Long-Range Transformer Architectures for Document Understanding

Thibault Douzon, Stefan Duffner, Christophe Garcia and Jérémy Espinas

August 25, 2023

ICDAR 2023 – VINALDO

Motivations

- Transformers provide best performance ;

Motivations

- Transformers provide best performance ;
- Transformers don't scale well with long sequences ;

Motivations

- Transformers provide best performance ;
- Transformers don't scale well with long sequences ;
- How to process long documents?
  - Multi-page documents
  - Dense text (scientific, legal, . . .)
  - Both?

Datasets - Business Documents

JMART

SALES ORDER

ORDER # 953487
DATE: JANUARY 13, 2013

JMART
1600 Boston Road
Springfield, MA 01129
(413) 543-0601

TO IDES US Inc.
1230 Lincoln Avenue
New York, NY 10019
Customer ID 300717

SHIP TO JMART
1600 Boston Road
Springfield, MA 01129
Customer ID 300717

SALES PERSON	JOB	SHIPPING METHOD	DELIVERY DATE	PAYMENT TERMS	DUE DATE
Terry Schmidt	953487	OVERNIGHT SHIPPING	1/15/13	Due on receipt	

QTY	ITEM #	DESCRIPTION	UNIT PRICE	LINE TOTAL
3	L-40C	Light Bulb 40 Watt clear 220/235V	438.00/carton	1314.00
4	L-40F	Light Bulb 40 Watt frosted 220/235V	432.00/carton	1728.00
4	L-60C	Light Bulb 60 Watt frosted 220/235V	430.15/carton	1720.60
3	L-60F	Light Bulb 60 Watt frosted 220/235V	435.08/carton	1305.24
2	L-80C	Light Bulb 80 Watt clear 220/235V	440.35/carton	880.70
4	L-80F	Light Bulb 80 Watt frosted 220/235V	453.21/carton	1812.84
12	PK-100	PK-100 Special carton high tech	14.40/PC	172.80
12	PK-102	PK-102 Pallet 120 x 80 x 12,5 Type B	13.15/PC	157.80
4	R-1141	PAQ Monitor, 20”, Color	300.00/PC	1200.00
2	R-1002	Maxitec R 3133 Personal Computer	1392.40/PC	2784.80
4	R-5002	Processor Pentium	530.00/PC	2120.00
3	M-15	SEC Multisync XV15	1187.70/PC	3563.10
4	R-1141	PAQ Monitor, 20”, Color	300.00/PC	1200.00
2	M-12	MAG DX 12F/FE	855.14/PC	1710.28
2	M-13	MAG DX 13F/FE	855.14/PC	1710.28

This is an offer to purchase. Written acceptance of the order or shipment signifies acceptance of the Terms and Conditions of Purchases printed on the back or otherwise available at www.jmart.com. Any additional or different terms proposed by Seller are rejected unless expressly agreed to in writing by the Buyer.

JMART

SALES ORDER

ORDER # 953487
DATE: JANUARY 13, 2013

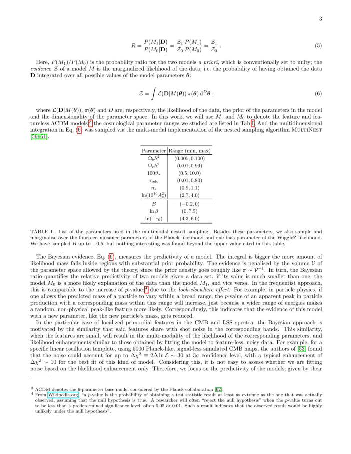
SALES PERSON	JOB	SHIPPING METHOD	DELIVERY DATE	PAYMENT TERMS	DUE DATE
Terry Schmidt	953487	OVERNIGHT SHIPPING	1/15/13	Due on receipt	

QTY	ITEM #	DESCRIPTION	UNIT PRICE	LINE TOTAL
3	L-40C	Light Bulb 40 Watt clear 220/235V	438.00/carton	1314.00
4	L-40F	Light Bulb 40 Watt frosted 220/235V	432.00/carton	1728.00
4	L-60C	Light Bulb 60 Watt frosted 220/235V	430.15/carton	1720.60
3	L-60F	Light Bulb 60 Watt frosted 220/235V	435.08/carton	1305.24
2	L-80C	Light Bulb 80 Watt clear 220/235V	440.35/carton	880.70
4	L-80F	Light Bulb 80 Watt frosted 220/235V	453.21/carton	1812.84
2	M-12	MAG DX 15F/FE	855.14/PC	1710.28
2	M-13	MAG DX 15F/FE	855.14/PC	1710.28
2	M-14	MAG DX 15F/FE	855.14/PC	1710.28
			TOTAL DISCOUNT	–
			SUBTOTAL	37272.66
			SALES TAX	1863.63
			TOTAL	39136.29

This is an offer to purchase. Written acceptance of the order or shipment signifies acceptance of the Terms and Conditions of Purchases printed on the back or otherwise available at www.jmart.com. Any additional or different terms proposed by Seller are rejected unless expressly agreed to in writing by the Buyer.

Multi-page customer order with annotations for information extraction.

Datasets - Docbank



3

$$R = \frac{P(M_1|\mathbf{D})}{P(M_0|\mathbf{D})} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\frac{P(M_1)}{P(M_0)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\ . \tag{5}$$

Here, $P(M_1)/P(M_0)$ is the probability ratio for the two models *a priori*, which is conventionally set to unity; the *evidence* $\mathcal{Z}$ of a model M is the marginalized likelihood of the data, i.e. the probability of having obtained the data $\mathbf{D}$ integrated over all possible values of the model parameters $\boldsymbol{\theta}$:

$$\mathcal{Z} = \int \mathcal{L}(\mathbf{D}|M(\boldsymbol{\theta}))\,\pi(\boldsymbol{\theta})\,\mathrm{d}^D\boldsymbol{\theta}\ , \tag{6}$$

where $\mathcal{L}(\mathbf{D}|M(\boldsymbol{\theta}))$, $\pi(\boldsymbol{\theta})$ and D are, respectively, the likelihood of the data, the prior of the parameters in the model and the dimensionality of the parameter space. In this work, we will use M_1 and M_0 to denote the feature and featureless ΛCDM models[3] the cosmological parameter ranges we studied are listed in Tab.I. And the multidimensional integration in Eq. (6) was sampled via the multi-modal implementation of the nested sampling algorithm MultiNest [59–61].

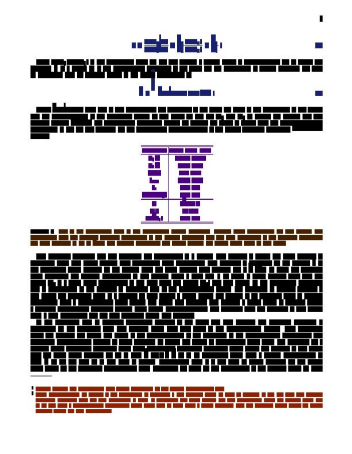
Parameter	Range (min, max)
$\Omega_b h^2$	$(0.005, 0.100)$
$\Omega_c h^2$	$(0.01, 0.99)$
$100\vartheta_*$	$(0.5, 10.0)$
$\tau_{\rm reio}$	$(0.01, 0.80)$
n_s	$(0.9, 1.1)$
$\ln(10^{10}A_s^2)$	$(2.7, 4.0)$
B	$(-0.2, 0)$
$\ln\beta$	$(0, 7.5)$
$\ln(-\tau_0)$	$(4.3, 6.0)$

TABLE I. List of the parameters used in the multimodal nested sampling. Besides these parameters, we also sample and marginalise over the fourteen nuisance parameters of the Planck likelihood and one bias parameter of the WiggleZ likelihood. We have sampled B up to -0.5, but nothing interesting was found beyond the upper value cited in this table.

The Bayesian evidence, Eq. (6), measures the predictivity of a model. The integral is bigger the more amount of likelihood mass falls inside regions with substantial prior probability. The evidence is penalised by the volume $\mathcal{V}$ of the parameter space allowed by the theory, since the prior density goes roughly like $\pi \sim \mathcal{V}^{-1}$. In turn, the Bayesian ratio quantifies the relative predictivity of two models given a data set: if its value is much smaller than one, the model M_0 is a more likely explanation of the data than the model M_1, and vice versa. In the frequentist approach, this is comparable to the increase of p-values[4] due to the *look-elsewhere effect*. For example, in particle physics, if one allows the predicted mass of a particle to vary within a broad range, the p-value of an apparent peak in particle production with a corresponding mass within this range will increase, just because a wider range of energies makes a random, non-physical peak-like feature more likely. Correspondingly, this indicates that the evidence of this model with a new parameter, like the new particle's mass, gets reduced.

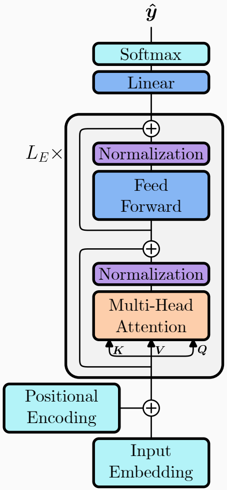
In the particular case of localized primordial features in the CMB and LSS spectra, the Bayesian approach is motivated by the similarity that said features share with shot noise in the corresponding bands. This similarity, when the features are small, will result in the multi-modality of the likelihood of the corresponding parameters, and likelihood enhancements similar to those obtained by fitting the model to feature-less, noisy data. For example, for a specific linear oscillation template, using 5000 Planck-like, signal-less simulated CMB maps, the authors of [53] found that the noise could account for up to $\Delta\chi^2 \equiv 2\Delta\ln\mathcal{L} \sim 30$ at 3σ confidence level, with a typical enhancement of $\Delta\chi^2 \sim 10$ for the best fit of this kind of model. Considering this, it is not easy to assess whether we are fitting noise based on the likelihood enhancement only. Therefore, we focus on the predictivity of the models, given by their

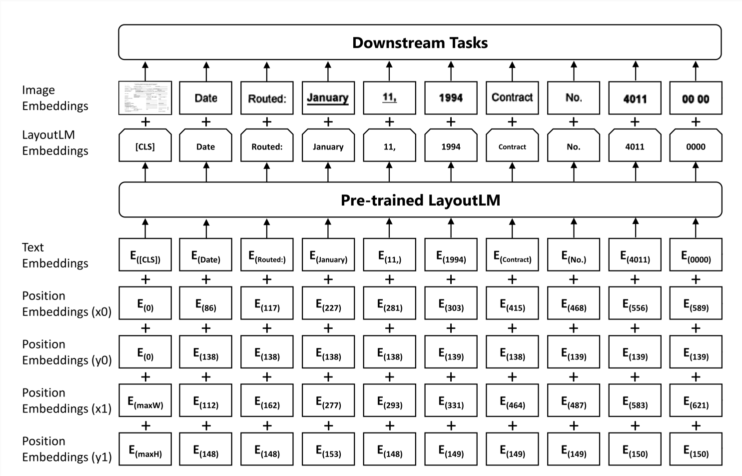
[3] ΛCDM denotes the 6-parameter base model considered by the Planck collaboration [62].
[4] From Wikipedia.org, "a p-value is the probability of obtaining a test statistic result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. A researcher will often "reject the null hypothesis" when the p-value turns out to be less than a predetermined significance level, often 0.05 or 0.01. Such a result indicates that the observed result would be highly unlikely under the null hypothesis".



Single-page document from Docbank [Li et al., 2020] with its layout annotation mask.

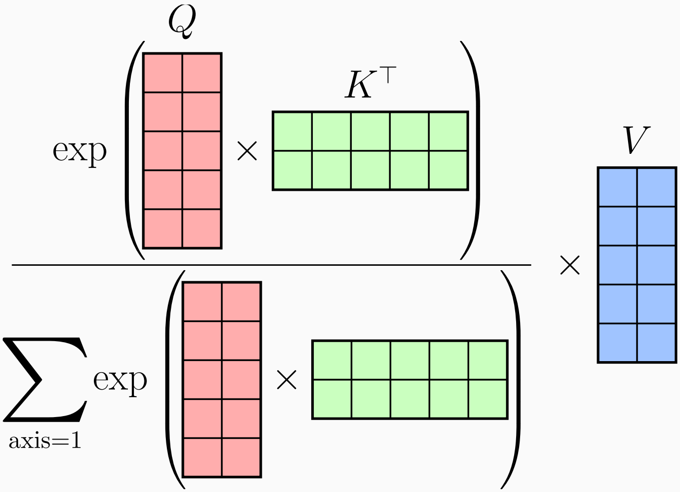
LayoutLM





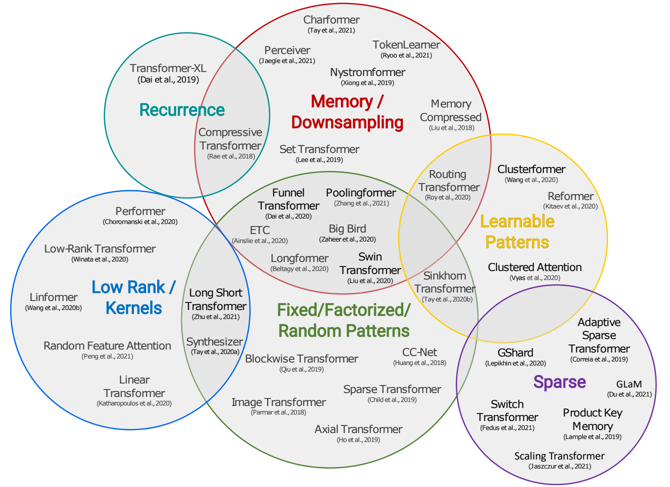
Transformer [Vaswani et al., 2017] and LayoutLM [Xu et al., 2020] architectures.

Classical Attention

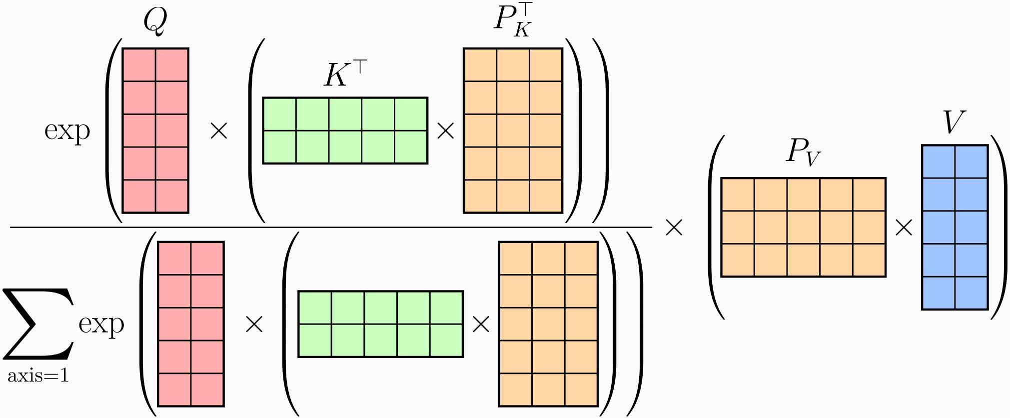


Default dot-product attention with $O(N^2)$ complexity.

Efficient Transformers

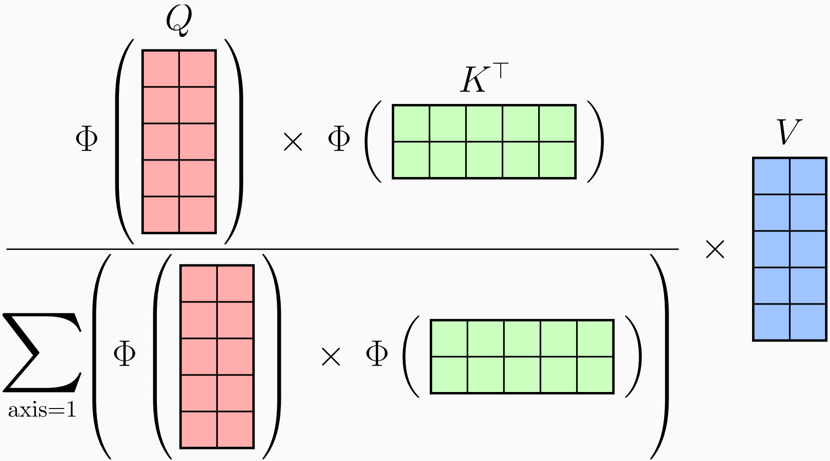


Efficient transformer architectures Venn diagram from Tay et al. [2022].



Linformer [Wang et al., 2020] attention approximation.

Key's and Value's sequence length dimensions are projected onto a smaller space.



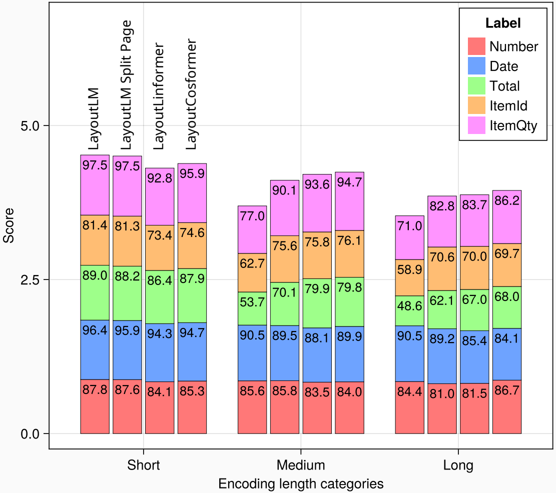
Cosformer [Qin et al., 2022] kernel attention approximation.
Where Φ is a kernel function, ReLU is used in this work.

Efficiency

Model Name	Time (s) / *Memory (GiB)* Sequence Length					
	512	1024	2048	4096	8192	16384
LayoutLM	1.41/*1.25*	2.83/*2.50*	7.39/*5.01*	23.43/*13.69*	-	-
LayoutLinformer	1.18/*1.35*	1.92/*2.26*	3.54/*3.28*	6.90/*5.19*	13.08/*8.96*	25.65/*16.78*
LayoutCosformer	2.03/*1.36*	2.50/*2.37*	4.68/*3.38*	9.00/*5.38*	17.23/*9.59*	33.96/*17.59*

Duration and memory use for various sequence lengths on a reference inference task.

Results - Business Documents



Cumulated F1-scores per document length category.

Relative Attention

Item	Quantity	Description	Unit Price	Total Price
1	10	T5F.L, DN15 5.015.F.L.K	$ 101,31	$ 1 013,10
2	4	T5F.L, DN25 5.025.F.L.K	$ 153,04	$ 612,16
3	2	T5F.L, DN32 5.032.F.L.K	$ 172,97	$ 345,94
4	10	T6F.L, DN15 6.015.F.L.K	$ 96,55	$ 965,50
5	4	T6F.L, DN25 6.025.F.L.K	$ 142,87	$ 571,48
6	2	T6F.L, DN32 6.032.F.L.K	$ 157,46	$ 314,92
7	3	T38V.E, DN15 38.015.V.E	$ 210,80	$ 632,40

A closer look at an order's line items table.

Relative Attention

Squircle (left) and Cross-shaped (right) attention bias patterns.

Results - Docbank

Model Name	F1 Score									
	Categories									Macro Average
	Abst.	Auth.	Capt.	Equa.	Footer	List	Sect.	Table	Title	
LayoutLM	97.8	87.5	94.9	87.2	90.5	84.0	92.8	85.7	88.6	91.6
LayoutLM$_{\text{SQUIRCLE}}$	**98.4**	90.2	**96.1**	89.7	92.0	**88.9**	**94.6**	87.7	90.3	**93.2**
LayoutLM$_{\text{CROSS}}$	**98.4**	**90.3**	96.0	89.6	**92.1**	88.7	**94.6**	87.5	**90.7**	**93.2**

Results - Docbank

Model Name	F1 Score									
	Categories									Macro Average
	Abst.	Auth.	Capt.	Equa.	Footer	List	Sect.	Table	Title	
LayoutLM	97.8	87.5	94.9	87.2	90.5	84.0	92.8	85.7	88.6	91.6
LayoutLM$_{\text{SQUIRCLE}}$	**98.4**	90.2	**96.1**	89.7	92.0	**88.9**	**94.6**	87.7	90.3	**93.2**
LayoutLM$_{\text{CROSS}}$	**98.4**	**90.3**	96.0	89.6	**92.1**	88.7	**94.6**	87.5	**90.7**	**93.2**
LayoutLinformer	97.9	88.9	93.7	90.0	91.1	87.9	91.3	87.6	88.7	92.3
LayoutCosformer	97.2	87.2	91.0	88.1	90.6	87.4	81.4	87.0	88.3	90.7
LayoutCosformer$_{\text{SQUIRCLE}}$	97.0	85.4	92.4	89.2	90.7	84.2	85.6	87.9	86.8	90.7
LayoutCosformer$_{\text{CROSS}}$	97.4	86.9	93.8	**91.2**	91.7	87.5	87.4	**89.0**	88.1	91.9

Takeaway

- Efficient transformer architectures can be valuable for document understanding ;

Takeaway

- Efficient transformer architectures can be valuable for document understanding ;
- But the cost of attention approximation negatively impact performance on short documents ;

Takeaway

- Efficient transformer architectures can be valuable for document understanding ;
- But the cost of attention approximation negatively impact performance on short documents ;
- 2D relative attention is hard to tune on efficient architectures and does not perform as anticipated.

Thanks everyone !

Any questions ?

Paper & contact information

Any in-depth question about this work ? Please contact me !

Thibault Douzon, Stefan Duffner, Christophe Garcia and Jérémy Espinas

thibault.douzon@esker.com

Code and models will be available on
github.com/thibaultdouzon/long-range-document-transformer.git

Presentation theme

Get the source of this theme and the demo presentation from

`github.com/matze/mtheme`

The theme *itself* is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

References

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A Benchmark Dataset for Document Layout Analysis, November 2020.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosFormer: Rethinking Softmax in Attention. *arXiv:2202.08791 [cs]*, February 2022.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient Transformers: A Survey. *arXiv:2009.06732 [cs]*, March 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.*, June 2017.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-Attention with Linear Complexity. *arXiv:2006.04768 [cs, stat]*, June 2020.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, August 2020. doi: 10.1145/3394486.3403172.

Results - Business Documents - Relative Attention