Investigating Training Dynamics via Token Loss Trajectories

Alex Foote

AI Testing Hackathon Report

Apart Research

PIs: Esben Kran, Haydn Belfield, Fazl Barez

Date: 18th December, 2022

Abstract

Evaluations of ML systems typically focus on average statistical performance on a dataset measured at the end of training. However, this type of evaluation is relatively coarse, and not does not provide insight into the training dynamics of the model. We present tools for stratifying tokens into groups based on arbitrary functions and measuring the loss on these token groups throughout the training process of a Language Model. By evaluating the loss trajectory of meaningful groups of tokens throughout the training process, we can gain more insight into how the model develops during training, and make interesting observations that could be investigated further using interpretability tools to gain insight into the development of specific mechanisms within a model. We use this lens to look at the training dynamics of the region in which induction heads develop. We also zoom in on a specific region of training where there is a spike in loss and find that within this region the majority of tokens follow the loss trajectory of a spike, but a small set follow the inverse trajectory.

Investigating Training Dynamics via Token Loss Trajectories

Motivation and Methods

A significant recent finding in ML interpretability was the discovery of induction heads in Language Models [Olsson et al.], which play a major role in in-context learning. The paper found that these induction heads develop during a short window early on in training by looking at the loss of the model on the 500th token in a set of prompts minus the loss on the 50th token in a set of prompts. By comparing these two groups of tokens, they saw a region where there was a much greater improvement in the model's ability to predict the 500th token than the 50th token, which corresponded to the development of induction heads as identified by interpretability techniques.

This method of evaluation can be generalized to comparing the loss over training for two groups of tokens, and could be useful for identifying similar interesting regions of training to more deeply investigate, or for identifying broader training dynamics. We develop simple tools for visualizing the loss trajectories of two groups of tokens, given a pair of arbitrary functions which each select a group of tokens.

Firstly, we took 600 checkpoints of a GPT-2 small model provided by Mistral. We then took the WikiText-v2 dataset [Merity et al.], and computed the token-level loss of each model on the dataset, before saving this to disk.

The user can then define a pair of boolean functions which select tokens to include in each of the two groups. We provide some simple templates that can be used to quickly create common functions. For each of the 600 models, the functions are run over the dataset processed by the given model, and where a function selects a token by returning True, the loss of the model on that token is included in the running average loss. The average loss of the models on each group, as well as the difference between these loss trajectories, is then plotted.

We also provide a function for zooming in on a region of interest in the loss curve over all tokens, which plots a histogram of the per-token change in loss between two checkpoints, as well as the loss trajectories for random tokens and extreme tokens. This can be used to look for patterns in the tokens that are undergoing similar trajectories during a specific region of training.

Results

Firstly, we used this method to investigate differences between the 500th and 50th token, following Olsson et al. As shown in Figures 1 and 2, we similarly find that after the initial steep drop in loss during very early training, there is a region where the loss on the 500th token decreases significantly more than the loss on the 50th token. Beyond this point, loss on the 500th token increases slightly before staying essentially constant over the rest of training, whereas the loss on the 50th token gradually decreases. This corresponds to the finding that induction heads are the primary driver of in-context learning and develop early on in training.

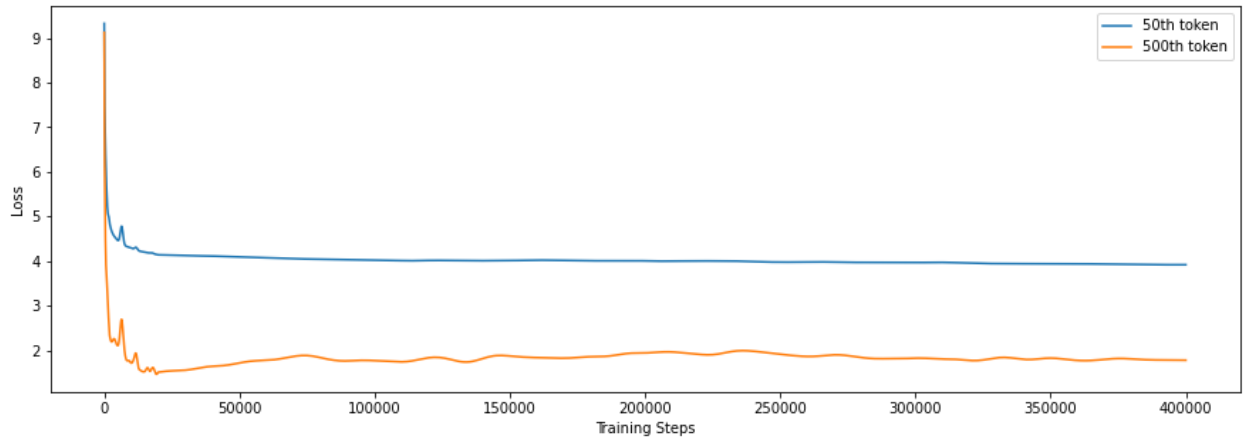

Figure 1

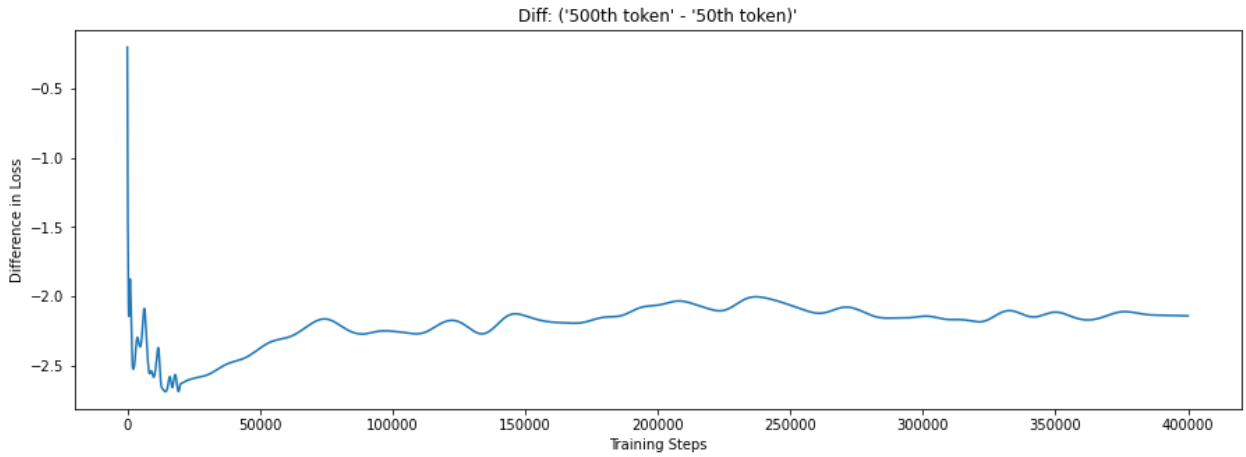

Figure 2: Difference in loss between the 500th token and the 50th token

We then look at specific groups of tokens to further investigate this behavior. We compare tokens that begin a word (single or multi part) with tokens that do not begin a word. For example, if the tokenizer splits " Sakimoto" into " Sak" and "imoto". In this case " Sak" would go into the first group (tokens that start a word) and "imoto" would go into the second group (tokens that do not start a word. Additionally, a single part word such as " the" would go into the first group as well. As shown in Figures 3 and 4, we find that tokens that do not start a word experience a greater improvement in loss than the start tokens, and after this region the loss on the non-start tokens stays relatively constant whilst the loss on the start tokens continues to fall. This behavior is very similar to that of the 500th vs 50th token, and fits with the hypothesis that induction heads are developing during this region - non-start tokens are commonly preceded by the same start token, so induction heads can look at previous instances of a start:non-start pair in the context and use this to increase the probability of the non-start token, given the same start token.

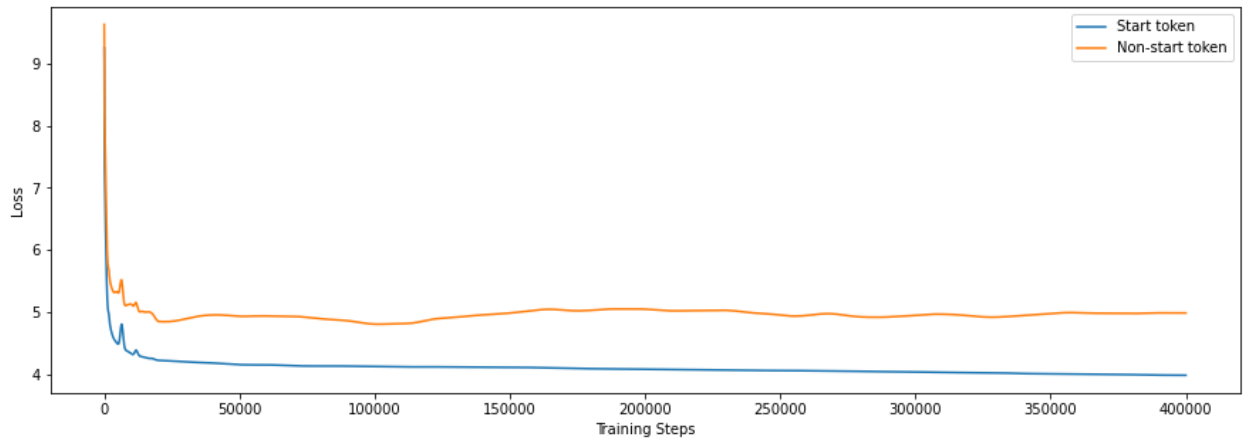

Figure 3

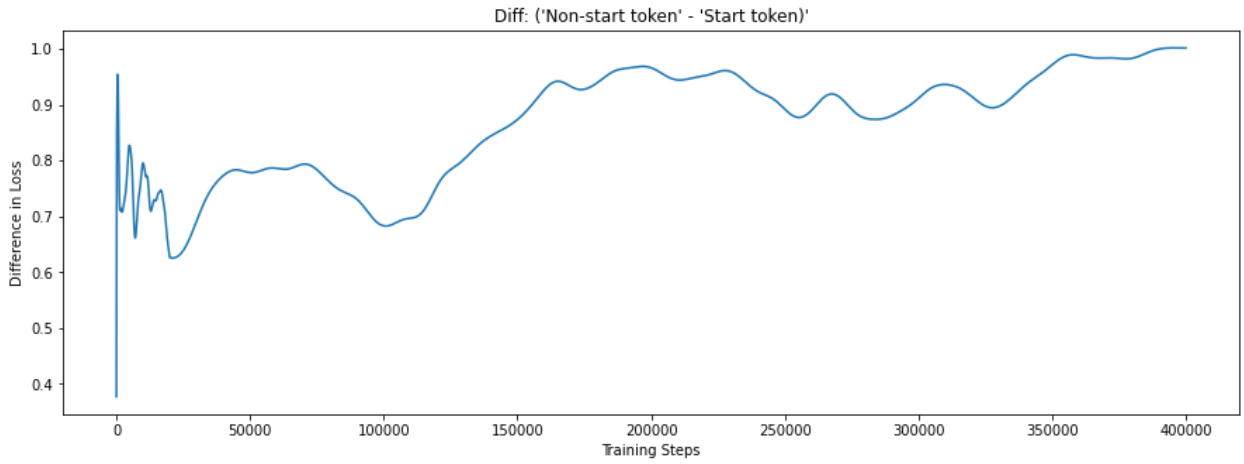

Figure 4

We then looked at other tokens which are likely to benefit from induction head, specifically the surnames of US presidents (specifically ["Obama", "Trump", "Biden", "Clinton", "Bush", "Reagan"]). These surnames are likely to be preceded by one of a small set of tokens, and a given pair is likely to occur multiple times in an input, so would benefit from the induction head mechanism. As shown in Figures 5 and 6, these tokens experience a sharp decrease in loss during the induction head region, but then show a different regime during the rest of training, where they undergo a gradual improvement in loss, potentially from the model slowly learning better n-gram statistics (or more complex contextual token statistics) , rather than a more general mechanism like the induction head. This shows how this token group trajectory lens can provide insight into different training dynamics.

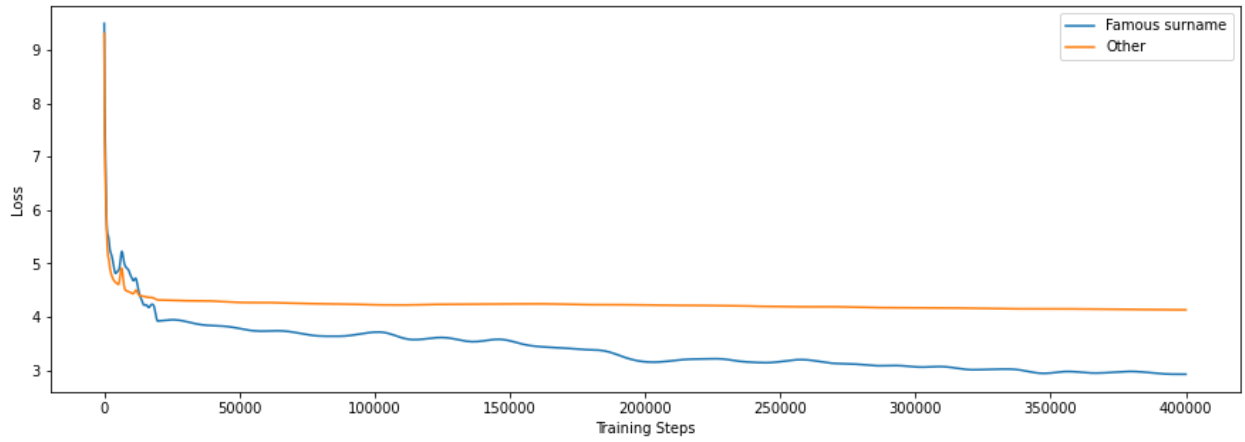

Figure 5

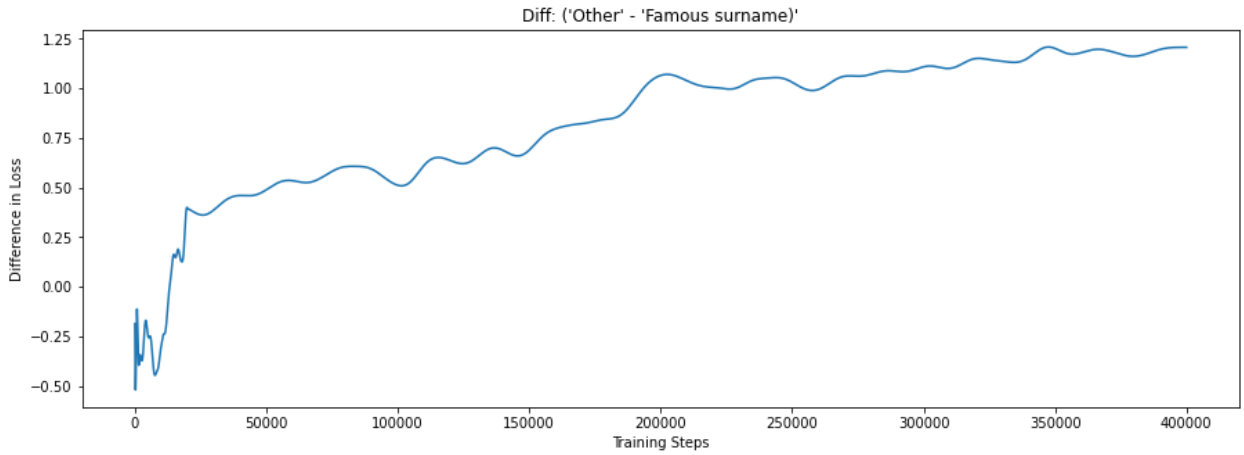

Figure 6

Finally, we zoom in on part of the induction head region, where there is a sharp spike in loss. As expected, Figure 7 shows that in this region most tokens undergo an increase in loss, however a small number experience a significant decrease. Comparing the trajectories of 5 random tokens vs the tokens that experienced the largest decrease in loss in Figure 8, we can see that at the same point, where the spike begins, the random token trajectories start to increase in loss, whereas the other tokens begin to decrease in loss. This suggests that the same underlying mechanism is driving these changes, rather than performance on some tokens randomly improving whilst most regress.

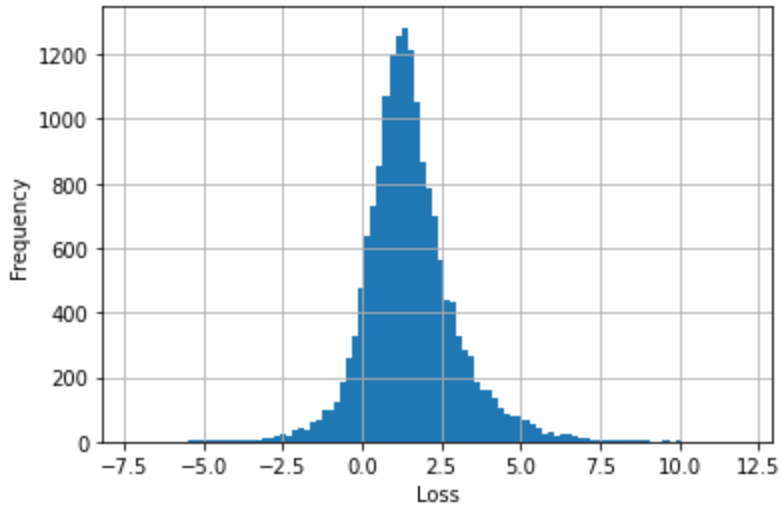

Figure 7

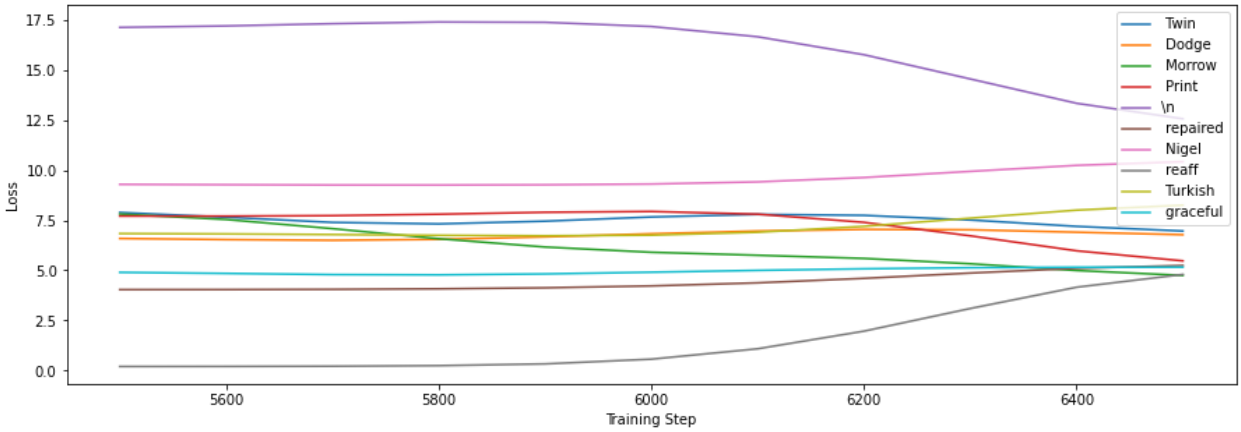

Figure 8

When we look at the other side of the spike where loss starts decreasing again, we see the inverse trend (Figures 9 and 10). Additionally, when we look at the top 50 tokens that decreased in loss during the start of the spike, and compare them to the top 50 tokens that increased in loss during the end of the spike, there is significant overlap between the groups. This provides further evidence that this spike is not a random artifact, and is instead the result of an underlying mechanism. This shows how the token trajectory view can provide insight into training regions that could be useful to investigate via interpretability tools.

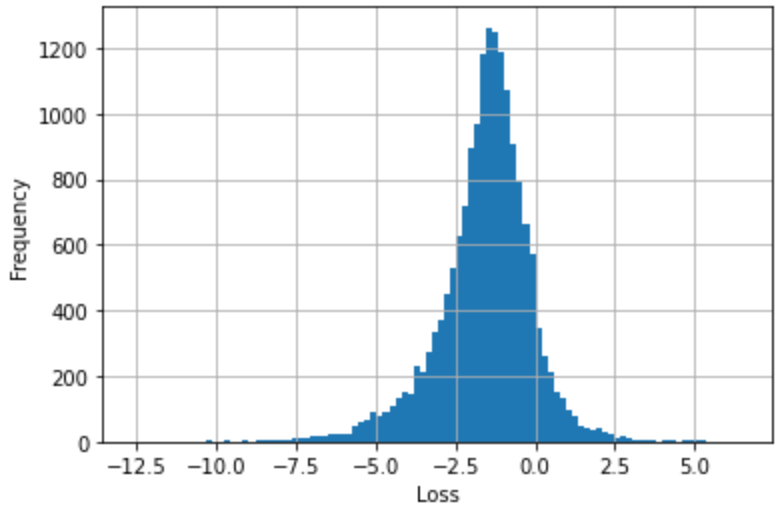

Figure 9

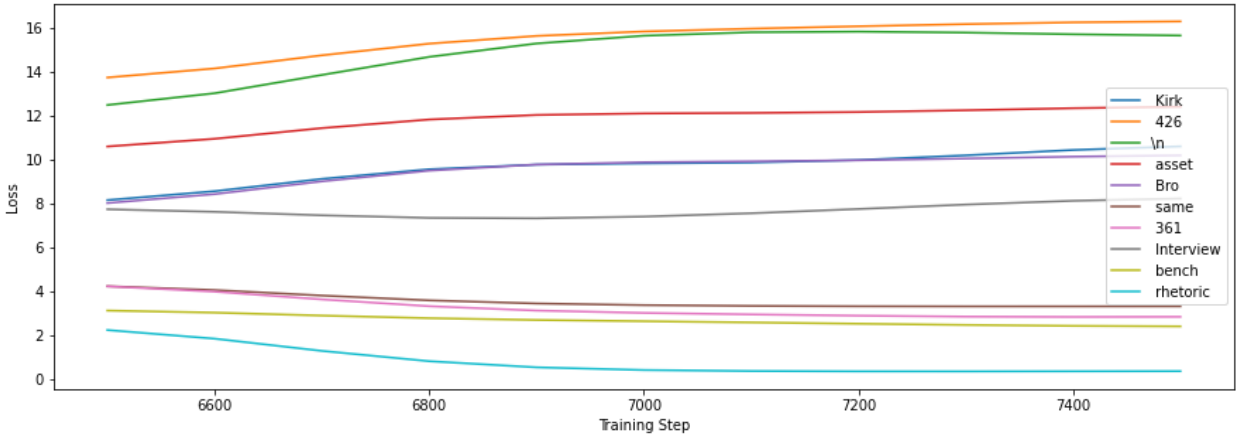

Figure 10

References

See links in text