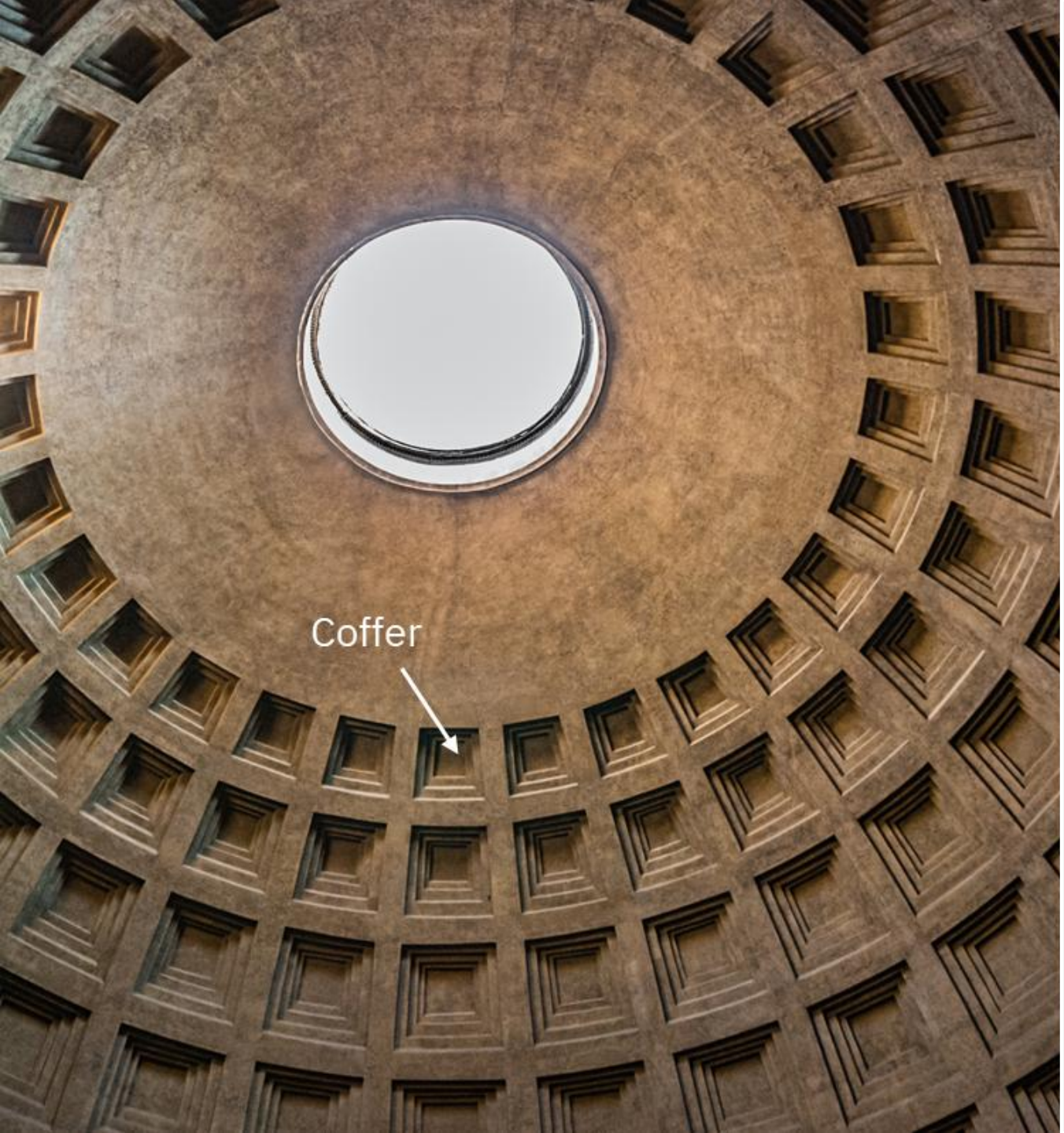
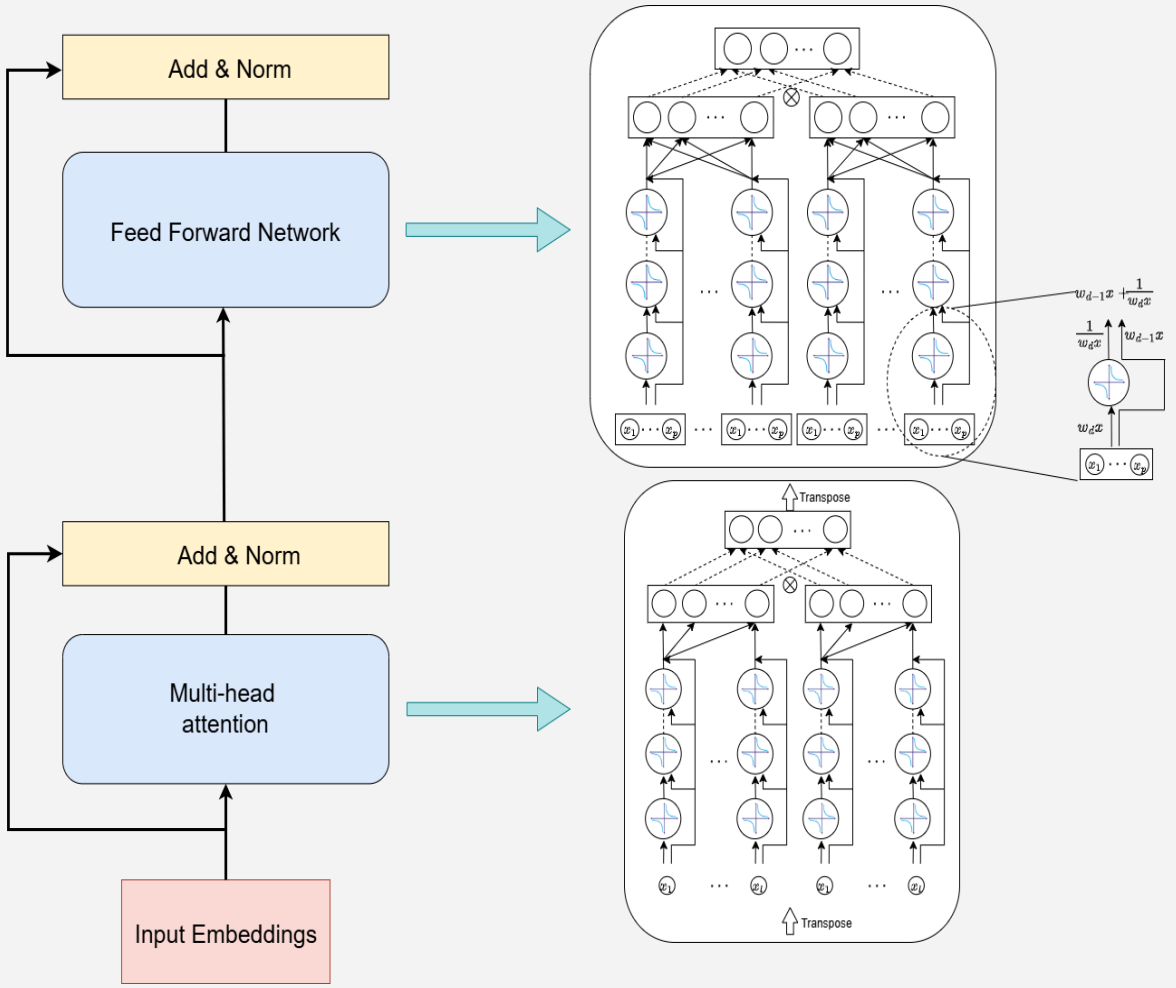


CoFrGeNets: Continued Fraction Architectures for Language Generation



Authors

Amit Dhurandhar

Enara Vijil

Dennis Wei

Tejaswini Pedapati

Karthikeyan Ramamurthy

Rahul Nair

But thanks to many other folks: David Cox, Kush Varshney, Hajar’s team, Nirmits Team, Jayants Team, Ahmed Nassar, Fabian Lim,...

CoFrGeNets

Contributions:

- Novel architectures for modeling Attention and MLPs in Transformers.
- Custom gradients using “continuants” that speed up training.
- Incremental training strategy to stabilize deep ladder training.

CoFrNet

- Represented as “ladder-like” sequence: $a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\ldots}}$
- can represent any real number, analytic function, including:
  trigonometric functions, polynomials, exponential functions, power function, special functions (gamma, hypergeometric, Bessel functions)
- best rational approximation of numbers and functions
- Fast convergence of approximations to real numbers

$$\phi = \frac{1+\sqrt{5}}{2} = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1+\cdots}}}$$

$$\tan(z) = \cfrac{z}{1 - \cfrac{z^2}{3 - \cfrac{z^2}{5 - \cfrac{z^2}{7 - \ddots}}}}.$$

Ratio of polynomials which is a rational function

CoFrNet

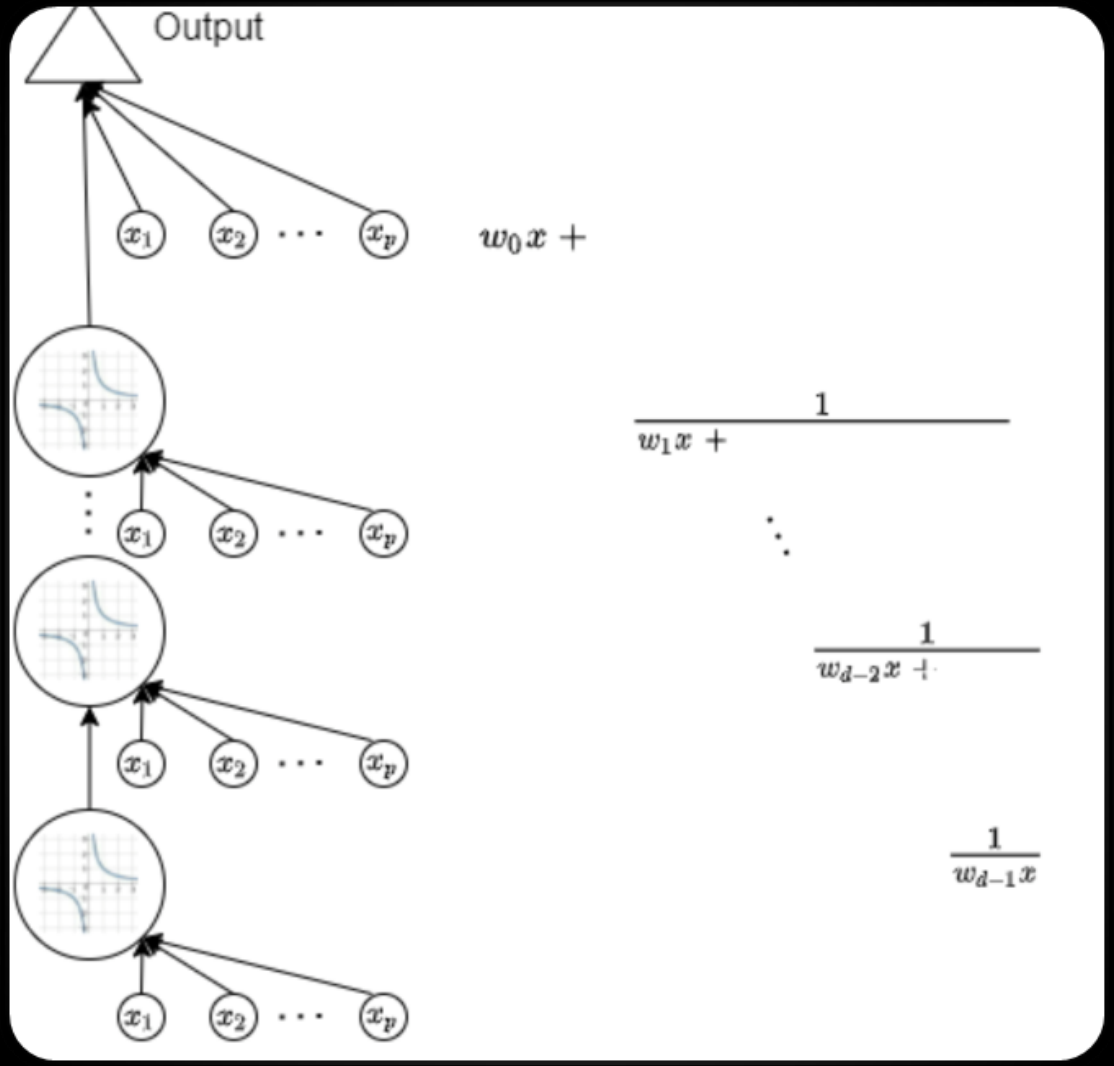
What is a CoFrNet?

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_{2+\ldots}}}$$

proposed "ladder-like" architecture replaces a_ks in CFs with linear functions - the input x multiplied by a weight vector w_k in each layer k

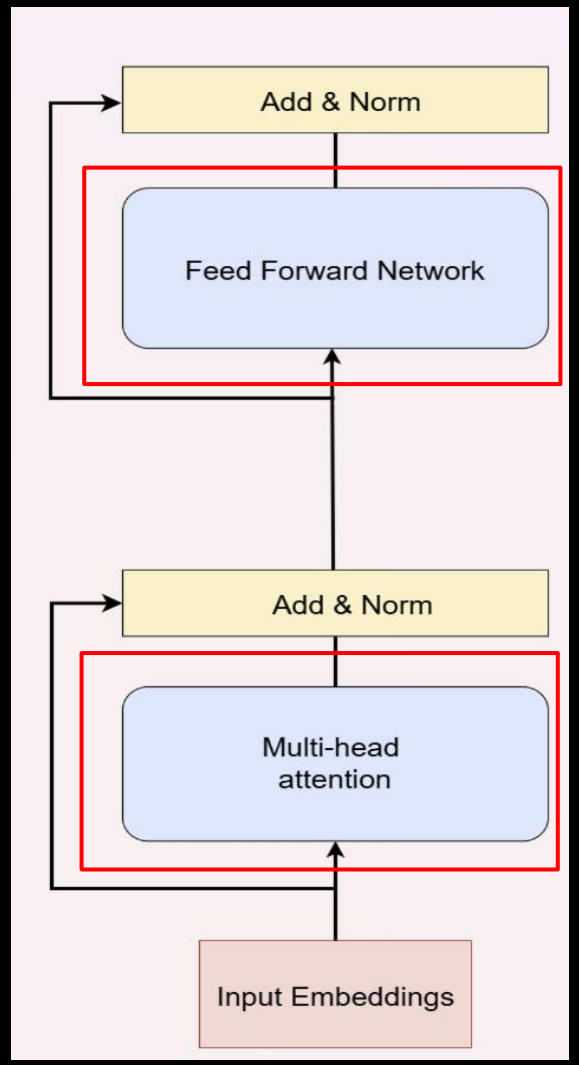
Reciprocal $\frac{1}{x}$ of the function is the nonlinearity in each layer (differs from commonly used ReLU, sigmoid, etc).

We show that linear functions are sufficient for universal approximation with a finite number of ladders



CoFrNets (Puri et. al. NeurIPS, 2021)

Can it be extended for generation?



NanoGPT (Shakespear Dataset)

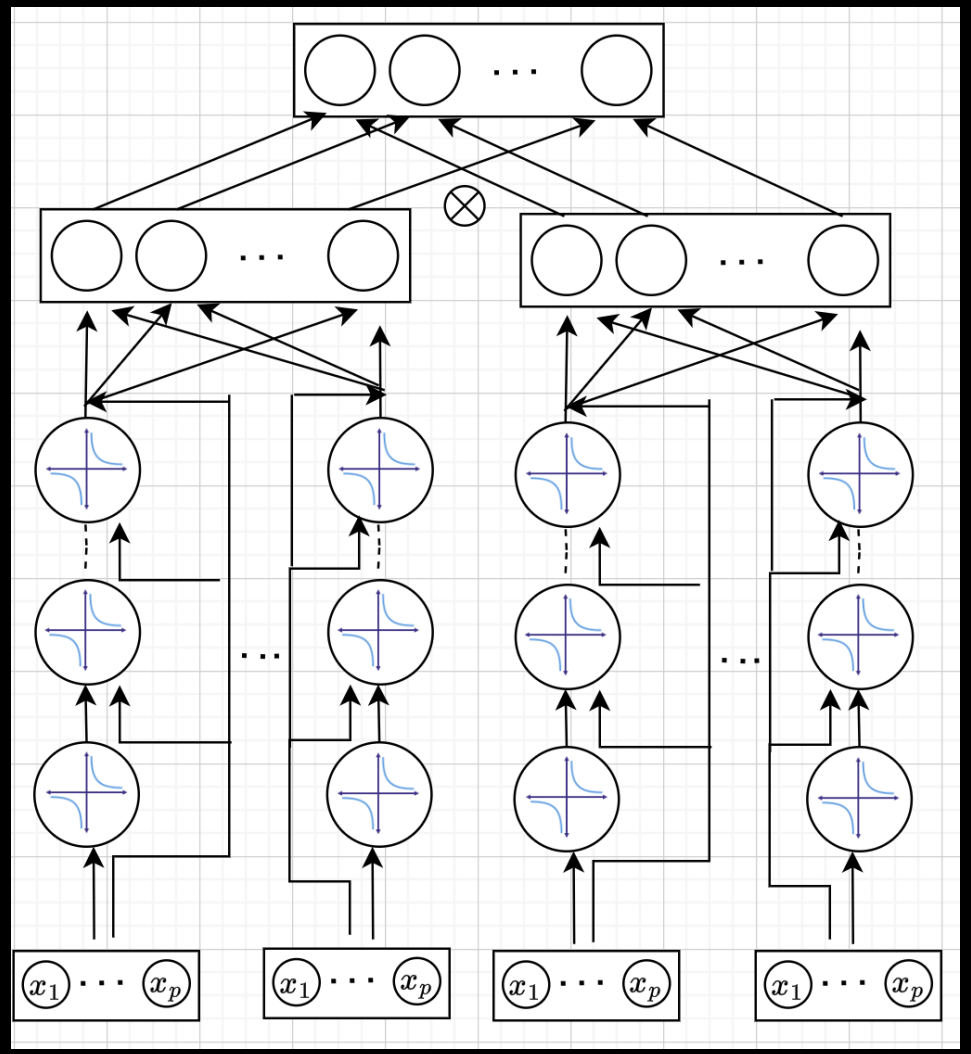
```
---------------

How chance it to repair it?

DUKE VINCENTIO:
If I am here be young as I lief as your
common a gap of prosperous settled as best
your tent?

DUKE VINCENTIO:
Honour, masters, and
the rod victors, are the characters of his
son: therefore lies like a rise desert ado her fee
him to the way, you will not visit there: 'tis in a very time a vile must
fellow on the voice; which he is not a wise, he comes for the particular, he with
the great Bohemia is devisitor.

ROMEO:
Against that Hermione, in the Cre
---------------
```

Val loss = 1.46

CoFrNet – MLP Replaced

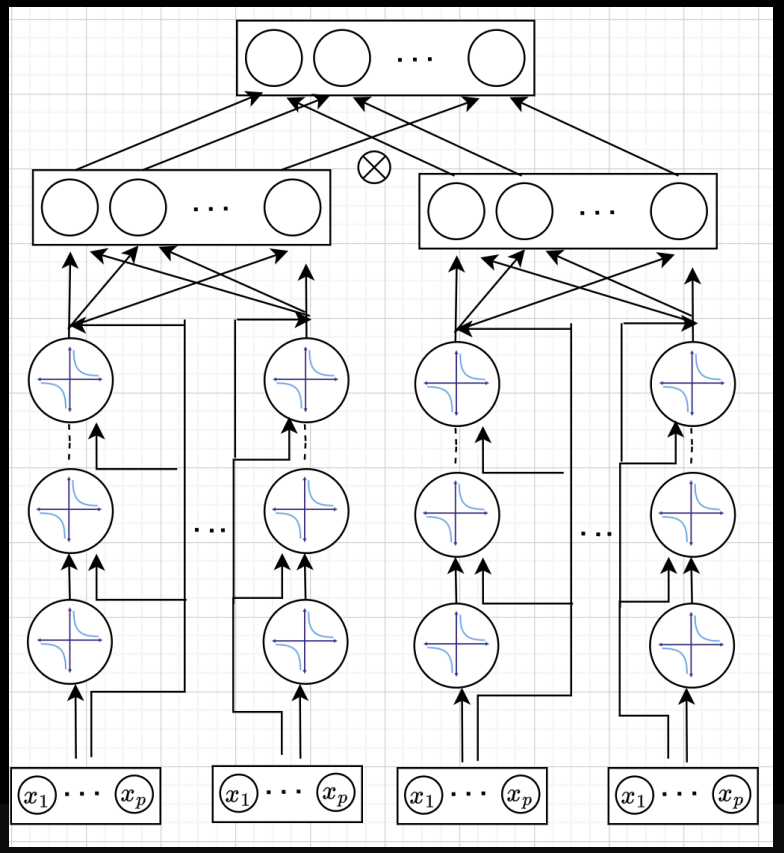
2 full ladder ensembles

```
---------------

He have not seat resay
The west did not of the part of this hold.

QUEEN MARGARET:
Well, the igness of the cousin the isabs,
If you gracious a thousand consul?

KING EDWARD IV:
The red Henry sorruble her should not well.

KING RICHARD II:
Play your she, sir, and you fee
And land--
Why, you were wars that rumber appy,
That becomes no come to make on the end,
Ox your said merst the honour, your sinces,
And the cause what is in dead,
If you have not speak to the it be hath end;
And child you shall
---------------
```



Val loss = 1.47

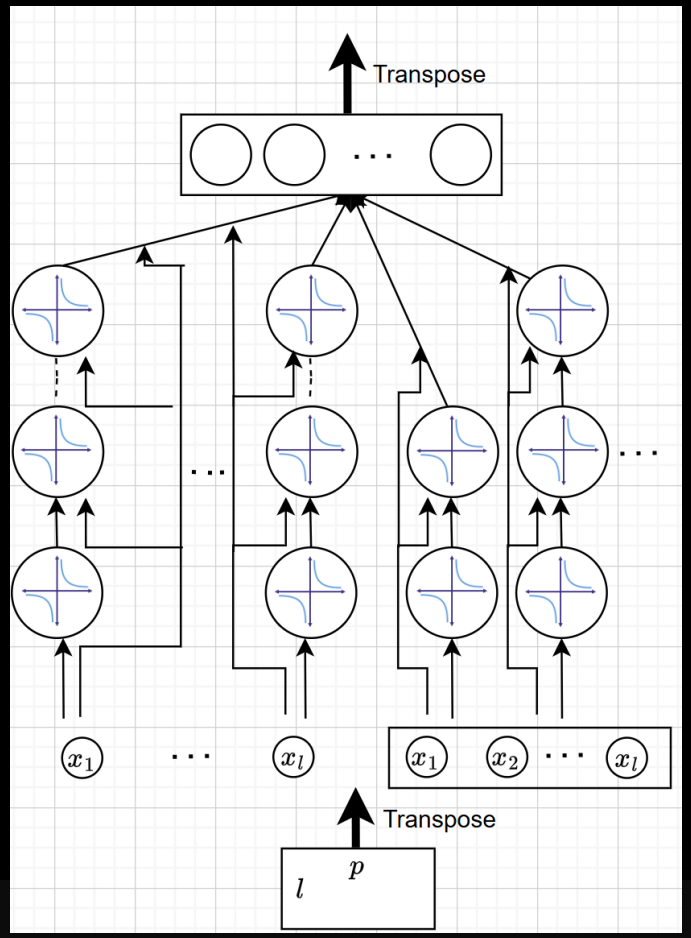
CoFrNet – with full ladders for attn.



Val loss = 0.009

```
---------------

 t    eepp eet   ppi e  e et  e   e iu  M  ti l  t: ehh   oetu ioo   Euo l   tio  teeheee  o te se i les   et iefa ooais
'ne ep mssio et hi3 tt e    tp strfoea  emhe suf lsM a liuI f  omtf at h re  ttfs oeeuloplYDtI Aoy  eucehrussqrnuueo iYe
etM u io b fs ie he do he.o ne
And
And -atw yofy the ucs ar in ath om he the ton wfores then ay for mu ahe mod the tou poord gowsaty mer thelled nome, yo i
t ine supra ht sa sines haue tounget, ton tory his th. the to Hfo the thaully,'se loce buo, Iad bots then
---------------
```

CoFrNet – with full ladders for attn.

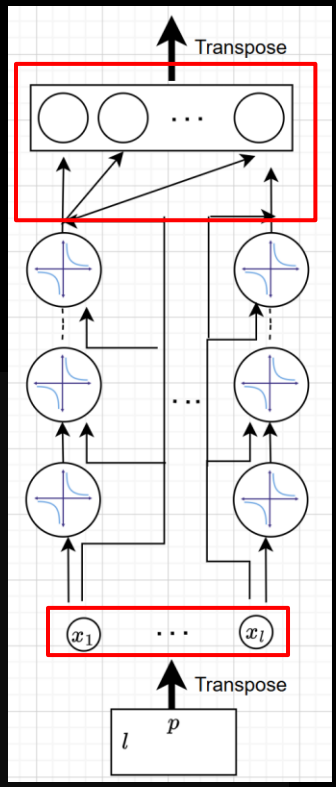


Val loss = 0.009

```
---------------

 t    eepp eet   ppi e  e et  e   e iu  M  ti l  t: ehh   oetu ioo   Euo l   tio  teeheee  o te se i les   et iefa ooais
'ne ep mssio et hi3 tt e    tp strfoea  emhe suf lsM a liuI f  omtf at h re  ttfs oeeuloplYDtI Aoy  eucehrussqrnuueo iYe
etM u io b fs ie he do he.o ne
And
And -atw yofy the ucs ar in ath om he the ton wfores then ay for mu ahe mod the tou poord gowsaty mer thelled nome, yo i
t ine supra ht sa sines haue tounget, ton tory his th. the to Hfo the thaully,'se loce buo, Iad bots then
---------------
```

CoFrNet – Attn. Replaced

Causal linear layer, diagonalized cofrnet ladders



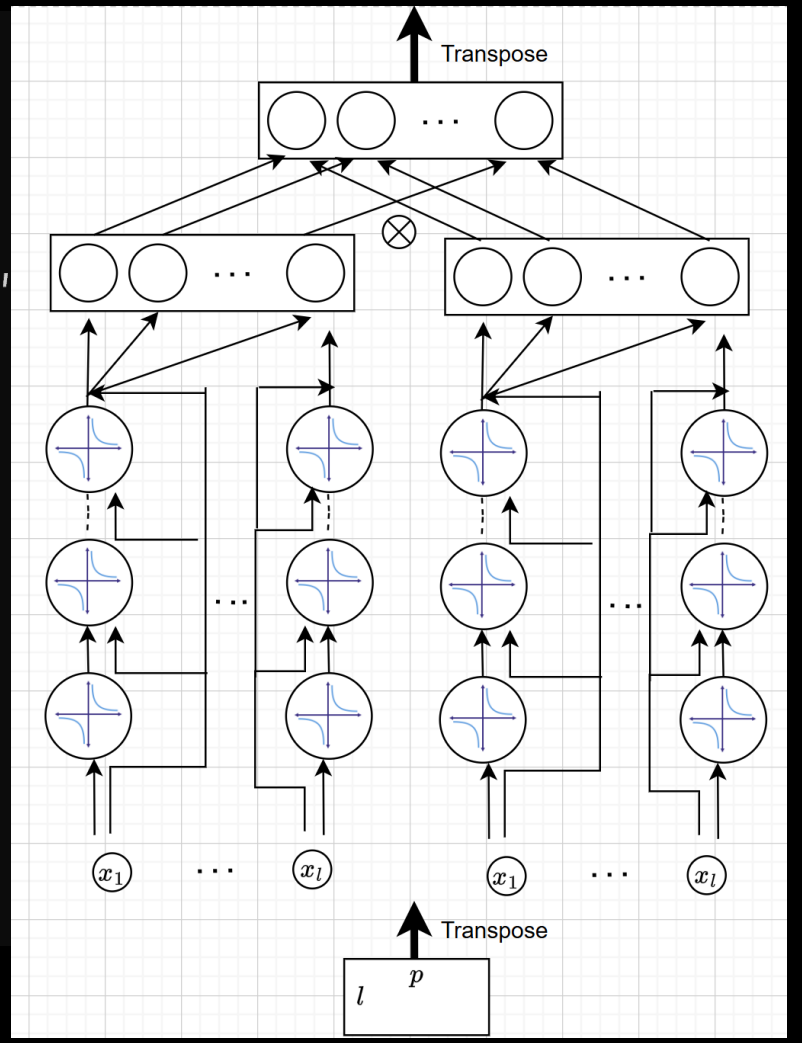

```
--------------

GLOUCESTER:
Lord, I will not be so, 'tis set them like envious like such vanity
His counter'd soon sound me in the state
In cretain soften times me to the bloody of the mother,
That shall not you call the crown her hope,
Is let man our sween believes?

Second Sercument be to be in and smile very hell; I wound him arrant to does as in the duke.

CAMMILLO:
O, but know's could not a rhim all the givorcester'ds more on
To relike to sit with the bond.
Lourney under his wrong Richmond,
That hands it w
--------------
```

Val loss = 1.59

CoFrNet – Attn. Replaced

Causal linear layers, 2 diagonalized cofrnet ladders, product of causal linear layers

```
--------------

lI never servant him.

Second Servingman:
You are to desire the gates and the rest state the father bows, and let him never groans,
in the like a breast.
Ah, and is heaven after sounds at once
Takes to seen the death offenced to be so strawberried?

HENRY BOLINGBROKE:
What thinks their dead!
Alack proud the duke.

KING RICHARD III:
I think of Gaunt and fear
As Marcius, editioner,
To thrust the extrembling in in this war,
When the officers of mount my good looks my l
--------------
```

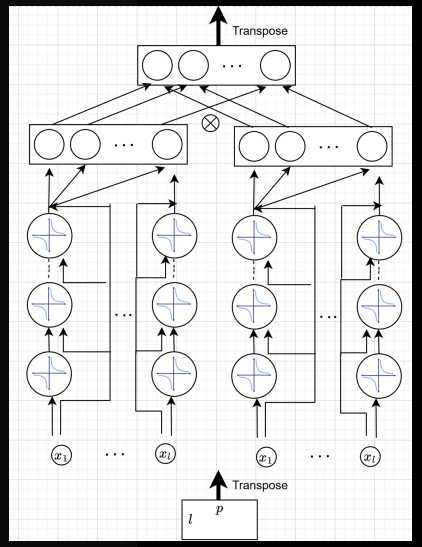


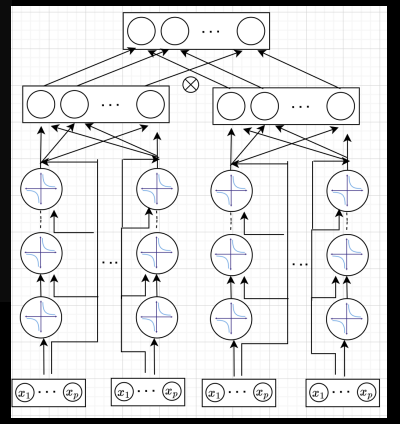
Val loss = 1.55

CoFrNet – Attn. and MLP Replaced

Causal linear layers, 2 (diagonalized for attn. and diag ladder of ladders for mlp) cofrnet ladders, product of causal linear layers

```
--------------

He have not traitors, and her revengerous. How not the king,
He would constain'd they slave a gave the to the sweet so sacred love foul thing.

ROMEO:
I know now the groans,
For heart more wronger what shall worthy would act
Out you be content your shed such a like.

KING EDWARD--

RIVERD:
Her shall I am rome;
And your with on the creder'd answer
Thought of harm;
Answer mercy the consentent it is to pray the party in being thee, to his fortune.

LUCIDIUS:
You still hath encounter, sir,
How my re
--------------
```





Val loss = 1.61

CoFrNet

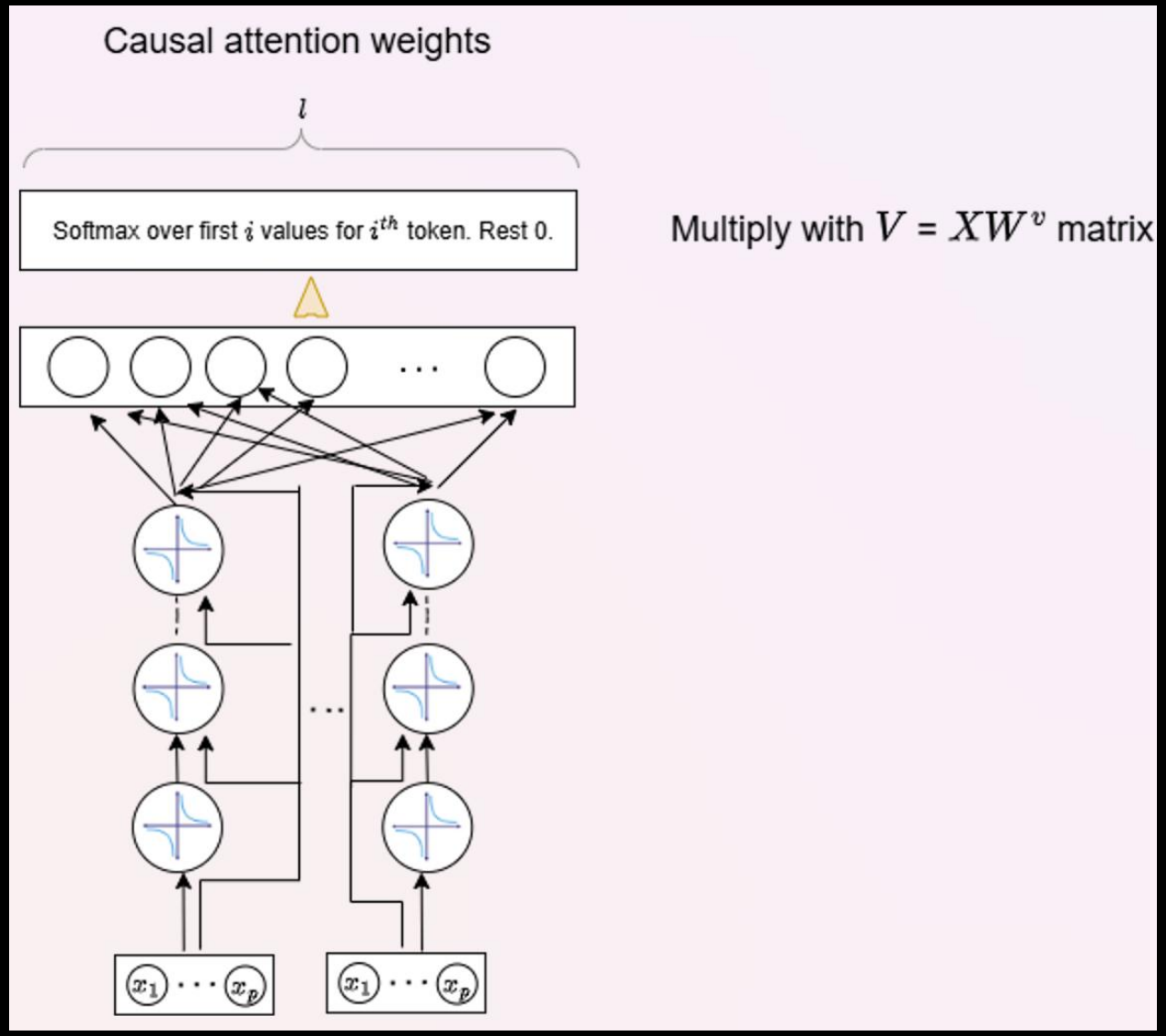
New attention implementation:

Parms: $L(p+l)+p^2$

L=1 gives similar or better results
to our previous architecture
with parms $l(l+2d+1)$ i.e. order l^2

Model	Attn.	Val loss	Data
GPT2 (124M)	Old	3.16	OWT (IT)
GPT2 (124M)	New	3.10	OWT (IT)
Llama (165M)	Old	3.22	docling
Llama (165M)	New	3.08	docling
GPT2-xl (1.5B)	Old	2.66	OWT (IT)
GPT2-xl (1.5B)	New	2.64	OWT (IT)

Parms (Standard attention): $4p^2$



GPT2 Both Replaced

Attention	CAttnU	CAttnM	FFN	Cffn
$4p^2$	$l(2d + l + 1)$	$L(p + l) + p^2$	$2\alpha p^2$	$2Lp(d + 1)$

The second reason for the need for the patent was that the court accepted the facts that the patent does not give the same benefits as the patent does. The court established that the patent does not give the same benefits as the patent does. The court also established that the patent does not give the same benefits as the patent does.

The third reason for the necessary appeal was that the patent did not give the same benefits as the patent does. The court established that the patent does not give the same benefits as the patent does. The court also established that the patent does not give the same benefits as the patent does.

The fourth reason for the need for the patent was that the patent did not give the same benefits as the patent does. The court established that the patent does not give the same benefits as the patent does. The court also established that the patent does not give the same benefits as the patent does.

The Fifth reason for the need for the patent was that the patent did not give the same benefits as the patent does. The court established that the patent does not give the same benefits as the patent does.

The Sixth reason for the need for the patent was that the patent did not give the same benefits as the patent does. The court established that the patent does not give the same benefits as the patent does.

The Seventh reason for the need for the patent was that the patent does not give the same benefits as the patent does. The court also established that the patent does not give the same benefits as the patent does.

The Eighth reason for the need for the patent was that the patent does not give the same benefits as the patent does. The court also established that the patent does not give the same benefits as the use of the software and the patent does not give the same benefits as the use of the software and the patent does not give the same benefits as the use of the software and the patent does not give the same benefits as the use of the software and the patent does not give the same benefits as the use of the software and the patent does not give the same benefits as the use of the software and the patent does not give the same benefits as the use of the software and the patent does not give the same benefits as the use of the software and the patent does not give the patent does not give the patent does not give the patent does not give the patent does not give the patent does not give the patent does
----

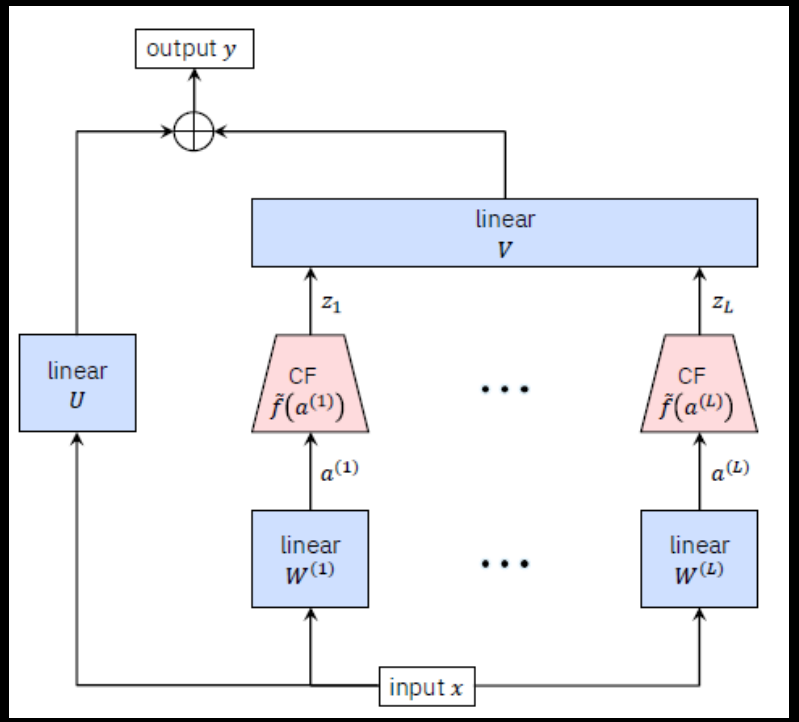
Val loss = 3.11

Continuant Formalism

Proposition 1. *The partial derivatives of continued fraction $\tilde{f}(a)$ defined in equation 2 are given by*

$$\frac{\partial \tilde{f}(a)}{\partial a_k} = (-1)^k \left(\frac{K_{d-k}(a_{k+1}, \ldots, a_d)}{K_d(a_1, \ldots, a_d)} \right)^2, \qquad k = 1, \ldots, d. \tag{9}$$

This implies for a d-depth ladder we now have to perform **only 1 division** as opposed d divisions.




```
class CoFrContinuant(torch.autograd.Function):
    """
    PyTorch Function to compute continued fractions and their gradients given partial denominators
    (and optionally partial numerators)
    """

    #Note that both forward and backward are @staticmethods
    @staticmethod
    def forward(ctx, a, b=None, epsilon=0.01):#, mask=None, output_direct=None):
        """
        Compute continued fractions given partial denominators (and optionally partial numerators)

        Parameters
        ----------
        ctx :
            PyTorch context object
        a : (batch_size, seq_len, width, depth) Tensor
            Partial denominators
        b : (batch_size, seq_len, width, depth) Tensor or None
            Partial numerators
        epsilon : float
            Parameter that caps reciprocal function at 1/epsilon in magnitude

        Returns
        -------
        output : (batch_size, seq_len, width) Tensor
            Continued fraction outputs

        """
```

Results

Table 2. Downstream task accuracies (best results bolded) on GLUE benchmark after finetuning. The first column is the pre-training dataset. Stds are reported in Table 9 in the appendix.

Model	MNLI	QQP	QNLI	SST2	COLA	MRPC	RTE	WNLI
OWT								
G-xl (1.5B)	86.89	88.93	91.35	93.56	81.78	79.83	60.27	58.28
C-F (972M)	**87.24**	**89.95**	**91.87**	**94.12**	**82.57**	**80.17**	**61.36**	**58.30**
C-A (1.21B)	86.94	89.31	91.74	93.83	81.77	79.89	60.91	58.28
C (790M)	87.11	89.36	91.79	93.91	81.97	79.93	61.25	58.29
S-D (1.2B)	84.93	86.82	90.13	91.34	80.15	77.95	59.83	58.28
S-A (1.21B)	85.27	86.38	90.93	92.72	80.76	77.42	59.36	58.27
GW								
G-xl (1.5B)	78.28	86.83	**82.93**	91.82	74.18	77.72	60.19	**58.33**
C-F (972M)	**79.23**	**87.24**	82.69	**92.38**	**74.79**	**78.04**	**61.37**	**58.33**
C-A (1.21B)	78.42	86.17	82.51	91.86	74.15	77.37	60.85	**58.33**
C (790M)	79.05	86.98	82.12	92.13	74.38	77.95	61.11	**58.33**
S-D (1.2B)	77.56	86.35	80.38	91.25	73.27	76.73	59.26	58.24
S-A (1.21B)	77.67	86.41	80.77	91.16	72.83	76.62	59.39	58.28

Table 3. Perplexities of the different variants with GPT2-xl.

Model	PTB	Wikitxt2	Lbda	AgNews	Lm1b	Wiki103
OWT						
G-xl (1.5B)	30.12	18.30	8.66	37.13	41.20	17.50
C-F (972M)	**29.94**	**17.09**	**8.15**	**35.72**	**40.11**	**16.17**
C-A (1.21B)	30.02	18.22	8.54	37.02	41.03	17.26
C (790M)	30.03	17.96	8.55	36.47	40.86	17.17
S-D (1.2B)	31.47	19.35	9.92	39.84	41.94	18.91
S-A (1.21B)	31.23	18.78	9.13	38.82	42.05	18.82
GW						
G-xl (1.5B)	29.07	19.12	31.78	45.62	52.36	18.93
C-F (972M)	29.83	**18.08**	30.55	**41.77**	**46.59**	**18.11**
C-A (1.21B)	**28.89**	18.77	30.98	43.91	48.37	18.67
C (790M)	29.08	18.29	30.71	42.55	48.01	18.42
S-D (1.2B)	30.83	19.25	31.92	46.81	52.99	19.03
S-A (1.21B)	29.36	18.95	31.23	46.38	52.83	19.45

Table 4. Training time and inference time. C_B is our basic implementation not using continuants. As can be seen using the continuants formalism speeds up training and inference.

Data	Model	Train Time (hrs)	Inf. Time (μs)
OWT	G-xl	190	643.93±1.73
	C-F	186	627.48±1.85
	C-A	186	638.26±1.76
	C	178	628.73±1.66
	C_B	203	5898.72±3.91
GW	G-xl	413	638.26±2.73
	C-F	397	627.34±1.65
	C-A	396	625.86±1.78
	C	387	619.78±1.49
	C_B	424	5877.87±4.52

Results

Dyadic Training Schedule: Start training the linear layer and then after ½ the number of total iterations start training depth 1, after 3/4th iterations start training depth 2 and so on...

Table 5. Perplexities of CoFrGeNet (GPT2-xl) variants with (top number) and without (below number) incremental training. As seen our training schedule has significant impact.

Model	PTB	Wikitxt2	Lbda	AgNews	Lm1b	Wiki103
			OWT			
C-F (972M)	**29.94**	**17.09**	**8.15**	**35.72**	**40.11**	**16.17**
	33.72	26.71	12.56	42.18	47.28	22.65
C-A (1.21B)	30.02	18.22	8.54	37.02	41.03	17.26
	38.24	21.82	10.92	45.52	46.21	24.25
C (790M)	30.03	17.96	8.55	36.47	40.86	17.17
	36.77	23.87	15.23	42.72	49.44	23.33
			GW			
C-F (972M)	29.83	**18.08**	30.55	**41.77**	**46.59**	**18.11**
	35.88	25.55	37.33	45.46	49.53	20.44
C-A (1.21B)	**28.89**	18.77	30.98	43.91	48.37	18.67
	33.71	23.72	36.28	45.29	52.51	21.67
C (790M)	29.08	18.29	30.71	42.55	48.01	18.42
	34.22	22.98	36.23	44.39	51.91	21.67

Results

Table 6. Zero-shot accuracies on open domain Q&A, reasoning and text understanding tasks. The docling data mix of 2 trillion tokens was used for pre-training.

Model	Opqa	Piqa	Arc	Wino	Hswag	Lbda	Boolq	Sciq
Llama (3.2B)	.282	.76	.77	**.654**	**.503**	**.581**	**.691**	**.941**
C-F (1.9B)	.292	**.764**	.765	.643	.482	**.581**	.659	**.941**
C-A (2.5B)	.304	.752	.757	.646	.463	.575	.633	.914
C (1.7B)	.283	.751	.751	.64	.464	.571	.633	.907
Mamba-2 (3.2B)	**.324**	.761	.768	.615	.486	.548	.655	.919

Table 7. Throughput for Llama-3.2B and our variants.

Model	Tokens/day	Train Time (days)
Llama (3.2B)	235B	8.5
C-F (1.9B)	288B	7
C-A (2.5B)	250B	8
C (1.7B)	315B	6.4

CoFrNet

Value Proposition:

- Novel function class for modeling Attention and MLPs.
- Can be seamlessly integrated into current architectures
- Requires much fewer parameters (2/3 to 1/2 in multiple cases)
- Given the structure possibly much more efficient hardware implementations possible (main operation is matrix-vector multiplication and non-linearity is reciprocal)

CoFrNet

Tons of exciting future work:

- Replace components in other architectures such as Mamba, MOEs, etc.
- Come up with simpler strategies to stabilize divisions rather than incremental training in software and hardware.
- Test on other modalities, multimodal models, world models.
- Explore if having multiple intervention points in the architecture can be exploited to train more robust and safe models.
- Study how quantization maybe used to further reduce size given the unique functional form.
- Study if custom white box distillation strategies can be developed.

CoFrNet

Where we were ...

➔ Implementation that was slow (because of divisions) and architectures restricted to classification

Where we are now...

➔ Generative architectures, faster training and smaller models at billions of parameter scale

Thank You