Executive Summary

The PyTorch repository is facing various challenges, including performance regressions, compatibility issues, and security vulnerabilities. These issues are affecting different aspects of the project, such as tensor parallelism, autograd, and CUDA support. To address these challenges, the team needs to prioritize and tackle the root causes, which include missing optimized ops, incompatible backend combinations, and insufficient testing coverage.

Open Questions

* What is the root cause of the performance regression in AMP static shape default wrapper with multiple threads on CPU?

* How can we improve the naming and commenting scheme for triton fusion ops with custom triton kernel?

* What is the impact of the security vulnerabilities in the protobuf version used by PyTorch?

Key Challenges

Performance Regressions

Possible Causes
* Changes in the torch or torchvision libraries
* Changes in the model or benchmarking code
* Changes in the compiler or build flags

Remediations
* Investigate the suspected guilty commit and revert or fix the changes
* Verify that the model and benchmarking code are correct and up-to-date

Affected issues: **134679134686**

Compatibility Issues

Possible Causes
* Incompatible version of caffe2 with the current environment
* cuDNN version incompatibility between PyTorch and the runtime environment
* Incompatible NVIDIA H100 hardware with current PyTorch version

Remediations
* Try installing an older version of caffe2 that is compatible with the current environment
* Ensure PyTorch can find the bundled cuDNN by removing incompatible versions from the LD_LIBRARY_PATH
* Update PyTorch to a version compatible with NVIDIA H100

Affected issues: **134640134682134684**

Security Vulnerabilities

Possible Causes
* Using the affected version of protobuf (3.20.2)
* Processing maliciously crafted messages during data serialization/deserialization
* Using ONNX with the affected version of protobuf

Remediations
* Update protobuf to a version mentioned in the CVE reports (e.g., 3.21.7, 3.20.3, 3.19.6, or 3.16.3)
* Avoid using ONNX with the affected version of protobuf

Affected issues: **134664**

Tensor Parallelism and Autograd

Possible Causes
* Lack of understanding of tensor parallelism in PyTorch
* Incorrect handling of dynamic shapes in compiled autograd

Remediations
* Improve documentation on tensor parallelism and overlapping communication calculations
* Update compiled autograd to handle dynamic shapes correctly

Affected issues: **134668134676**

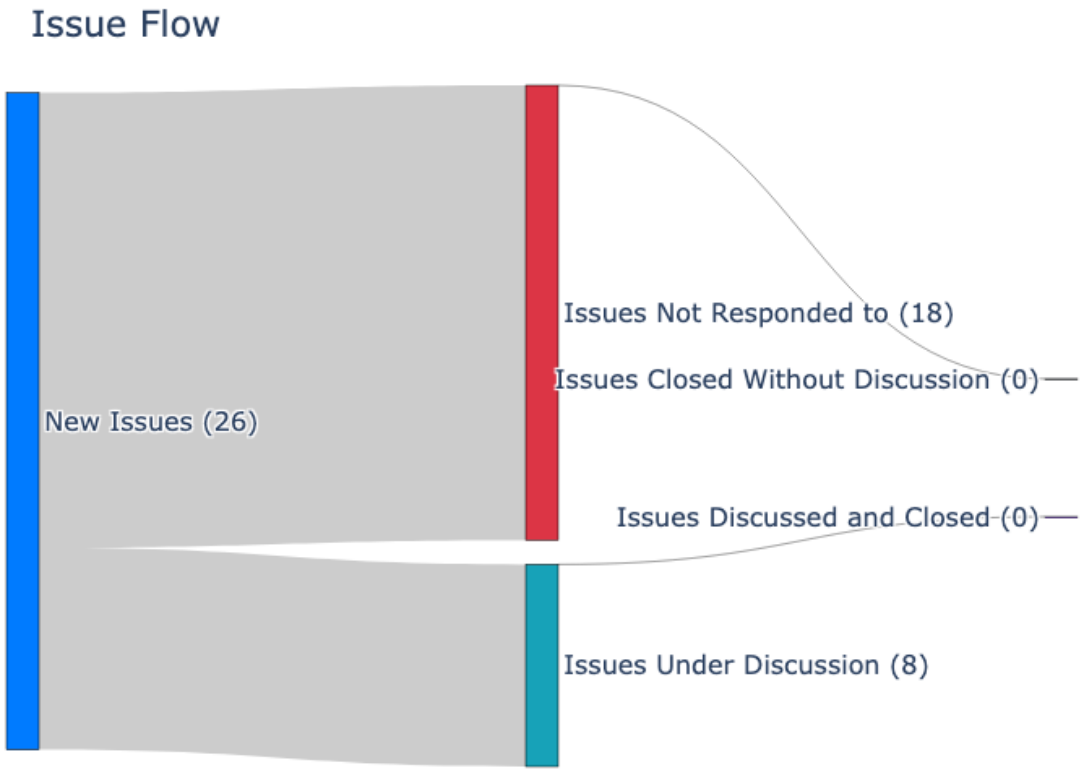
CUDA Support

Possible Causes
* cuDNN version incompatibility between PyTorch and the runtime environment
* Incompatible NVIDIA H100 hardware with current PyTorch version

Remediations
* Ensure PyTorch can find the bundled cuDNN by removing incompatible versions from the LD_LIBRARY_PATH
* Update PyTorch to a version compatible with NVIDIA H100

Affected issues: **134682134684**

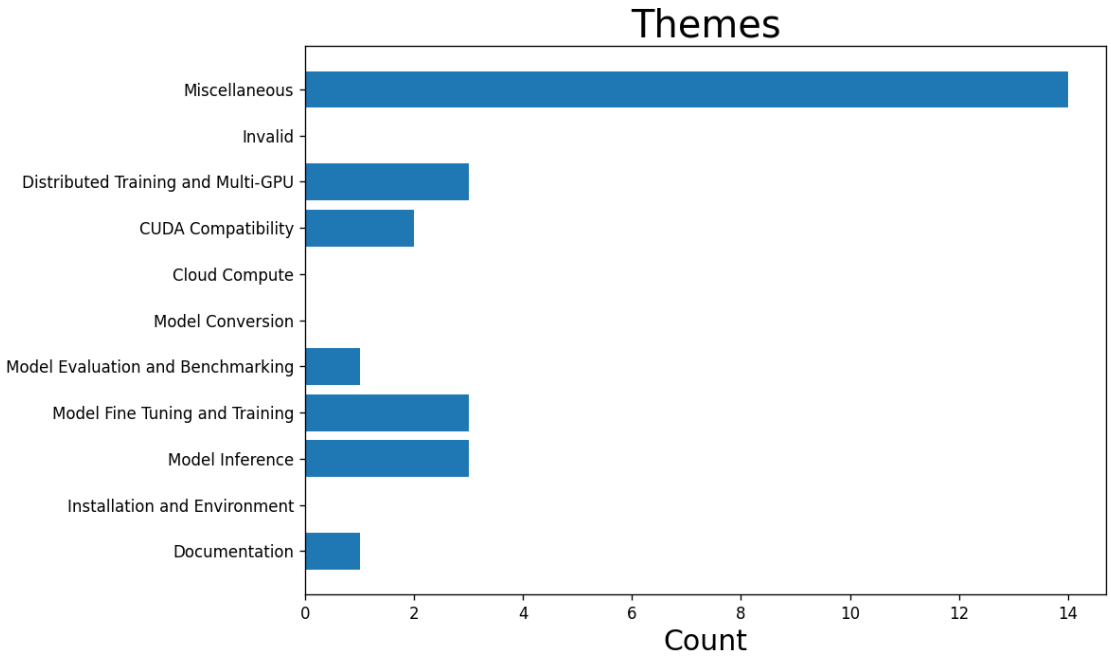
[Viz] Repo Maintenance

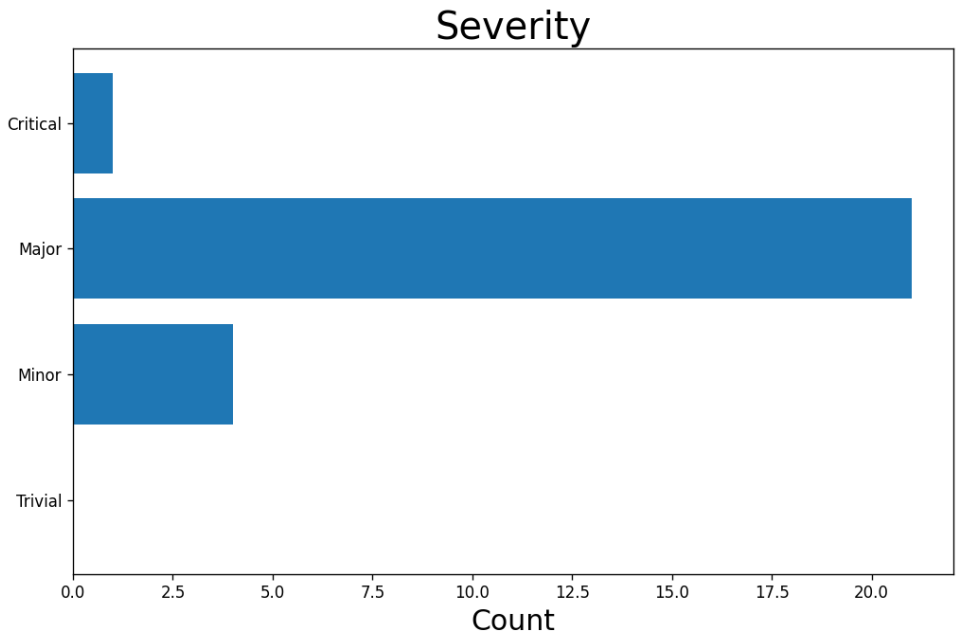


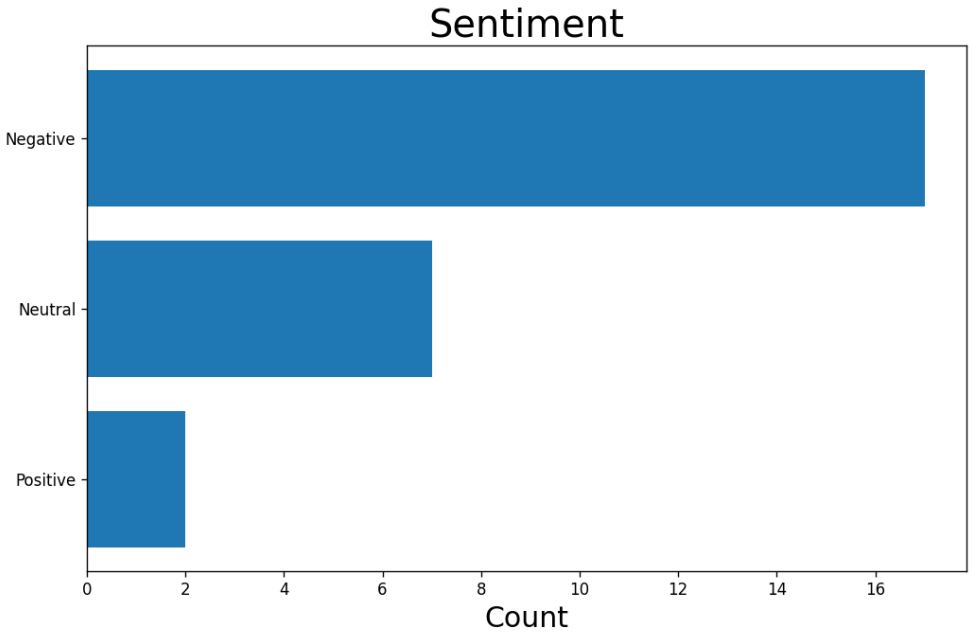
[Viz] Traffic in the last 2 weeks

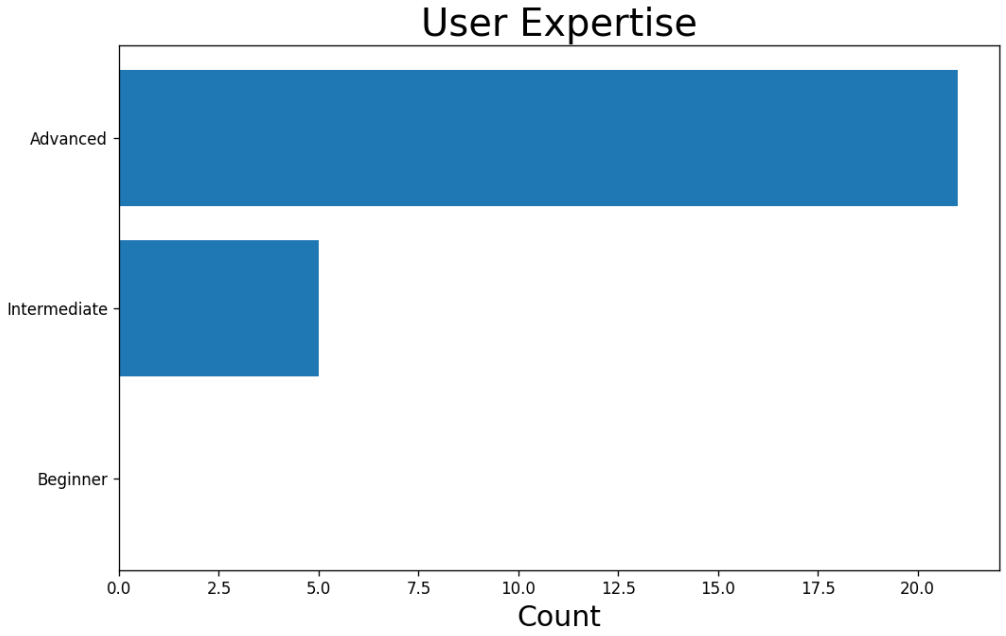
< Plot not found, make sure you have push-acces to this repo > ***< Plot***

[Viz] New issues in the last 2 weeks






Sentiment
Negative
Neutral
Positive
0
2
4
6
8
10
12
14
16
Count


User Expertise
Advanced
Intermediate
Beginner
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Count