War is 15% conflic, 15% DragonMagazine

Giles Edkins, Lauren Greenspan, Dan Valentine

Interpretability Hackathon Write-Up

Apart Research

PIs: Esben Kran, Neel Nanda, Fazl Barez

Date: 13th November, 2022

Abstract

How does a transformer network represent concepts? Are they localized in activation space or in the learned parameters of the network, or else totally unlocalized?

We determined that:
- average activations give information about the prompt topic
- casual tracing suggests concepts cannot be easily localized
- "concept diffing" may give information about which attention heads are dealing with semantic, as opposed to grammatical, information
- we can create a basis for the activation vector space and in some cases express non-basis vectors as linear combinations of semantically related basis vectors

War is 15% conflic, 15% DragonMagazine

Goals

We set out on a preliminary (for us) investigation into how concepts are encoded by transformer networks. In order to ensure alignment with humans, future neural networks must learn concepts and use them appropriately. Understanding how attention heads and MLPs move and process conceptual information is therefore an important piece of the interpretability puzzle, and one that would have a big impact on AI safety. The original idea, “Don’t mention the war” was to perform surgery on a GPT until it stopped talking about war, directly or indirectly.

As well as watching information be copied between tokens, it is important to understand the space that information is stored in. The safety angle is: if concepts aren't stored as linearly

independent vectors, the model might try to create a superposition of two things and end up with something else entirely, which could lead to robustness failures or adversarial attacks. And having a map of where in concept space things make sense and where they conflict might help guard against these problems.

Among other things, we quickly learned that it is hard to understand what a concept *is,* let alone create a prompt that would allow us to see where the network stores this information. “War”, for example, is a concept, but details about specific wars are facts. By trying our hands at interpretability tools like EasyTransformers, causal tracing, and those from *Interpretability in the Wild*, we learned a lot about how this idea fits (or doesn’t) with the state-of-the-art, and gained a better intuition for future research.

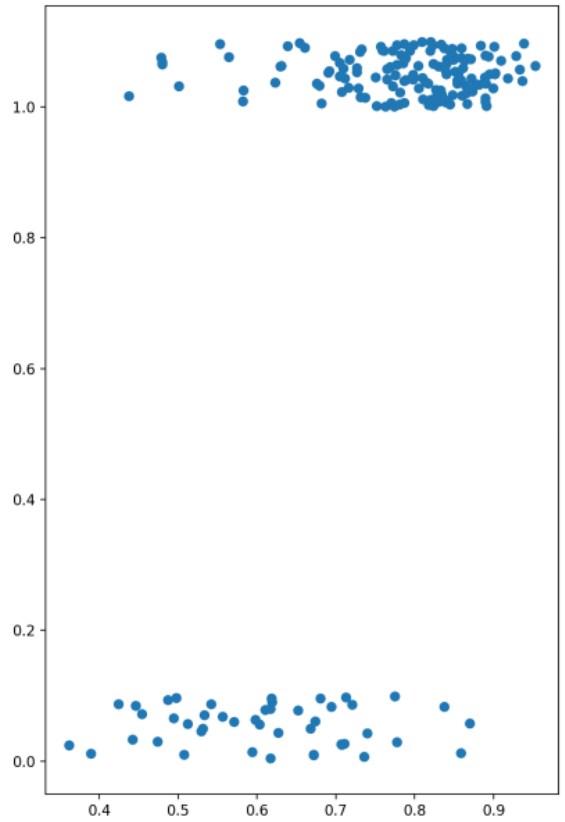
Many of our brief explorations are detailed below. Each was tested on no more than a handful of examples.

Investigations

Splitting data based on war/not war and looking for relevant activations (Giles)

Basic idea:

- Have a dataset, half of it is “about war”, half of it not
    - The “about war” half was Wikiquote: war
    - The rest was from various random wikiquote pages
- We want to find which parts of the transformer activate when it’s processing a war quote
    - The best predictor for whether it’s processing war might be a linear combination of neurons rather than a single neuron
- Split the data up into training/test
    - We’re not training the transformer; we’re training a simple linear model to predict war/not-war from the transformer activations
- Gather activations across all the training quotes
    - Choose an arbitrary layer, somewhere in the middle
    - Take the average across all tokens
- Train a linear classifier based on this
- Idea: whichever direction the linear classifier ends up pointing in, that’s the predictor of the war topic
- Test on the test set.



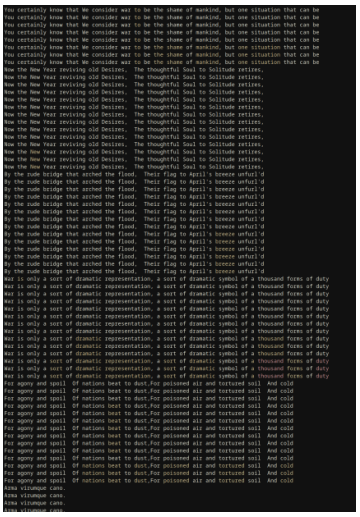
This plot shows the test set. The y axis is the ground truth with “not war” on the bottom and “war” on the top. The x axis is the prediction from the linear classifier. You can see that it accomplished something, but nothing spectacular.
https://github.com/fractal-pterodactyl/concept_detector/blob/main/scan/logreg.py

As a follow-up, I ran some prompts through the transformer, captured its activation, and then plotted the war/not-war prediction based on the linear model. Again, nothing spectacular - the red bits are maybe slightly more war-like but it’s nothing brilliant.


Your flaming torch aloft we bear,With burning heart an oath we swearTo keep the faith, to fight it through,To crush the foe or sleep with you In Flanders' fields.
When I'm watchin' my TV and a man comes on and tell meHow white my shirts can be.But, he can't be a man 'cause he doesn't smokeThe same cigarettes as me.
He Walks Through Walls Of Solid Steel And Stone... Into The 4th Dimension!
We think that the struggle against totalitarianism, Nazism and Communism, and the resistance movements, were the most important parts of 20th Century history. The expulsion, especially of Germans, was only a consequence of that.
When the rules of civilized society are suspended, when killing becomes a business and a sign of valor and heroism, when the wanton destruction of peaceable women and, children becomes an act of virtue, and is praised as a service to God and country, then it seems almost useless to talk about crime in the ordinary sense.
We are experiencing an accelerated obliteration of the planet's life-forms – an estimated 8,760 species die off per year – because, simply put, there are too many people. Most of these extinctions are the direct result of the expanding need for energy, housing, food and other resources. The Yangtze River dolphin, Atlantic gray whale, West African black rhino, Merriam's elk, California grizzly bear, silver trout, blue pike and dusky seaside sparrow are all victims of human overpopulation. Population growth, as E. O. Wilson says, is "the monster on the land." Species are vanishing at a rate of a hundred to a thousand times faster than they did before the arrival of humans. If the current rate of extinction continues, Homo sapiens will be one of the few life-forms left on the planet, its members scrambling violently among themselves for water, food, fossil fuels and perhaps air until they too disappear. Humanity, Wilson says, is leaving the Cenozoic, the age of mammals, and entering the Eremozoic – the era of solitude. As long as the Earth is viewed as the personal property of the human race, a belief embraced by everyone from born-again Christians to Marxists to free-market economists, we are destined to soon inhabit a biological wasteland.
I came, I saw, God overcame.
Those who expect to reap the blessings of freedom, must, like men, undergo the fatigues of supporting it.
Yes, the American people should hear this, $300 million a day for two decades. If you take the number of $1 trillion, as many say, that's still $150 million a day for two decades. And what have we lost as a consequence in terms of opportunities? I refused to continue in a war that was no longer in the service of the vital national interest of our people.
I may not use the formulas every day, but there are skills that I gained that I apply on a daily basis, even if I don't recognize that 'this is Statistics
And before we judge of them too harshly we must remember what ruthless and utter destruction our own species has wrought, not only upon animals, such as the vanished bison and the dodo, but upon its inferior races. The Tasmanians, in spite of their human likeness, were entirely swept out of existence in a war of extermination waged by European immigrants, in the space of fifty years. Are we such apostles of mercy as to complain if the Martians warred in the same spirit?
If we don't end war, war will end us.
De Beren sijnt suz, sä de Voss.
6 Proposition. The first Trumpet or Viall began at the Jubelee, in anno Christi 71.
{The Russians} dashed on towards that thin line tipped with steel.
I am going on to the Rhine. If you oppose me, so much the worse for you, but whether you sign an armistice or not, I do not stop until I reach the Rhine.
[T]hat one should never permit a disorder to persist in order to avoid war, for war is not avoided thereby but merely deferred to one's own disadvantage..
But let this fact burn its way into your brain to save you from hell and rouse you for the revolution—this fact:Nowhere on all that battlefield among the shattered rifles and wrecked canon, among the broken ambulances and splintered ammunition wagons, nowhere in the mire and mush of blood and sand, nowhere among the bulging and befouling carcasses of dead horses and swelling corpses of dead men and boys—nowhere could be found the torn, bloated and fly-blown carcasses of bankers, bishops, politicians, "brainy capitalists" and other elegant and eminent "very best people."Well, hardly.Naturally—these proud, cunning and intelligent people were not there, on the firing line.Listen, oh, listen—you betrayed multitude of toil-damned, war-blasted workers of all nations:If the masters want blood, let them cut their own throats.We don't want other people's blood and we refuse to wast our own.Let those who want "great victories" go to the firing line and get them.If war is good enough to vote or to pray for, it is good enough to go to—up close where bayonets gleam, swords flash, canon roar, rifles clash, flesh rips, blood spurts, bones snap, brains are dashed,—up close where men toil, sweat, freeze, starve, kill, groan, scream, pray, laugh, howl, curse, go mad and die,—up close where the flesh and blood of betrayed men and boys are pounded into a red mush of mud by shrieking canon balls, by the iron-shod hoofs of galloping horses and the steel-bound wheels of rushing gun-trucks."What is war?"They say "War is Hell."Well, then, let those who want hell, go to hell.
He who first called money the sinews of the state seems to have said this with special reference to war.
Politics is the domestication of war.

Colouring layers based on concept drift

Here the idea was very simple: see what the hidden layers are emitting. The same LayerNorm and unembedding was used as for the final output (that's what the last three lines are checking). The colour shows the probability of "war" in the unembedding: >1% is red, >0.1% is yellow.



Backpropagation to identify weights relevant to topic

This was intended as a different technique to isolate parameters relevant to a particular topic.

- Process prompts until it suggests "war" with >1% probability
- Perform backpropagation (but don't update the weights, just see what the gradients are)
- See if any gradients are especially big and make a note of them
- Continue processing prompts

https://github.com/fractal-pterodactyl/concept_detector/blob/main/scan/microlearn_any.py
You need to specify the topic token and threshold on the command line, e.g.
python3 microlearn_any.py war 0.01

Causal tracing

Colab (copied from ROME paper and tweaked with our prompts)

Technique used in the ROME paper to determine the location of a fact
We used it on war-related prompts with the goal of finding out where in the model the concept of "war" is located (and if it is even localised at all). We found a few things:

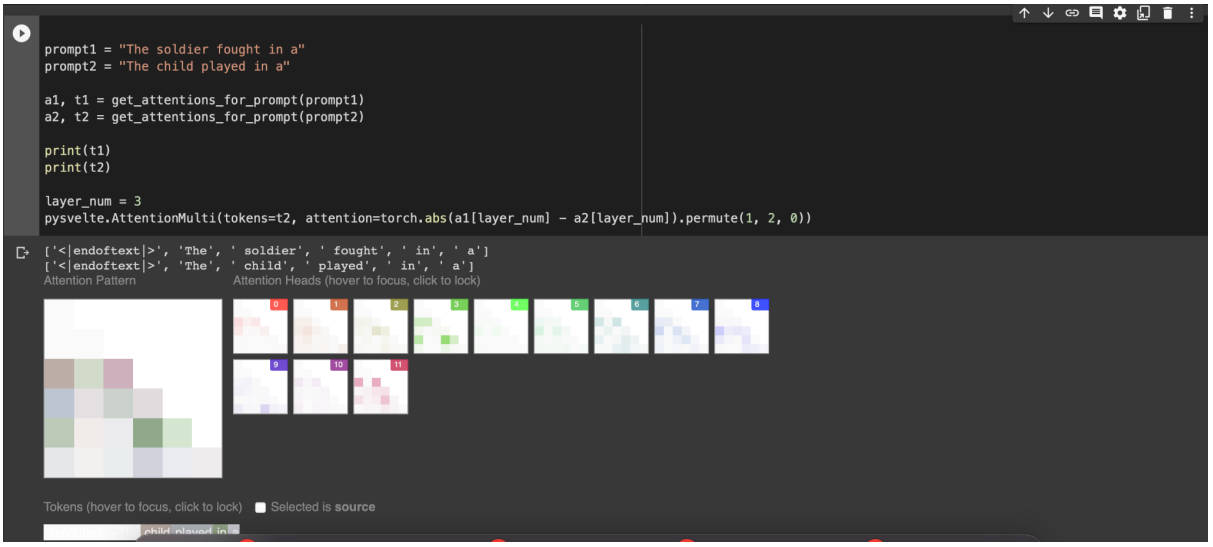
- It's hard to separate war as a concept from facts about wars. Ie "In 1914 the world went to **war"** - Completing this prompt just relies on knowledge of a simple historical fact. We need better prompts to more cleanly capture the concept. Related question - Does the model even have a concept of "war" separable from various facts about wars? Do humans? How could we test this? What exactly are concepts?
- War did not seem very localised
- Causal tracing takes a long time on the Colab free tier 🙁

Concept diffing

[Colab]

A much simpler approach that we came up with. We wanted to find the difference in attention maps between war and non-war prompts. We use 2 prompts which are structurally and grammatically the same, but with words changed so one prompt is talking about war and one is not. We then generate attention maps and diff them.

We just did this with a few prompt pairs and got some results that seem interesting, but more testing would be needed to see if we can do anything with this. The next step would be to create a lot more of these prompts and compare the diffs. We'd expect some randomness, but if there are a handful of places in the attention map that consistently show up on these diffs then maybe there is something war related there.



One caveat is that attention heads probably don't really store concepts. The ROME paper claims that facts are stored in the MLP layers. However, since attention layers also read from the residual stream, we think they may also pick up on information generated by the previous MLP layer.

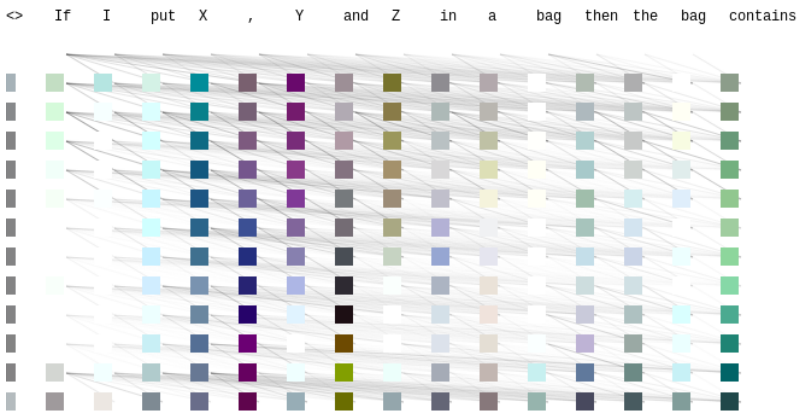
A lot of time was spent thinking about which prompts we should use. We mainly considered "war-like" and "non-war-like" prompts, but it might be interesting to investigate prompts on a single topic (like "war") and distinguish them as either "fact like" or "concept like". This may help us understand how a concept is treated by a neural network, and compare it to recent work like ROME.

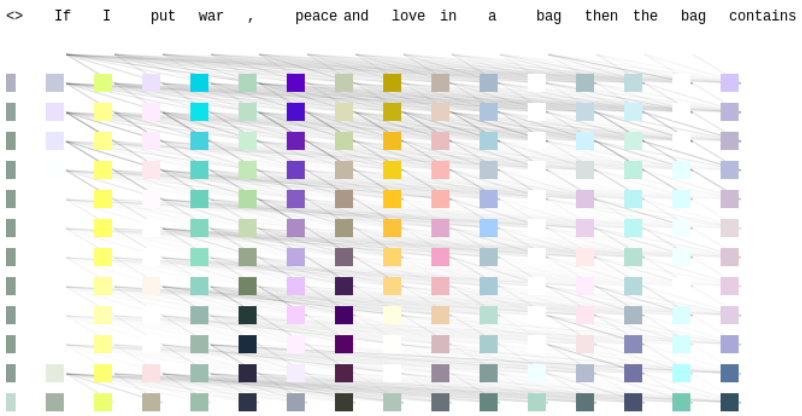
Information flow tracing

The idea behind this was to show how information propagates between tokens at each layer.

In these plots, the x axis corresponds to token position and the y axis corresponds to layer (with the first layer at the top). The colour of the squares shows the activations, unembedded, and then sampled at the position of key tokens. (red green and blue are “X”, “Y” and “Z” in the first image and “war”, “peace” and “love” in the second).

The lines connecting the squares show the sum of attention across all the relevant attention heads.





https://colab.research.google.com/drive/1zf7Uk3C4b774BGQKQXst32QLJbPCTilX?usp=sharing

Auspicious Basis

The idea here was to investigate the de-embedding matrix, and use it to infer structure of the embedding space.

In gpt2-small, the output of the MLP layers is a 768-entry vector, which can be thought of as a 768-dimensional vector space. This is passed to layernorm (which preserves dimensionality) and then to the unembedding matrix, which expands the dimension of the vector space to 50257, the number of tokens in the vocabulary.

So clearly each token doesn't get its own dedicated dimension. What then can we learn about how the values are organized in this reduced space?

For these purposes a vector is considered "auspicious" if its unembedding promotes one token significantly above all the rest. This can be tested by taking the softmax - one entry should end up close to 1 and the rest close to 0.

A basis is considered "auspicious" if it is made up of nearly-orthogonal auspicious vectors.

The first task was to see if an auspicious basis exists, and it turns out it does. This was discovered using one of Pytorch's optimizers (which might actually be overkill for this task, since there's no "data" that we're processing here, we're just trying to optimize our parameter matrix to satisfy two properties: near-orthogonality and the auspiciousness property of its component vectors.)

When printing out the softmaxed unembedding of the auspicious matrix we see an interesting property:

0.9993, ' bonded'
4.47e-05, ' bonding'
2.22e-05, ' bond'
5.48e-06, ' bonds'
1.72e-06, ' fused'

This is a fairly typical row, and we see the property that one entry is near 1 and the rest near 0, which is unsurprising as we were optimizing for that. We also see that the largest near-zero entries correspond to tokens that are very semantically similar to the main one. I don't know exactly why this is.

The next question then is: given an arbitrary auspicious vector (that's not in the basis), can we express it approximately as a linear combination of a small number of auspicious basis vectors, and if so are those vectors semantically related?

The answer is yes, and somewhat, respectively.

In an earlier version, the breakdown for " war" included a lot of garbage: 0.15 " conflic" but also 0.15 "Dragon magazine". The numbers also didn't tail off to zero as quickly as I expected. This was fixed by changing the vector norm in the optimizer from 1 to 0.8 (smaller values seem to break the optimizer).

Here is the breakdown for the " war" vector:
0.163 propag
0.138 conflic
0.127 Wars
0.102 strikes
0.079 unrest
0.056 dehuman
0.049 financial

And " peace":
0.117 unrest
0.081 enjoyment
0.072 lihood
0.060 financial
0.024 mutual

And " banana":
0.194 cone
0.097 Ghana
0.090 pudding
0.083 Paragu
0.074 snowball
0.070 chnology
0.050 reaction
0.047 frogs

And " science":
0.088 ⁄
0.076 gadgets
0.074 blending
0.052 athi
0.046 financial
0.036 promoting
0.003 mathemat

The words seem somewhat related in some cases, and not in others. (Remember that the vocabulary of the basis vectors is quite limited, so there might simply not be enough concepts available that are adjacent to e.g. a banana).

Note: these were found with another optimiser, not just by inverting the basis matrix. Inverting the basis results in a more or less even spread across the basis elements, without favouring the most meaningful ones.

https://colab.research.google.com/drive/1EYHJcfXbSbZH6GpS5DTE6mL6PPLASuNt?usp=sharing

Resources

- Easy Transformer Demo
- SERI MATS IOI Demo
- ROME
- Unpacking Large Models with Conceptual Consistency