Understanding the ultra low frequency cluster in sparse autoencoders

Author - Pavan Katta

Main results

- Ultra high frequency cluster majorly consists of features which are activated when the model's token prediction has very high loss.
- The average loss curve associated with feature activation displays a phase-transition-type behaviour, particularly noticeable when transitioning from high to low frequency clusters.
- Features within the low frequency cluster that have a low average loss appear more likely to be interpretable, though this aspect requires further research.
- I pick some features in the ultra low frequency cluster using the above heuristic and demonstrate their interpretability.

Background

Recently anthropic's MI team found great success extracting(Bricken et al) interpretable features out of a MLP layer of a 1L transformer. One interesting detail is that the feature density histogram seems to have 3 clusters. Namely

- Dead neurons
- Ultra low density (Not Interpretable)
- Others (Interpretable)

Nanda et al replicated the results and found a similar bimodal graph with a ultra low frequency cluster of features.

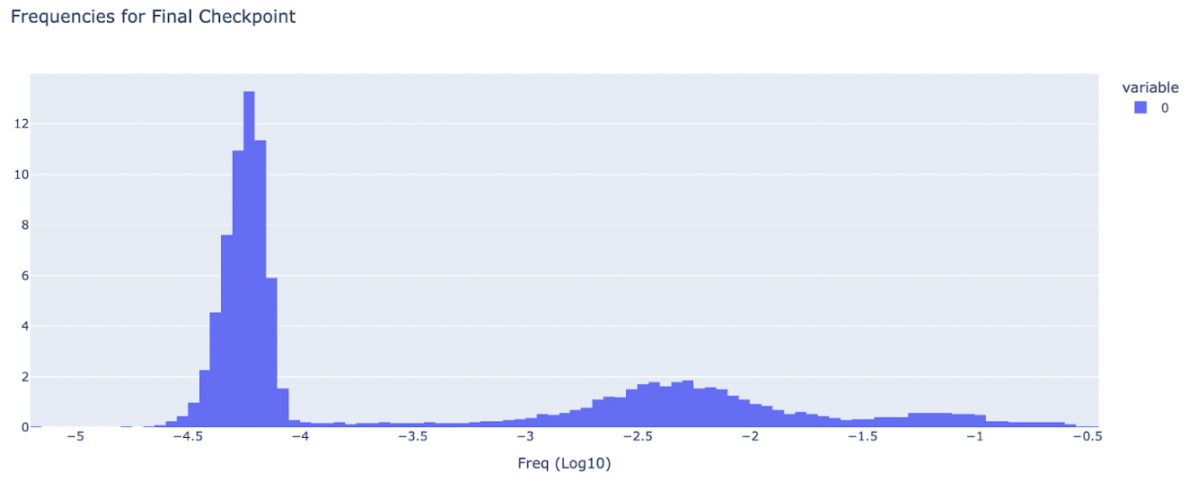


Image from Nanda et el

Ultra low frequency cluster is defined as features with frequency < 1e-4

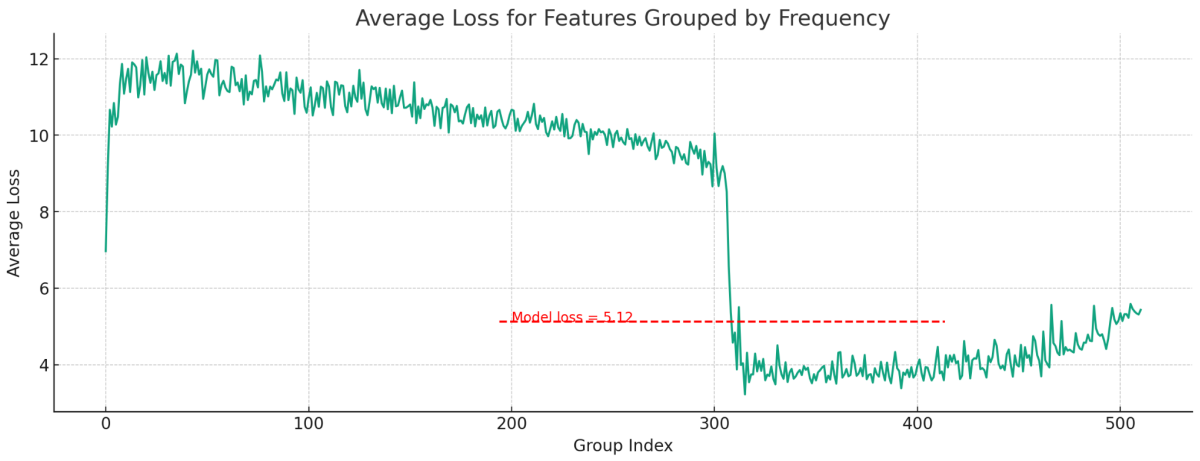
All the experiments were done using the open sourced autoencoder(run-1), 1L Gelu and training data from the [replication tutorial from] Nanda et al.

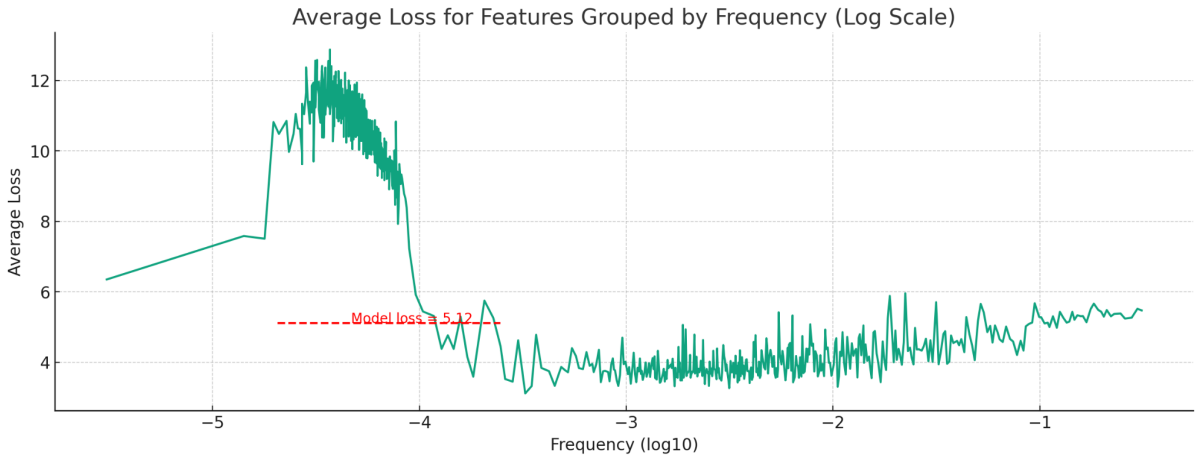
Results:

For each feature in the sparse autoencoder I looked at the avg loss when it was activated.

- Low frequency cluster is primarily activated on the tokens where the prediction loss of the model is very high.

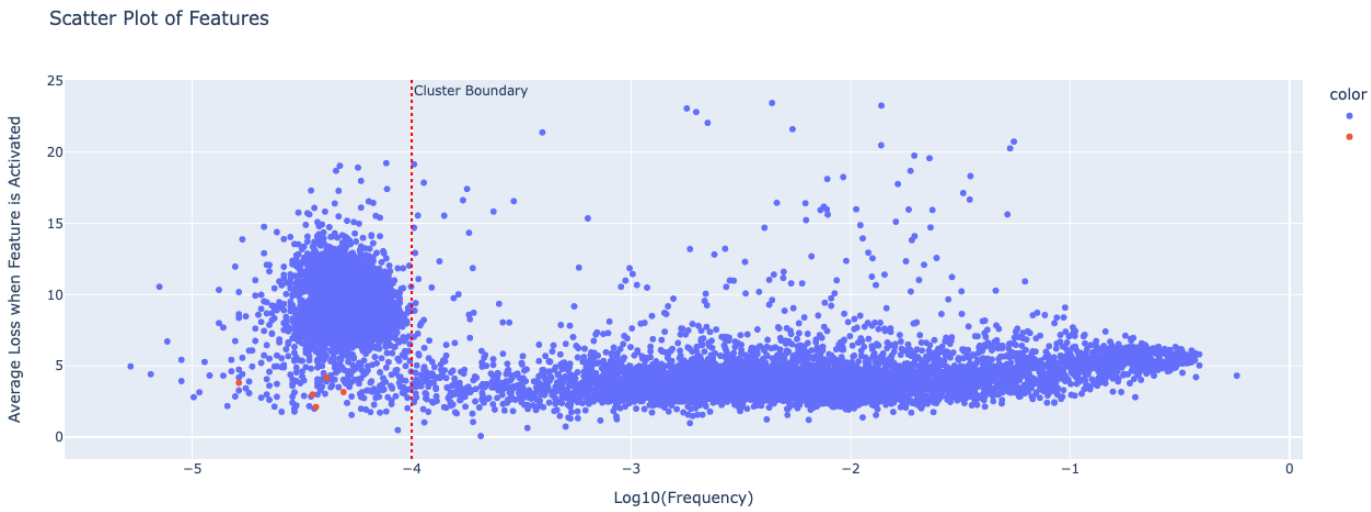
- There seems to be a sharp bump on avg loss when activated around the threshold of low frequency cluster(1e-4)





(Note: To mitigate noise, I grouped features into frequency-based buckets for the plots.)

The entire scatter plot looks as below



We see that

- Low frequency cluster features have very high loss, almost double the model's loss.
- High frequency cluster features have close to model loss.

Interestingly, while the majority of low frequency cluster features correspond to high loss, there are outliers with both low frequency and low average loss.

Given Bricken et al's heuristic finding that "Anecdotally, almost all of the features in the high density cluster are interpretable, but almost none of the features in the ultralow density cluster are."
It's interesting to see if there's any relation between loss and a feature being interpretable. The principled way to do this would be to use an autointerp (Bills et al)score for all features and check it with loss. But it's costly and will check with Bricken et al if they can run the analysis.

Another suggestive evidence could use the above loss information to help find interpretable features in the low frequency cluster which are considered uninterpretable.

We look at the features marked red in the scatter plot. We will look at the following features including a feature which is almost at the left end of the scatter plot.

Feature	Frequency	Percentile
2124	2.0e-05	0.598145
3560	2.8e-05	3.247070
15501	4.3e-05	24.005127
9686	4.6e-05	28.637695
11950	5.3e-05	42.089844

I find these features as interpretable as the ones in high frequency cluster.

Now let's take a look at one of these in detail and others are in the appendix.

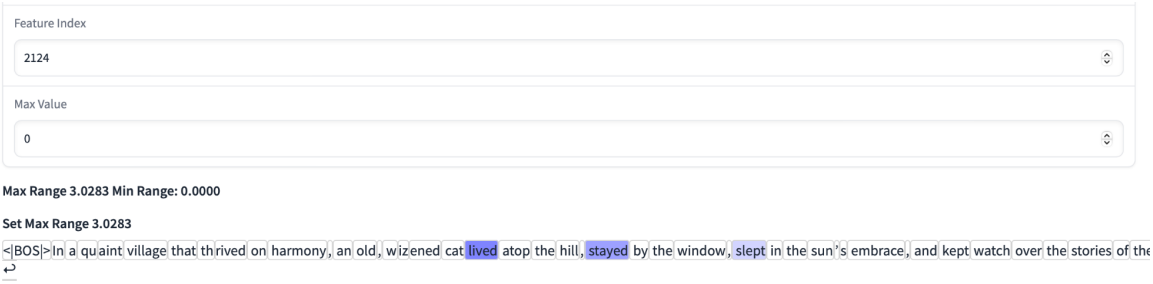
I used the visualisations used by Nanda et al in the tutorial.
Namely - Looking at activations, Looking at a sentence we generate, logits boosted.

Feature no - 2124
Frequency - 2e-05
Interpretation - lived and its synonyms.

Activations

	str_tokens	context	label	feature
782	·lived	·were·little.·We·did·not·have·much,·we$·lived$·in·a·house·that·was·barely·900·square	6/14	4.034832
4621	·lived	↩ Richard·(b.·1946)$·lived$·on·Grand·Street·from·1975-1	36/13	3.998774
3764	·lived	·accept·our·word.·And·now·that·you've$·lived$·away·from·home,·you've·changed!"·—	29/52	3.645726
5262	·lived	·back·home·after·having·gone·to·law·school·and·he$·lived$·in·Washington·as·a·public·defender·for·a·few·years	41/14	3.623543
3821	·lived	↩ When·my·mother·uttered·those·words,·I·had$·lived$·away·from·home·for·nearly·ten·years·since·starting·college	29/109	3.487822
3337	·lived	<|BOS|>·defence·lawyer·also·observed·that·had·the·woman$·lived$·in·any·other·region·of·the·UK,·she·"	26/9	3.232780
2272	·lived	ela·consigned·himself·to·the·same·fate.·He$·lived$·there·as·her·kokua,·or·help,	17/96	3.098993
4863	·lived	.·Her·mom·was·like·the·greatest·person·who·ever$·lived$	37/127	2.917739
3179	·stayed	·forth.·,↩ I·did·the·work.·I$·stayed$·calm.·Tried·not·to·use·words·that·would	24/107	2.551043
3854	·stayed	·each·run·in·5th·gear,·and·because·we$·stayed$·at·wide·open·throttle·for·several·hundred·feet·after·the	30/14	2.470767
5440	·stayed	·past·year.·But·the·vast·majority·of·them·have$·stayed$·in·the·Republic,·joining·tech·companies·(Google·has	42/64	2.401649
513	·stayed	<|BOS|>$·stayed$·together,·which·was·nice.↩ Approximately·under	4/1	2.381135
1947	·lived	·been·revealed·that·5,879·wild·elephants$·lived$·in·Sri·Lankan·forests·by·the·island·wide	15/27	2.279312
4003	·stayed	·unit.The·kitchen·was·well·equipped·.We·have$·stayed$·at·the·Calypso·since·it·first·opened·and	31/35	2.224121
2841	·stayed	·My·husband·and·I·and·our·9-year·old$·stayed$·with·my·sister·and·her·two·young·children·and·it	22/25	2.210056
5570	·lived	·not·far·from·the·safety·of·the·campus·where·she$·lived$.↩ Ms·Maasawre·was·attacked·at	43/66	1.802278
1220	·stayed	·implement·a·system·of·verifying·whether·the·respective·guest·actually$·stayed$·at·the·hotel·because·guess·what,·over·3	10/40	1.664600

1329	·stayed	·implement·a·system·of·verifying·whether·the·respective·guest·actually$·stayed$·at·the·hotel·because,·guess·what,·over·3	10/49	1.664600
3955	·stayed	·remained·around·210·degrees,·and·transmission·temps$·stayed$·around·183.·Without·the·goosene	30/115	1.517593
5082	·remained	·Jewish·scholars.·The·interests·of·the·Ottomans$·remained$·selective,·however,·because·of·their·feelings·of·moral	39/90	1.146454
2114	·slept	00·AM,·but·after·a·restless·night·I$·slept$·until·the·alarm·went·off.·For·some·reason·I	16/66	1.139289
4930	·slept	00·AM,·but·after·a·restless·night·I$·slept$·until·the·alarm·went·off.·For·some·reason·I	38/66	1.139289
4211	·kept	·my·then·2-year-old.·I·also$·kept$·buying·pants·in·larger·sizes·every·year.↩ My	32/115	1.033049
434	·staying	·her·story·here·+·on·her·blog.·She·is$·staying$·positive·+·upbeat·and·has·been·extremely·inspiring.	3/50	0.775939
1116	·remained	·all·their·goods;·it·is·not·recorded·that·any$·remained$·in·England.·In·1491·Henry·VII	8/92	0.768049
2421	·staying	·with·all·new·and·fresh·pics.·Hope·you·enjoy$·staying$·right·here.↩ Long·Casual·Summer·Dresses	18/117	0.715823
3699	·staying	·Hospital·Bratislava·initially·attributed·the·positive·effect$·staying$·in·a·cave·on·the·airways·that·inhaled·aeros	28/115	0.670062
2613	·remained	·father·Owen·to·the·executioner's·axe.·He$·remained$·in·touch·with·Margaret·of·Anjou,·Queen	20/53	0.635660
920	·staying	·Rod·and·OMG·their·crazy·happy·baby·Kyros$·staying$·with·us·during·this·glorious·wintry·weather.	7/24	0.559962
1447	·staying	·sugar.·I·have·not·been·all·that·successful·in$·staying$·away·from·processed·foods·with·them.·So,·I	11/39	0.464302
2465	F	·m0*ve*(fmass···1)/(F*fmass)·;print·"tend[	19/33	0.385344
20	·staying	·are·several·competent·organizations·with·professional·essay·penning·clubs$·staying$·recruited·by·individuals·principally·learners·with·your·purpose·of·essay	0/20	0.384806
1232	·staying	·OF·THESE·ACCOUNTS.↩ Instead·of$·staying$·in·the·box,·cracking·witty·jokes·and·p	9/80	0.374005
287	,	····def·__init__(self,·obj=None$,$·json=None,·value=None,·prop=	2/31	0.282910



Logits boosted

	token	logit
13096	·overnight	1.188174
8416	·closest	1.168526
7093	·opposite	1.124089
626	·there	1.117750
9884	·nearby	1.113974
24998	·halfway	1.097313
1045	·here	1.093630
3108	·behind	1.079099
21557	·awake	1.063767
48177	·dormant	1.059023
29758	·downstairs	1.036324

Discussion and Questions

- Are all the features in the low frequency cluster the same?
    - It is unlikely, as the features I have interpreted do not appear to be identical.
- What's happening in the average loss curve?

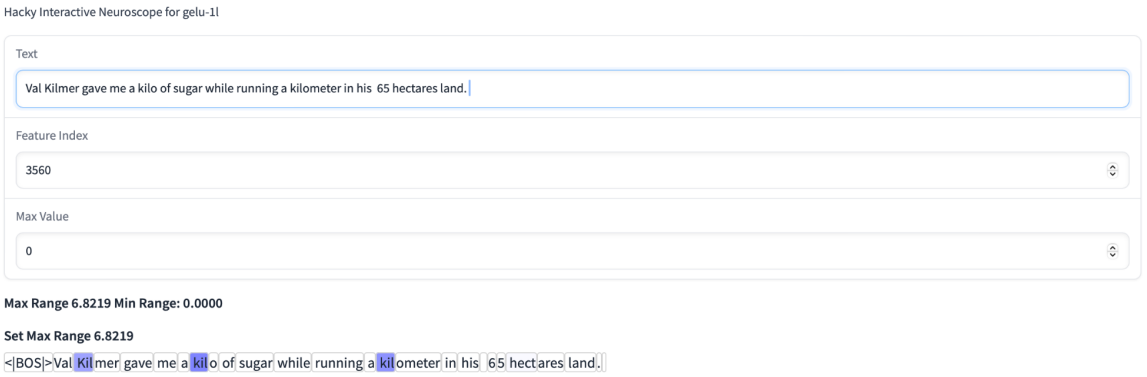
        - My intuitive guess would be that training the model (the original transformer) well to achieve a lower loss would get them to have low loss and move to high frequency cluster.
- Does average loss correlate with a feature's interpretability in a sparse autoencoder?
    - Currently, there is insufficient evidence to confirm this correlation, as a comprehensive analysis using autointerp scores for all features has not been done.
    - But some evidence is that we were able to randomly select some features in a low frequency cluster(which was previously thought to be uninterpretable) using this assumption and they were interpretable.
    - Would be great if it were true though!
- Are the high avg loss features in the high frequency cluster(there are quite some with loss >10) less interpretable
    - Working on this next

Replication - I primarily relied on the opensource models and colab tutorial made accessible by Neel Nanda. My results can be replicated through Sparseautoencoder analysis

Appendix

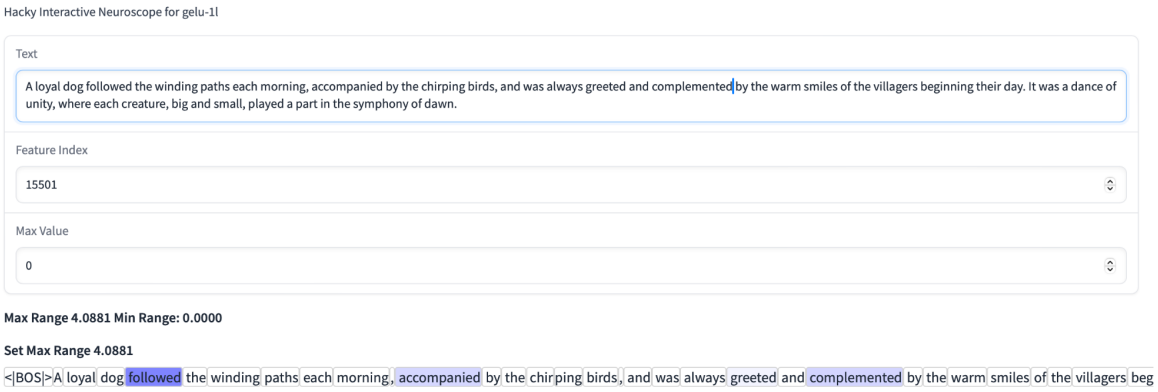
Other 4 interpretations

Feature no 3560
Interpretation - token ‘kil’ and very weakly on hect during hectare



-

Feature - 1501
Interpretation - Followed and words related to that.



Feature -9686

Interpretation - tokens ‘Er’ and ‘Ger’ and others like ‘Der’ ‘Unter’ .

- Note : this was not as interpretable and has likely a low interp score but not zero.

Feature - 11950

Interpretation - Conduct and words with conduct in context.