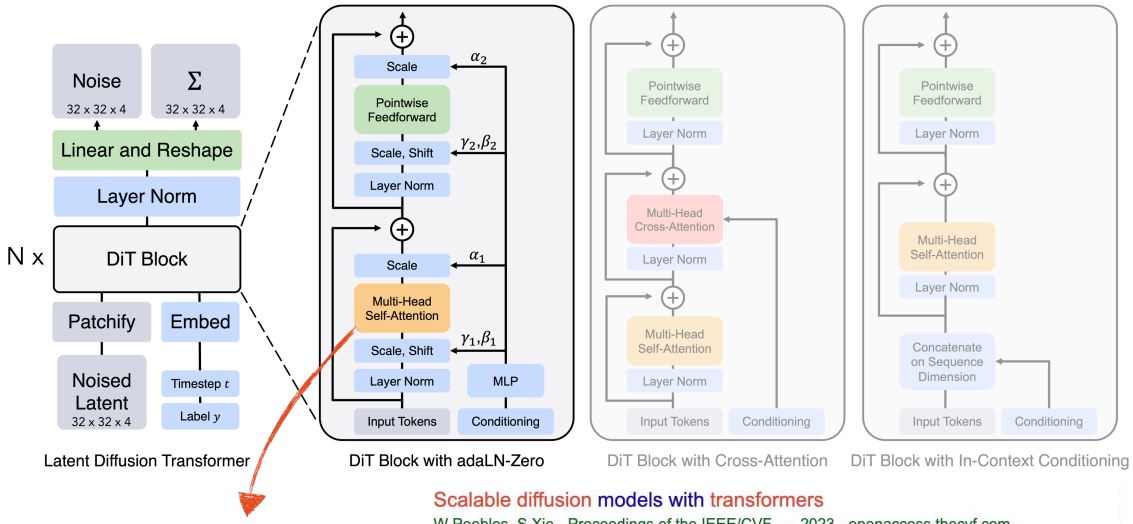
Attention is all you need, again!

Reference: http://cs231n.stanford.edu/slides/2023/lecture_9.pdf

31

31

Diffusion Transformer (DiT)



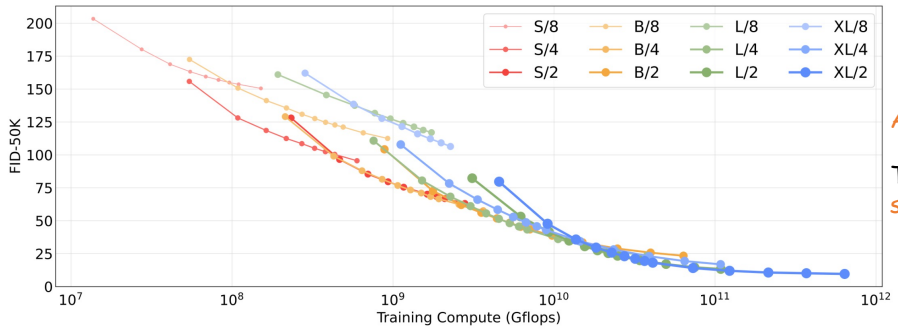
Replace the U-Net in latent diffusion models (LDMs) with a transformer.

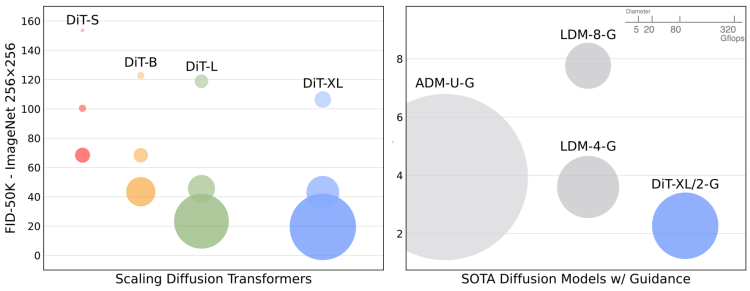
Scalable diffusion models with transformers
W Peebles, S Xie - Proceedings of the IEEE/CVF …, 2023 - openaccess.thecvf.com
… We train latent diffusion models of images, replacing the commonly-used U-Net … a transformer that operates on latent patches. We analyze the scalability of our Diffusion Transformers (…
☆ 保存 99 引用 被引用次数：1548 相关文章 所有 6 个版本 »

32

32

Diffusion Transformer (DiT) Performances





Attention is all you need again!

The bitter/sweet lesson: scaling laws.

33

33

Decision Transformer

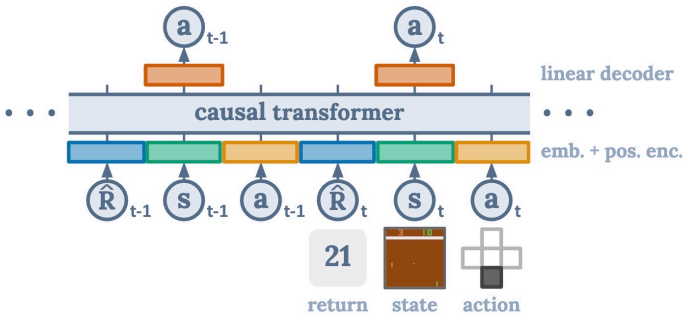
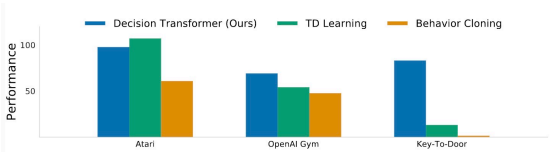


Figure 1: Decision Transformer architecture[1]. States, actions, and returns are fed into modality-specific linear embeddings and a positional episodic timestep encoding is added. Tokens are fed into a GPT architecture which predicts actions autoregressively using a causal self-attention mask.

- Mainstream RL: Learn value function or policy gradient.
- Decision transformer: Directly learn the next action based on an auto-regressive model.
- Multi-modal transformer: (reward to go, state, action)



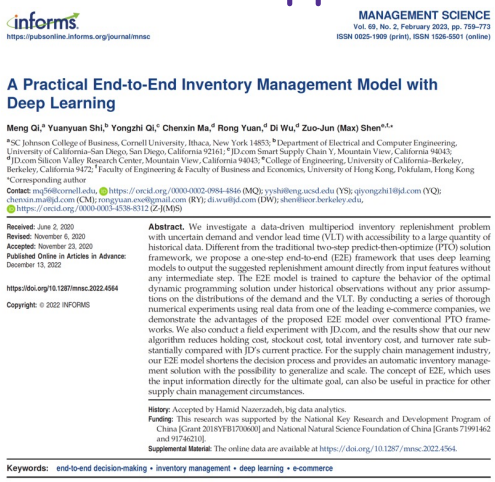
Reinforcement learning as conditional sequence modeling, like language models

Decision transformer: Reinforcement learning via sequence modeling
L Chen, K Lu, A Rajeswaran, K Lee… - Advances in neural …, 2021 - proceedings.neurips.cc
… of the Transformer architecture, and associated advances in language modeling such as GPT-x and BERT. In particular, we present Decision Transformer, … , Decision Transformer simply …
☆ 保存 99 引用 被引用次数：1748 相关文章 所有 11 个版本 »

34

34

BC Applied to Inventory Management



informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 69, No. 2, February 2023, pp. 759–773
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

A Practical End-to-End Inventory Management Model with Deep Learning

Meng Qi,[a] Yuanyuan Shi,[b] Yongzhi Qi,[c] Chenxin Ma,[d] Rong Yuan,[d] Di Wu,[d] Zuo-Jun (Max) Shen[e,f,*]

[a] SC Johnson College of Business, Cornell University, Ithaca, New York 14853; [b] Department of Electrical and Computer Engineering, University of California–San Diego, San Diego, California 92161; [c] JD.com Smart Supply Chain Y, Mountain View, California 94043; [d] JD.com Silicon Valley Research Center, Mountain View, California 94043; [e] College of Engineering, University of California–Berkeley, Berkeley, California 9472; [f] Faculty of Engineering & Faculty of Business and Economics, University of Hong Kong, Pokfulam, Hong Kong
*Corresponding author
Contact: mq56@cornell.edu, https://orcid.org/0000-0002-0984-4846 (MQ); yyshi@eng.ucsd.edu (YS); qiyongzhi18@jd.com (YQ); chenxin.ma@jd.com (CM); rongyuan.exe@gmail.com (RY); di.wu@jd.com (DW); shen@ieor.berkeley.edu, https://orcid.org/0000-0003-4538-8312 (Z-JM)S)

Received: June 2, 2020
Revised: November 6, 2020
Accepted: November 23, 2020
Published Online in Articles in Advance: December 13, 2022
https://doi.org/10.1287/mnsc.2022.4564

Abstract. We investigate a data-driven multiperiod inventory replenishment problem with uncertain demand and vendor lead time (VLT) with accessibility to a large quantity of historical data. Different from the traditional two-step predict-then-optimize (PTO) solution framework, we propose a one-step end-to-end (E2E) framework that uses deep learning models to output the suggested replenishment amount directly from input features without any intermediate step. The E2E model is trained to capture the behavior of the optimal dynamic programming solution under historical observations without any prior assumptions on the distributions of the demand and the VLT. By conducting a series of thorough numerical experiments using real data from one of the leading e-commerce companies, we demonstrate the advantages of the proposed E2E model over conventional PTO frameworks. We also conduct a field experiment with JD.com, and the results show that our new algorithm reduces holding cost, stockout cost, total inventory cost, and turnover rate substantially compared with JD's current practice. For the supply chain management industry, our E2E model shortens the decision process and provides an automatic inventory management solution with the possibility to generalize and scale. The concept of E2E, which uses the input information directly for the ultimate goal, can also be useful in practice for other supply chain management circumstances.

History: Accepted by Hamid Nazerzadeh, big data analytics.
Funding: This research was supported by the National Key Research and Development Program of China [Grant 2018YFB1700600] and National Natural Science Foundation of China [Grants 71991462 and 91746210].
Supplemental Material: The online data are available at https://doi.org/10.1287/mnsc.2022.4564.

Keywords: end-to-end decision-making • inventory management • deep learning • e-commerce

- Use multi-quantile RNNs to provide end-to-end predictions from features to the optimal inventory decisions, whereas most of the literature applies the predict-then-optimize paradigm.
- A field experiment shows that the e2e approach substantially reduces the inventory costs compared with some naïve benchmarks.
- Behavioral cloning (BC) applied to an MDP, with uncertainty completely removed.

A practical end-to-end inventory management model with deep learning
M Qi, Y Shi, Y Qi, C Ma, R Yuan, D Wu, ZJ Shen
Management Science, 2023 · pubsonline.informs.org

We investigate a data-driven multiperiod inventory replenishment problem with uncertain demand and vendor lead time (VLT) with accessibility to a large quantity of historical data. Different from the traditional two-step predict-then-optimize (PTO) solution framework, we propose a one-step end-to-end (E2E) framework that uses deep learning models to output the suggested replenishment amount directly from input features without any intermediate step. The E2E model is trained to capture the behavior of the optimal dynamic

SHOW MORE

☆ Save 99 Cite Cited by 65 Related articles All 4 versions Web of Science: 5 »

35

35

No Free Lunch: Quadratic Training Cost

- Training (and inference) cost of transformer: $O(n^2)$; cost of training RNN: $O(n)$.
  - Inference memory of transformer: $O(n)$; that of RNN: $O(1)$.
- **Sparse Attention** reduces the full pairwise attention calculations (such as sliding windows, fixed patterns, etc.) to lower computational complexity to O(nk); e.g., "Generating Long Sequences with Sparse Transformers" (2019)
- **Low-Rank Approximation** approximates the attention matrix using low-rank decomposition to reduce the computational complexity to O(nk); e.g., "Linformer: Self-Attention with Linear Complexity" (2020)
- **Kernel Methods** approximate self-attention calculations to reduce the complexity to O(nlogn); e.g., "Performer: Efficient Transformer with Linear Complexity" (2021)
- **Local Attention** uses sliding window-based attention to reduce the complexity to O(nk); e.g., "Longformer: The Long-Document Transformer" (2020)
- **Mixture of Experts (MoE)** uses multiple expert models and activate only a few at a time, adopted by the SOTA LLMs such as DeepSeek-V3; e.g., "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sharded Training"
- **Linear recurrence for parallel training** optimizes RNN architecture with parallelization; e.g., Mamba: Linear-Time Sequence Modeling with Selective State Spaces (2023)
  - Empirically, not as competitive as transformers for in-context learning.

36

36