UI screenshot classifier

by Yurii Voievidka
github repo: https://github.com/yvoievid/cv-project#ui-screenshot-classifier
data: https://huggingface.co/datasets/yuriivoievidka/classified-ui

Introduction

The world of UI design is vast and diverse, especially when it comes to desktop interfaces. Each UI serves a distinct purpose, and having a well-structured dataset of UI images is crucial for analysis and model training. Manually capturing screenshots of various UIs is not only tedious but also inefficient. Fortunately, an automated approach can be employed—extracting UI images directly from App Store data manifests of different applications.

However, there is a challenge: **not all screenshots contain valuable information**. Additionally, many of these images are from mobile applications, which are irrelevant to our study.

Objective

This repository aims to develop and evaluate methods for **classifying Desktop UI images** into three categories:

- **clean-ui** – Images that accurately and clearly represent the UI without unnecessary elements.
- **ui-to-crop** – Images that include redundant information, such as advertisements, macOS desktop backgrounds, or toolbars, which need to be cropped.
- **unnecessary** – Images that are irrelevant, such as mobile UI screenshots, text-heavy images without UI elements, or visuals that cannot be cleaned for meaningful use.

Approach

High-resolution image classification presents unique challenges, and this report explores multiple techniques to address them. For this purpose, a manually labeled dataset of **4,000 images** has been created, and the trained models will be evaluated on a larger dataset of **20,000 UI images**.

The classification experiments will be conducted using **four different approaches**:

1. **Vision Transformer (ViT) Fine-Tuning**
2. **CLIP Encoder + XGBoost**
3. **BLIP2 Encoder + XGBoost**
4. **SigLIP2 Encoder + XGBoost**

By comparing these methods, we aim to determine the most effective and efficient approach for classifying desktop UI images.

Hypothesis

I hypothesize that **SigLIP2** has the most advanced encoder architecture, which will lead to the highest performance metrics among the tested approaches. Given its ability to capture complex visual patterns and semantic information effectively, I expect it to outperform other models in classification accuracy.

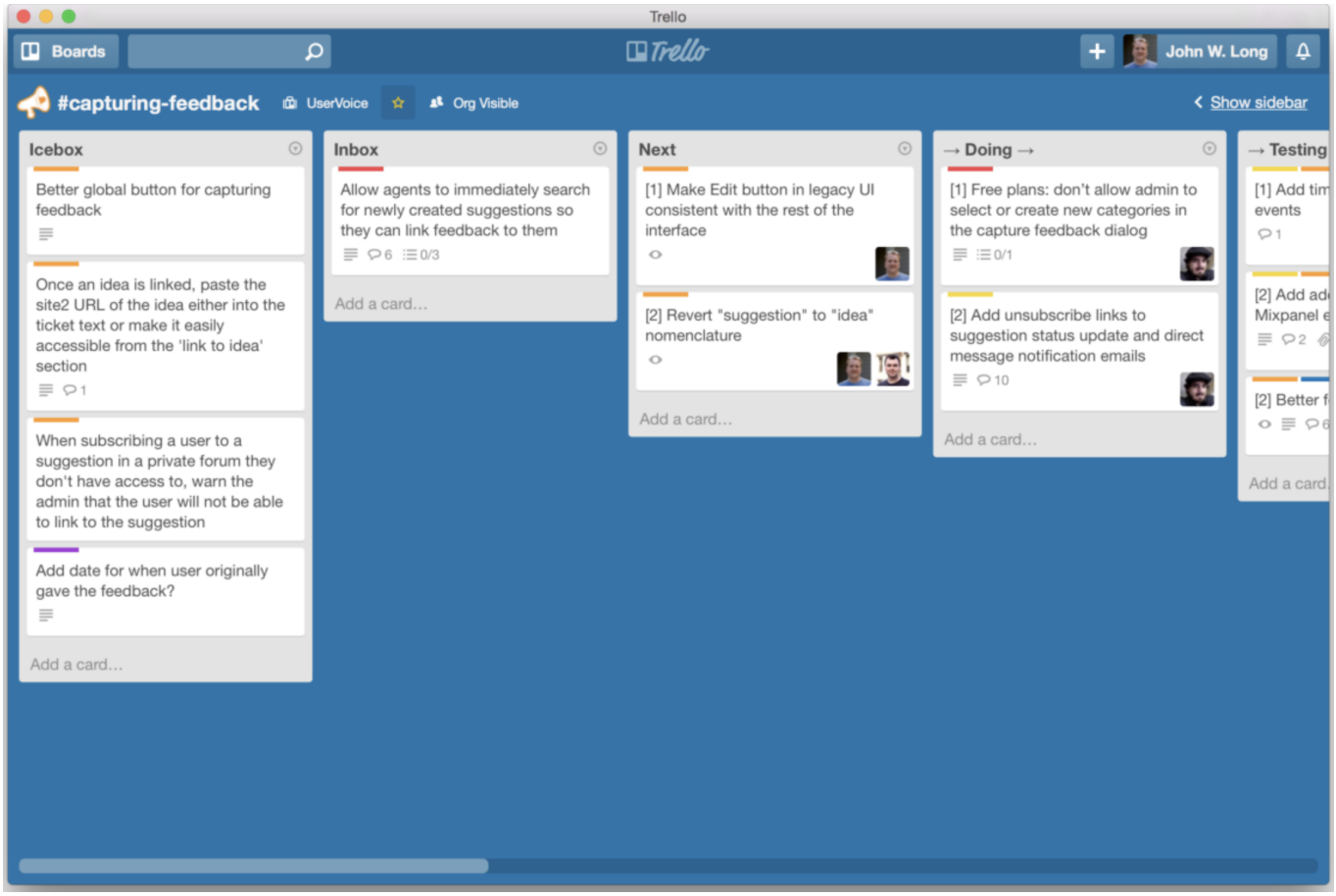
To evaluate model performance, the following metrics will be used:

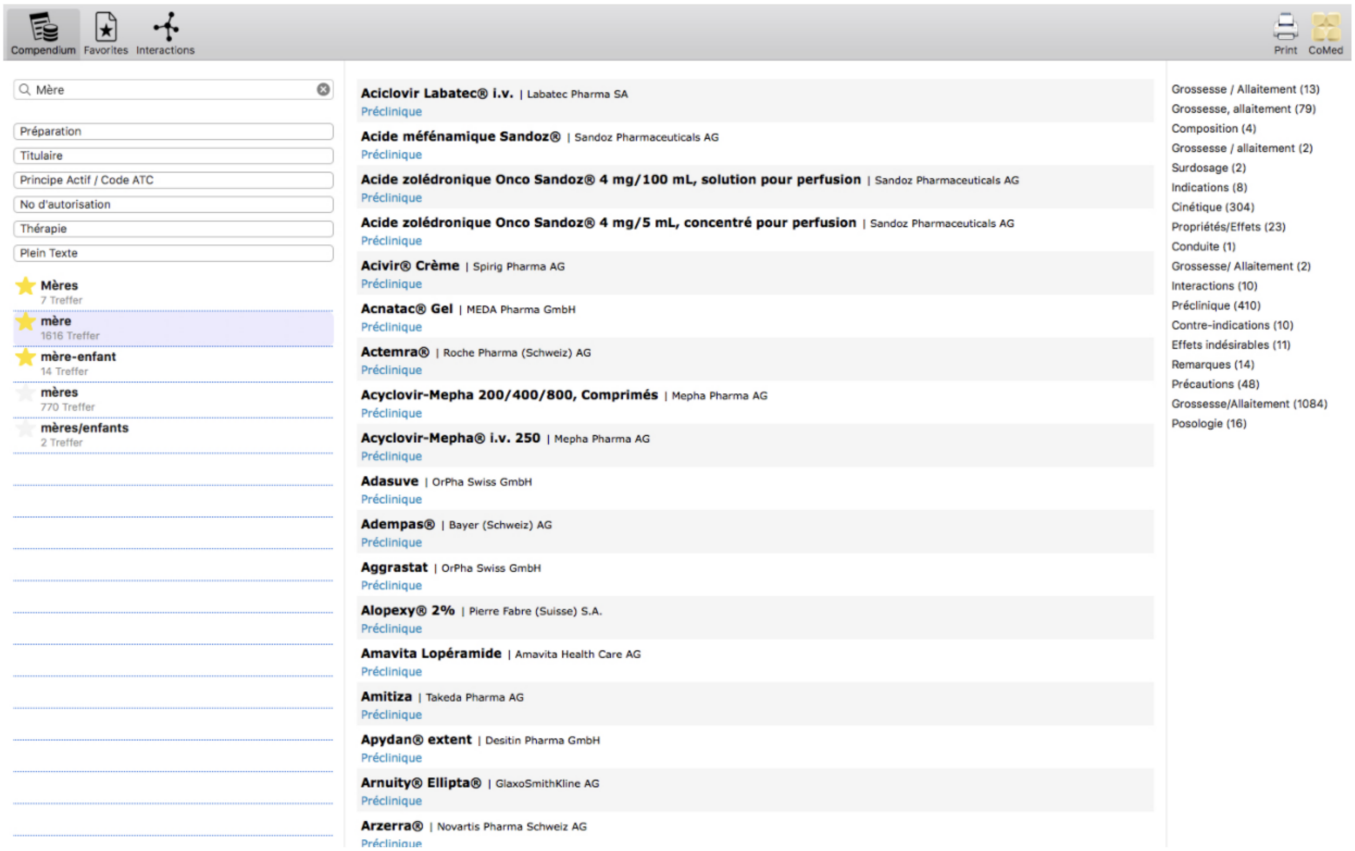
- **Accuracy** – Measures overall correctness of predictions.
- **Precision** – Assesses how many of the predicted positive instances are actually correct.
- **Recall** – Determines how well the model identifies all relevant instances.
- **F1 Score** – Balances precision and recall, providing a comprehensive assessment of classification performance.

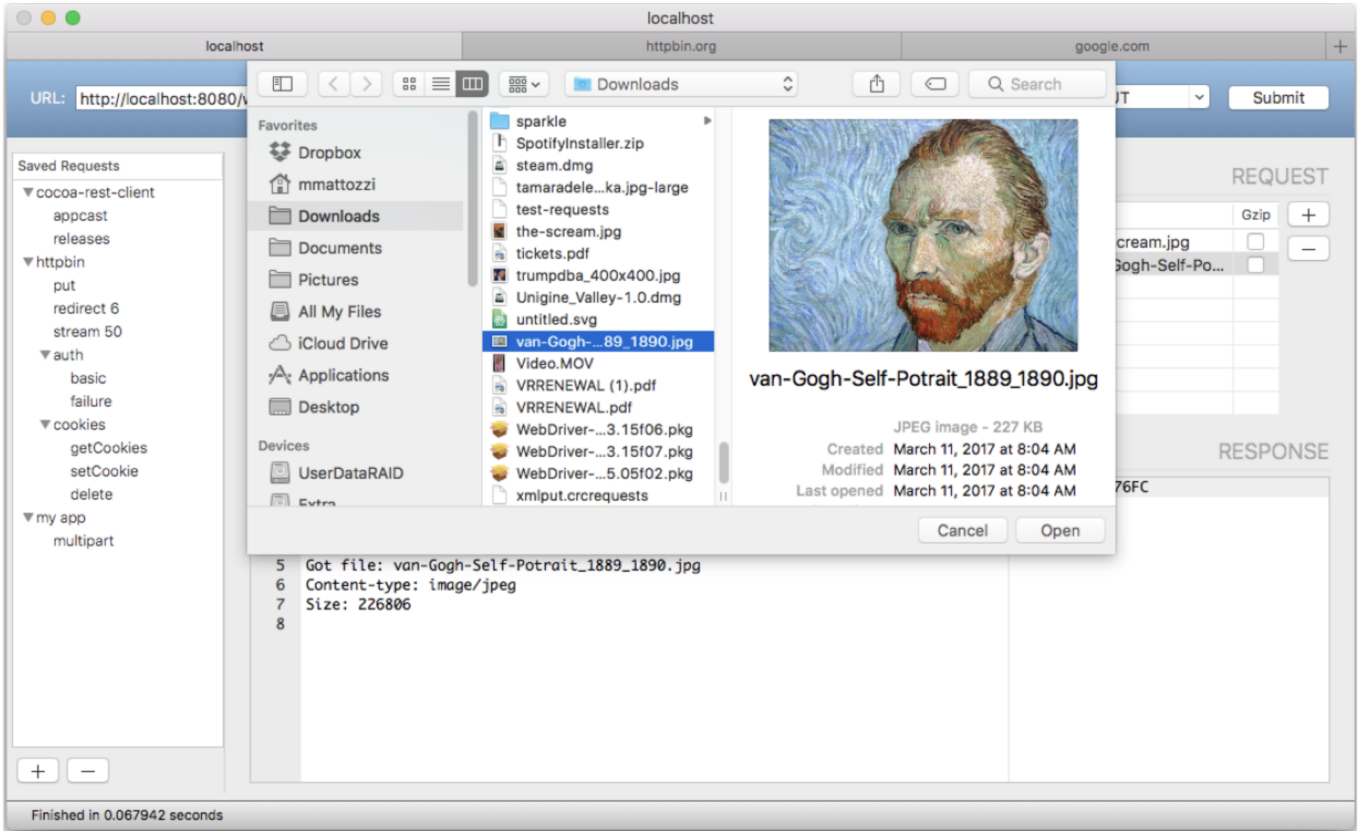
Through this experiment, I aim to validate whether **SigLIP2 truly provides superior classification performance** compared to other encoding methods.

Desktop UI images examples

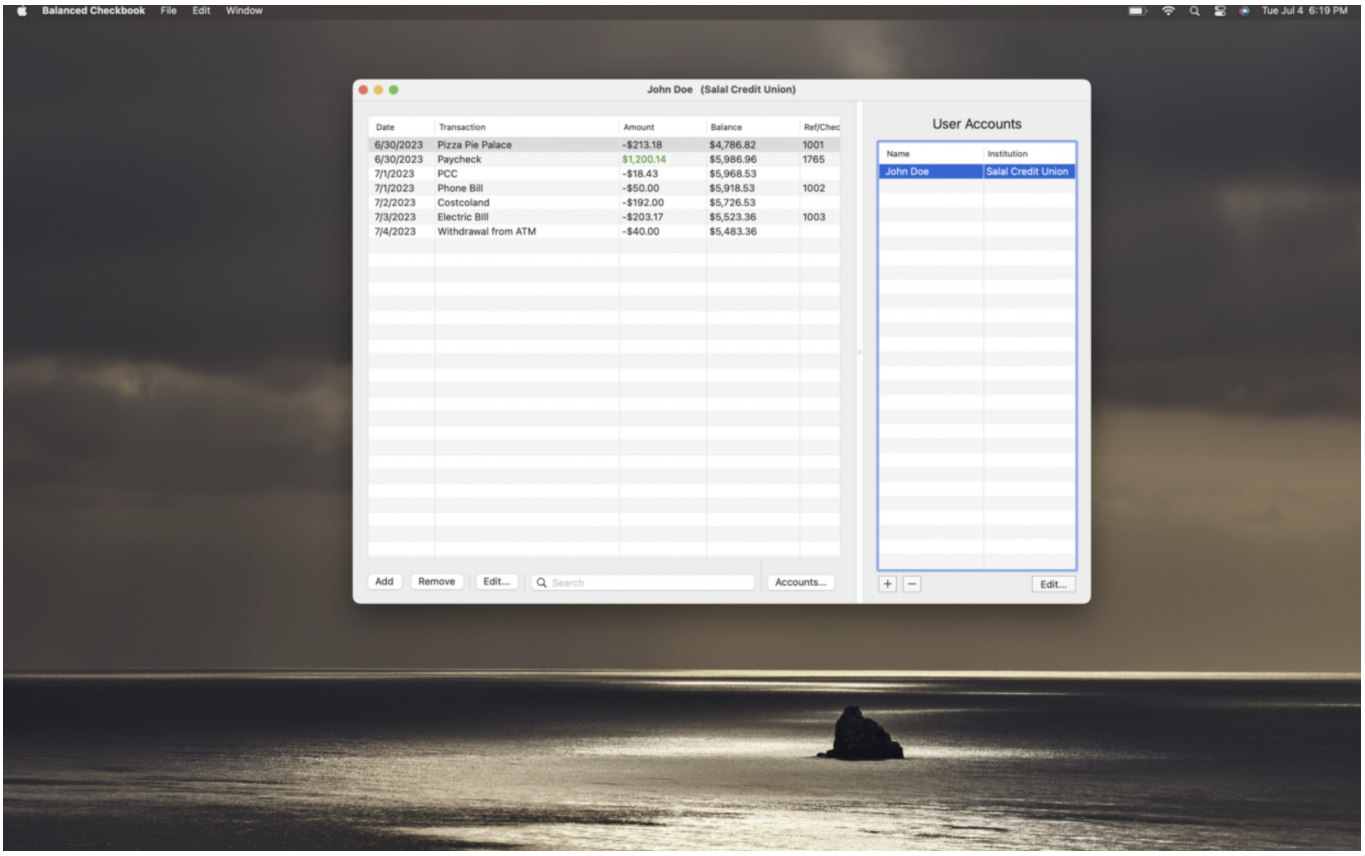
Clean UI :


Trello
Boards
Trello
John W. Long
#capturing-feedback
UserVoice
Org Visible
Show sidebar
Icebox
Better global button for capturing feedback
Once an idea is linked, paste the site2 URL of the idea either into the ticket text or make it easily accessible from the 'link to idea' section
When subscribing a user to a suggestion in a private forum they don't have access to, warn the admin that the user will not be able to link to the suggestion
Add date for when user originally gave the feedback?
Add a card...
Inbox
Allow agents to immediately search for newly created suggestions so they can link feedback to them
Add a card...
Next
[1] Make Edit button in legacy UI consistent with the rest of the interface
[2] Revert "suggestion" to "idea" nomenclature
Add a card...
→ Doing →
[1] Free plans: don't allow admin to select or create new categories in the capture feedback dialog
[2] Add unsubscribe links to suggestion status update and direct message notification emails
Add a card...
→ Testing
Add a card...


Compendium
Favorites
Interactions
Print
CoMed
Mère
Préparation
Titulaire
Principe Actif / Code ATC
No d'autorisation
Thérapie
Plein Texte
Mères
7 Treffer
mère
1616 Treffer
mère-enfant
14 Treffer
mères
770 Treffer
mères/enfants
2 Treffer
Aciclovir Labatec® i.v. | Labatec Pharma SA
Préclinique
Acide méfénamique Sandoz® | Sandoz Pharmaceuticals AG
Préclinique
Acide zolédronique Onco Sandoz® 4 mg/100 mL, solution pour perfusion | Sandoz Pharmaceuticals AG
Préclinique
Acide zolédronique Onco Sandoz® 4 mg/5 mL, concentré pour perfusion | Sandoz Pharmaceuticals AG
Préclinique
Acivir® Crème | Spirig Pharma AG
Préclinique
Acnatac® Gel | MEDA Pharma GmbH
Préclinique
Actemra® | Roche Pharma (Schweiz) AG
Préclinique
Acyclovir-Mepha 200/400/800, Comprimés | Mepha Pharma AG
Préclinique
Acyclovir-Mepha® i.v. 250 | Mepha Pharma AG
Préclinique
Adasuve | OrPha Swiss GmbH
Préclinique
Adempas® | Bayer (Schweiz) AG
Préclinique
Aggrastat | OrPha Swiss GmbH
Préclinique
Alopexy® 2% | Pierre Fabre (Suisse) S.A.
Préclinique
Amavita Lopéramide | Amavita Health Care AG
Préclinique
Amitiza | Takeda Pharma AG
Préclinique
Apydan® extent | Desitin Pharma GmbH
Préclinique
Arnuity® Ellipta® | GlaxoSmithKline AG
Préclinique
Arzerra® | Novartis Pharma Schweiz AG
Préclinique
Grossesse / Allaitement (13)
Grossesse, allaitement (79)
Composition (4)
Grossesse / allaitement (2)
Surdosage (2)
Indications (8)
Cinétique (304)
Propriétés/Effets (23)
Conduite (1)
Grossesse/ Allaitement (2)
Interactions (10)
Préclinique (410)
Contre-indications (10)
Effets indésirables (11)
Remarques (14)
Précautions (48)
Grossesse/Allaitement (1084)
Posologie (16)


localhost
localhost
httpbin.org
google.com
URL: http://localhost:8080/
Submit
Saved Requests
cocoa-rest-client
appcast
releases
httpbin
put
redirect 6
stream 50
auth
basic
failure
cookies
getCookies
setCookie
delete
my app
multipart
REQUEST
Gzip
RESPONSE
Downloads
Search
Favorites
Dropbox
mmattozzi
Downloads
Documents
Pictures
All My Files
iCloud Drive
Applications
Desktop
Devices
UserDataRAID
sparkle
SpotifyInstaller.zip
steam.dmg
tamaradele...ka.jpg-large
test-requests
the-scream.jpg
tickets.pdf
trumpdba_400x400.jpg
Unigine_Valley-1.0.dmg
untitled.svg
van-Gogh-...89_1890.jpg
Video.MOV
VRRENEWAL (1).pdf
VRRENEWAL.pdf
WebDriver-...3.15f06.pkg
WebDriver-...3.15f07.pkg
WebDriver-...5.05f02.pkg
xmlput.crcrequests
van-Gogh-Self-Potrait_1889_1890.jpg
JPEG image - 227 KB
Created March 11, 2017 at 8:04 AM
Modified March 11, 2017 at 8:04 AM
Last opened March 11, 2017 at 8:04 AM
Cancel
Open
5 Got file: van-Gogh-Self-Potrait_1889_1890.jpg
6 Content-type: image/jpeg
7 Size: 226806
8
Finished in 0.067942 seconds

UI To Crop:


Balanced Checkbook File Edit Window
Tue Jul 4 6:19 PM
John Doe (Salal Credit Union)
Date Transaction Amount Balance Ref/Chec
6/30/2023 Pizza Pie Palace -$213.18 $4,786.82 1001
6/30/2023 Paycheck $1,200.14 $5,986.96 1765
7/1/2023 PCC -$18.43 $5,968.53
7/1/2023 Phone Bill -$50.00 $5,918.53 1002
7/2/2023 Costcoland -$192.00 $5,726.53
7/3/2023 Electric Bill -$203.17 $5,523.36 1003
7/4/2023 Withdrawal from ATM -$40.00 $5,483.36
Add Remove Edit... Search Accounts...
User Accounts
Name Institution
John Doe Salal Credit Union
Edit...


软著代码文档生成工具
软著
软件名称 软著代码文档生成器
版本号 1.0.0
著作人名称 xcs
源码目录
目录 操作
/Users/xcs/project/code/app
添加源码目录
源码后缀名 .js
首尾行数增减行 0
输出位置 /Users/xcs/Downloads 选择
生成文档
MacBook Pro

AVG

IMPROVED!

Clean up
hidden cache, trash & junk

613 MB
IN OUR TESTS

- Application caches
- iPhoto Library Trash
- Trash
- Mail Downloads
- Junk files
- Downloads

Unnecessary:

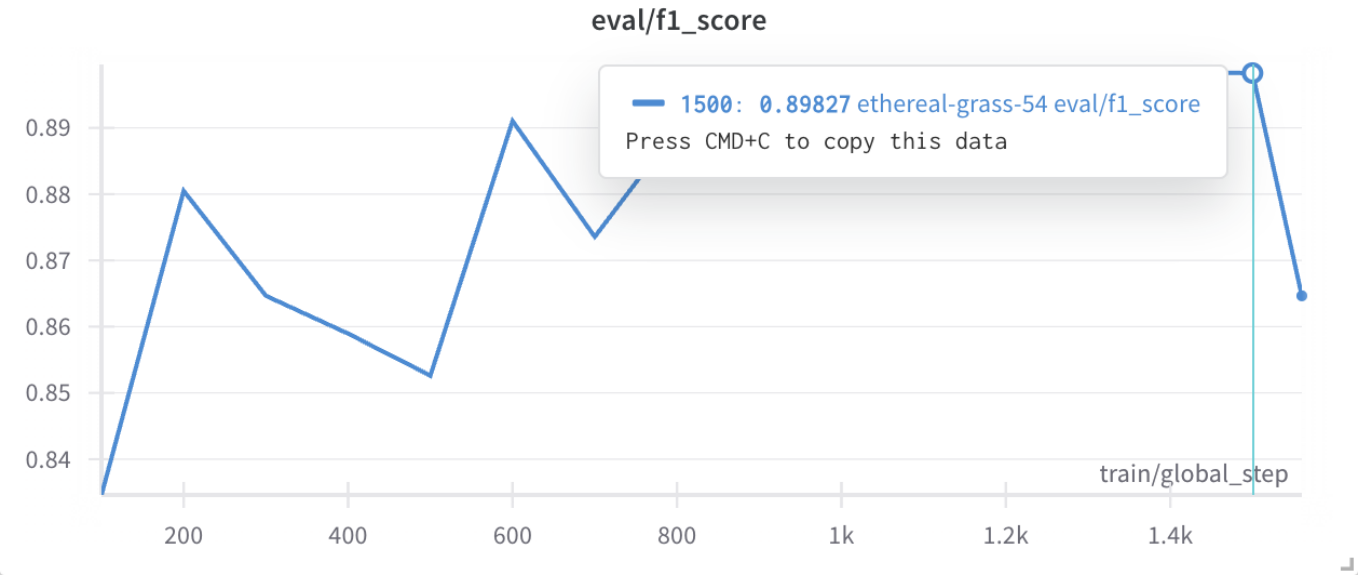
Visual Transformer Fine Tuning

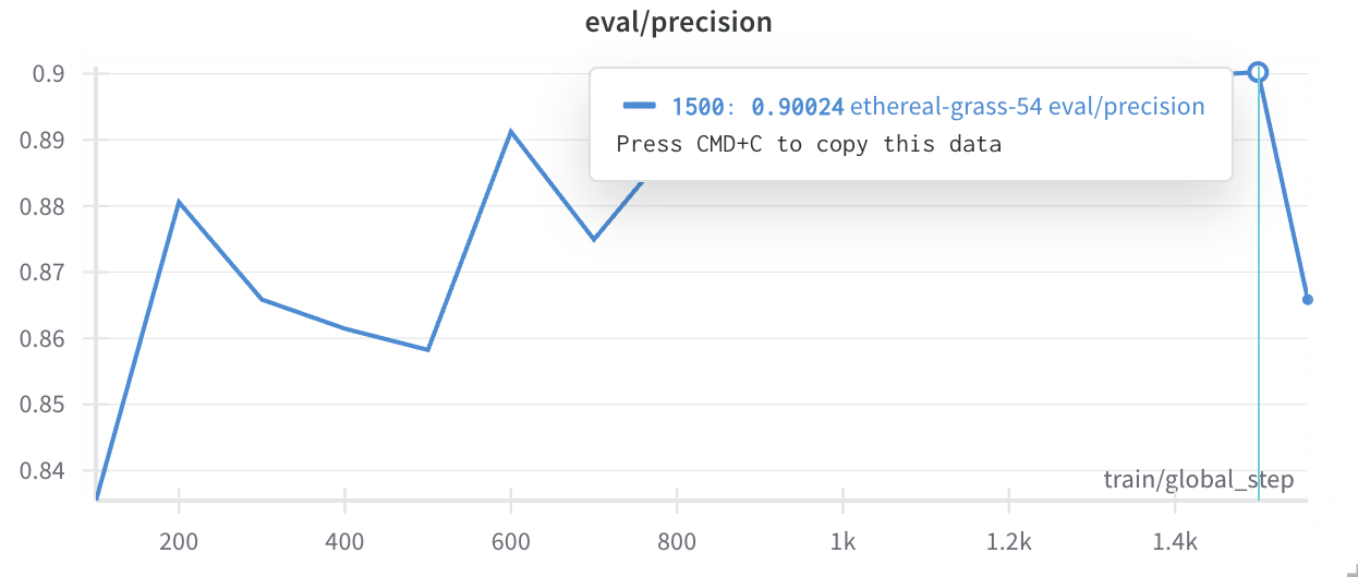
Start Date	2025-02-10
End Date	2025-02-11
dataset	DesktopUI

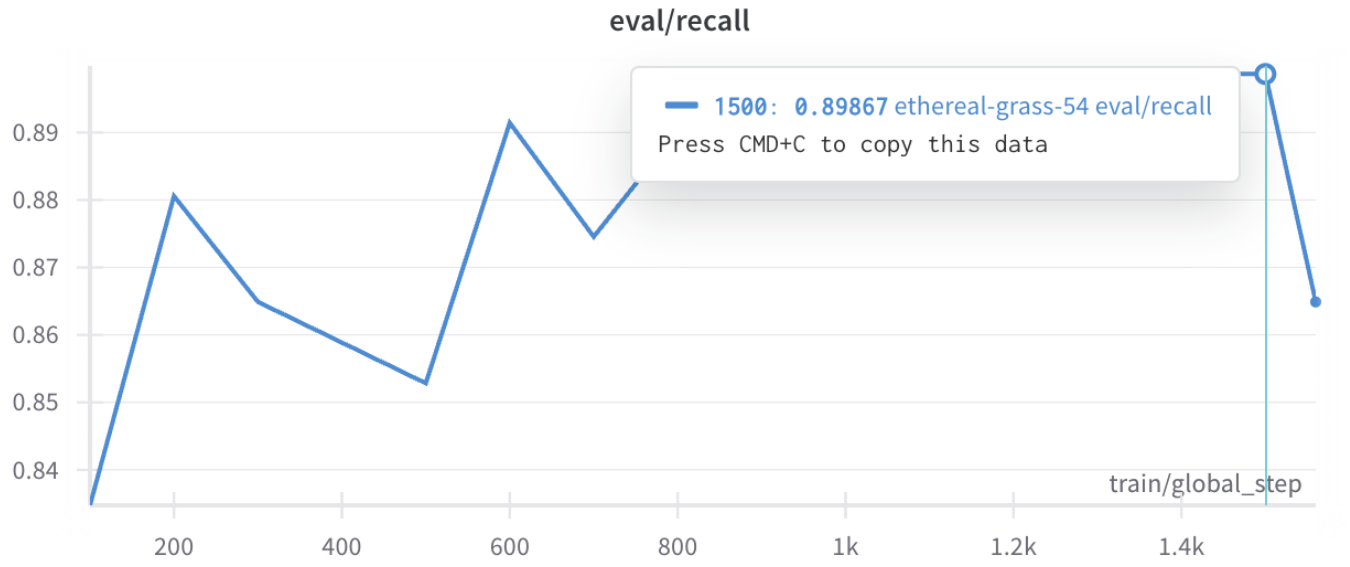
Motivation

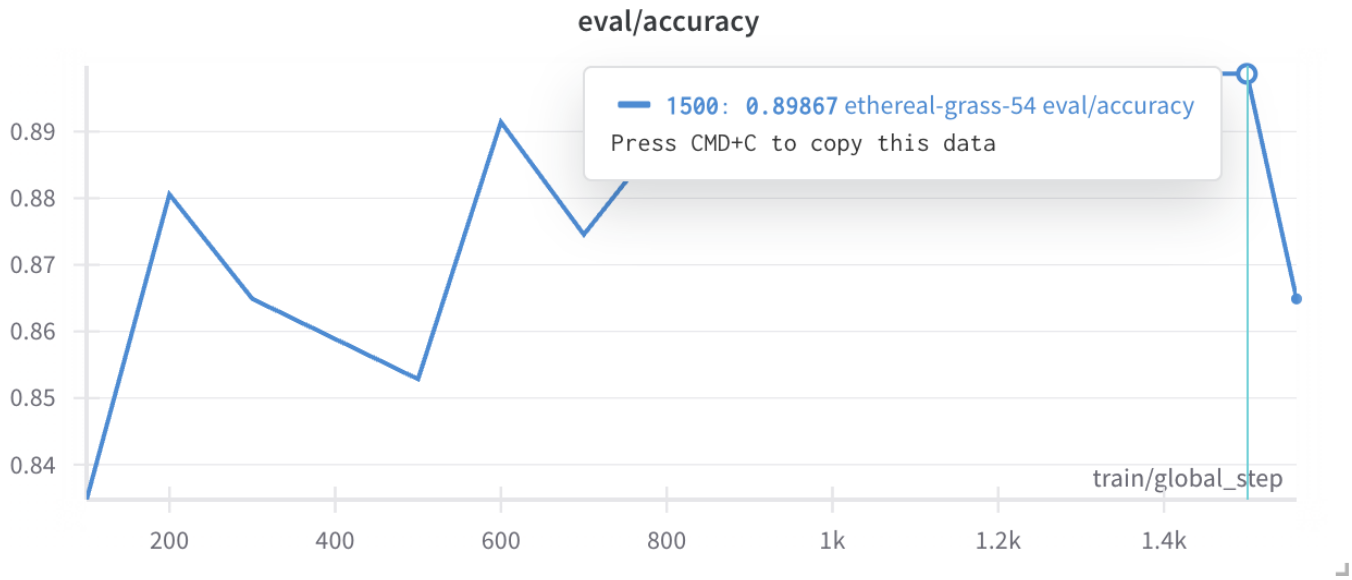
Try to fine tune Visual Transformer on custom dataset of desktop UI images.

Results


eval/f1_score
1500: 0.89827 ethereal-grass-54 eval/f1_score
Press CMD+C to copy this data
0.89
0.88
0.87
0.86
0.85
0.84
train/global_step
200
400
600
800
1k
1.2k
1.4k


eval/precision
1500: 0.90024 ethereal-grass-54 eval/precision
Press CMD+C to copy this data
0.9
0.89
0.88
0.87
0.86
0.85
0.84
train/global_step
200
400
600
800
1k
1.2k
1.4k

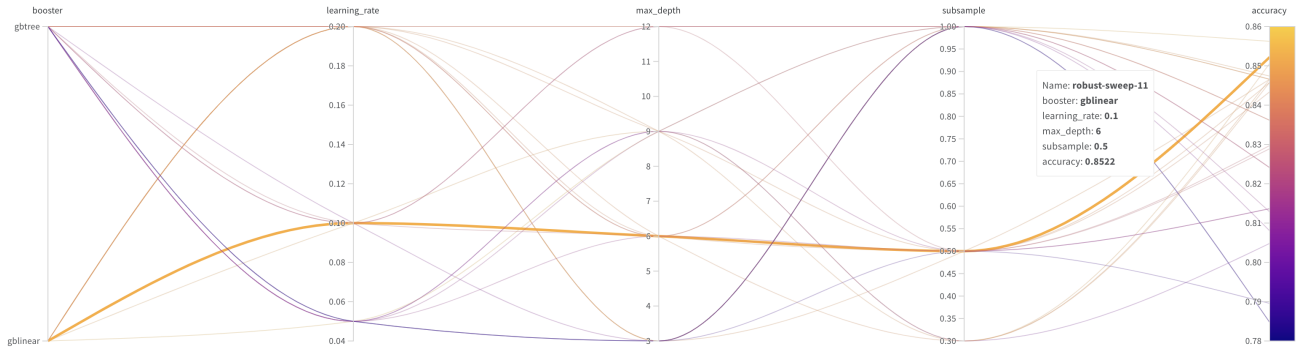


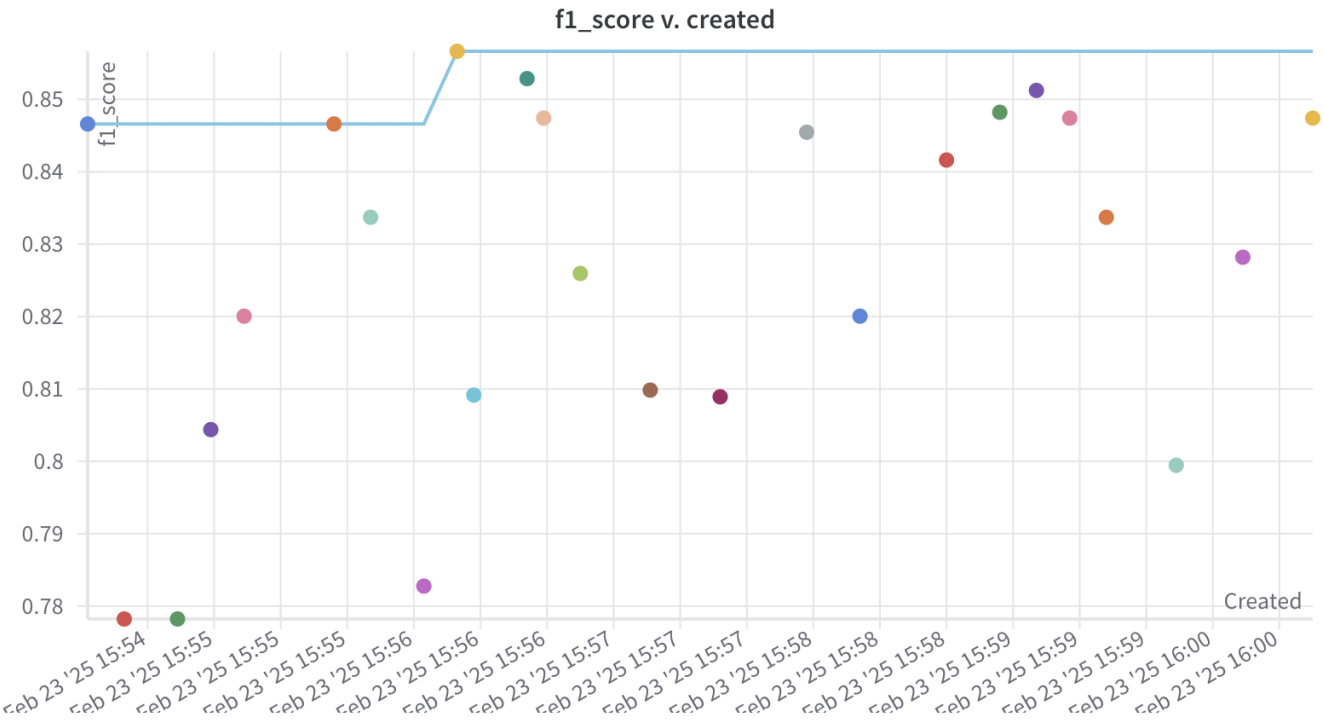


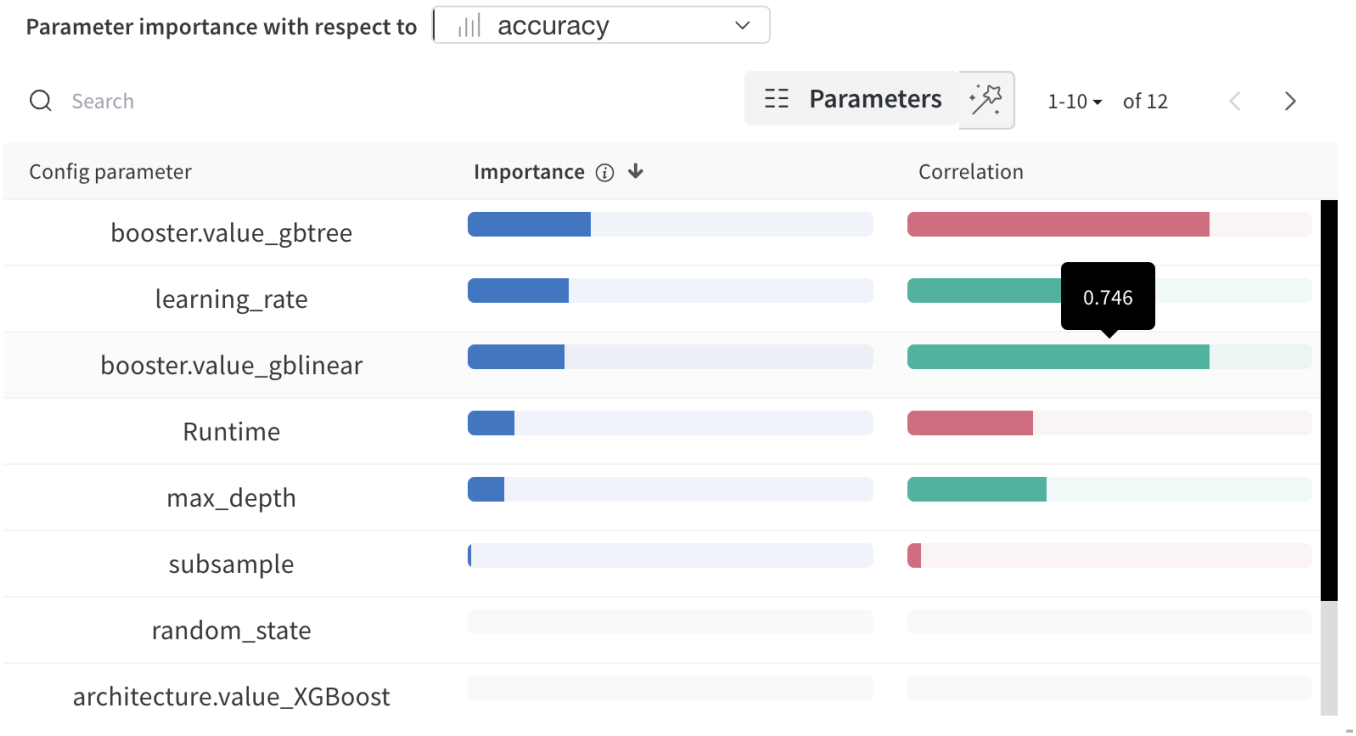
Metric	Value
step	1500
Eval Accuracy	0.898673
Eval F1 Score	0.898269
Eval Loss	0.541631
Eval Precision	0.900237
Eval Recall	0.898673

CLIP + XGBoost

Start Date	2025-02-13	
End Date	2025-02-15	
dataset	DesktopUI	





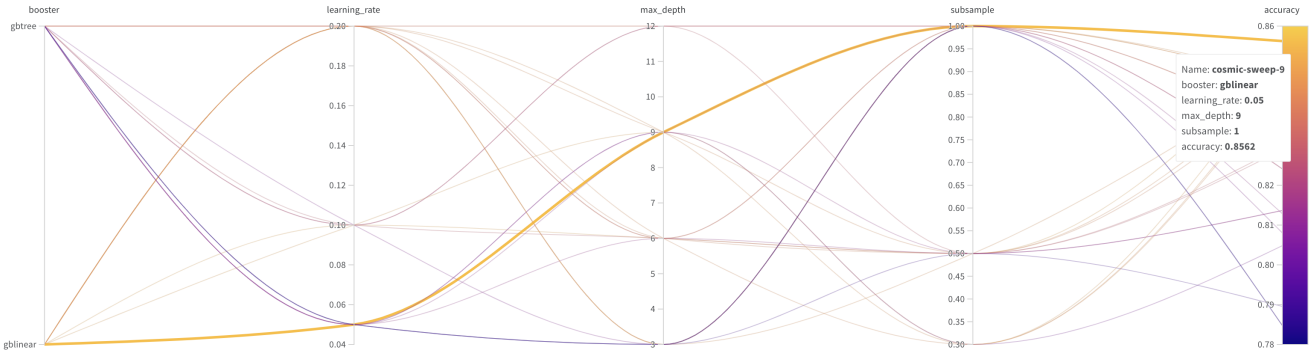


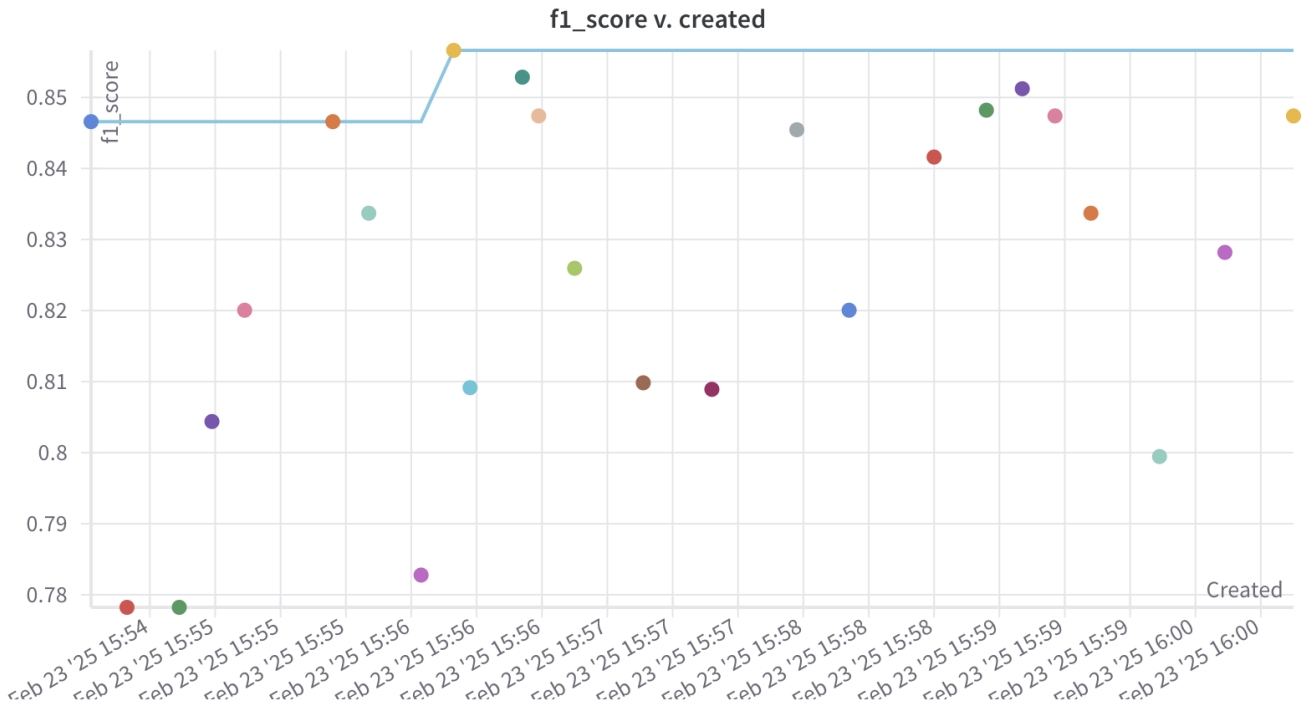
Results:
"Unfortunately, Weights & Biases did not save the history of metric progression. However, you can check them in the repository."

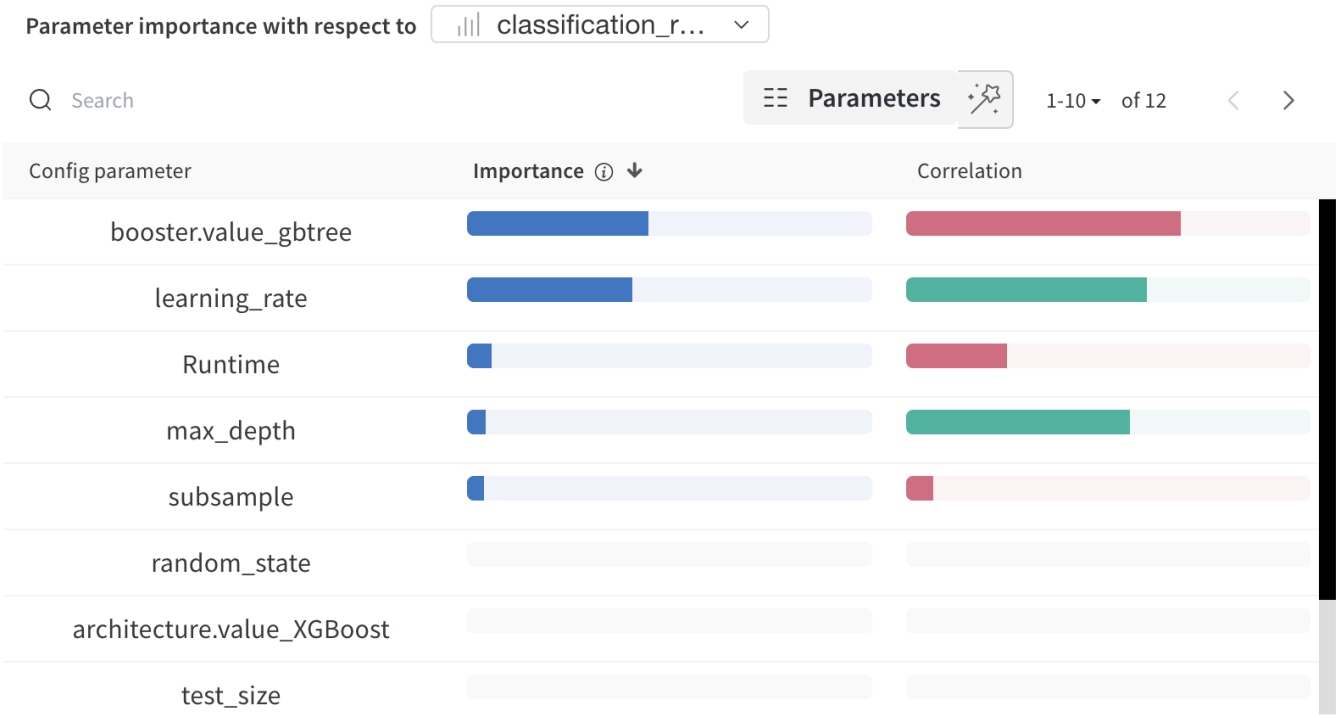
Metric	Value
Accuracy	0.43293
F1 Score	0.41927
Precision	0.41268
Recall	0.43293

BLIP2 + XGBoost

Start Date	2025-02-16
End Date	2025-02-19
dataset	DesktopUI


booster
learning_rate
max_depth
subsample
accuracy
gbtree
gblinear
Name: cosmic-sweep-9
booster: gblinear
learning_rate: 0.05
max_depth: 9
subsample: 1
accuracy: 0.8562


f1_score v. created
f1_score
Created
0.85
0.84
0.83
0.82
0.81
0.8
0.79
0.78
Feb 23 '25 15:54
Feb 23 '25 15:55
Feb 23 '25 15:55
Feb 23 '25 15:55
Feb 23 '25 15:56
Feb 23 '25 15:56
Feb 23 '25 15:56
Feb 23 '25 15:57
Feb 23 '25 15:57
Feb 23 '25 15:57
Feb 23 '25 15:58
Feb 23 '25 15:58
Feb 23 '25 15:58
Feb 23 '25 15:59
Feb 23 '25 15:59
Feb 23 '25 15:59
Feb 23 '25 16:00
Feb 23 '25 16:00

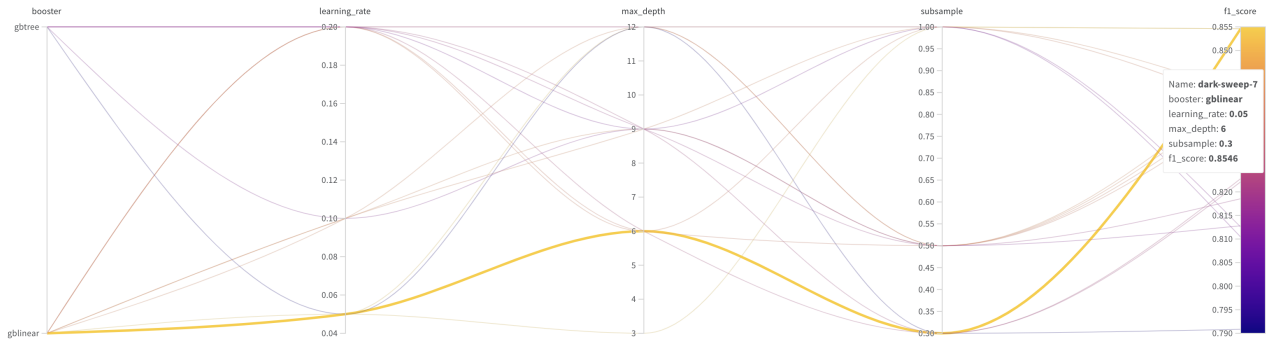


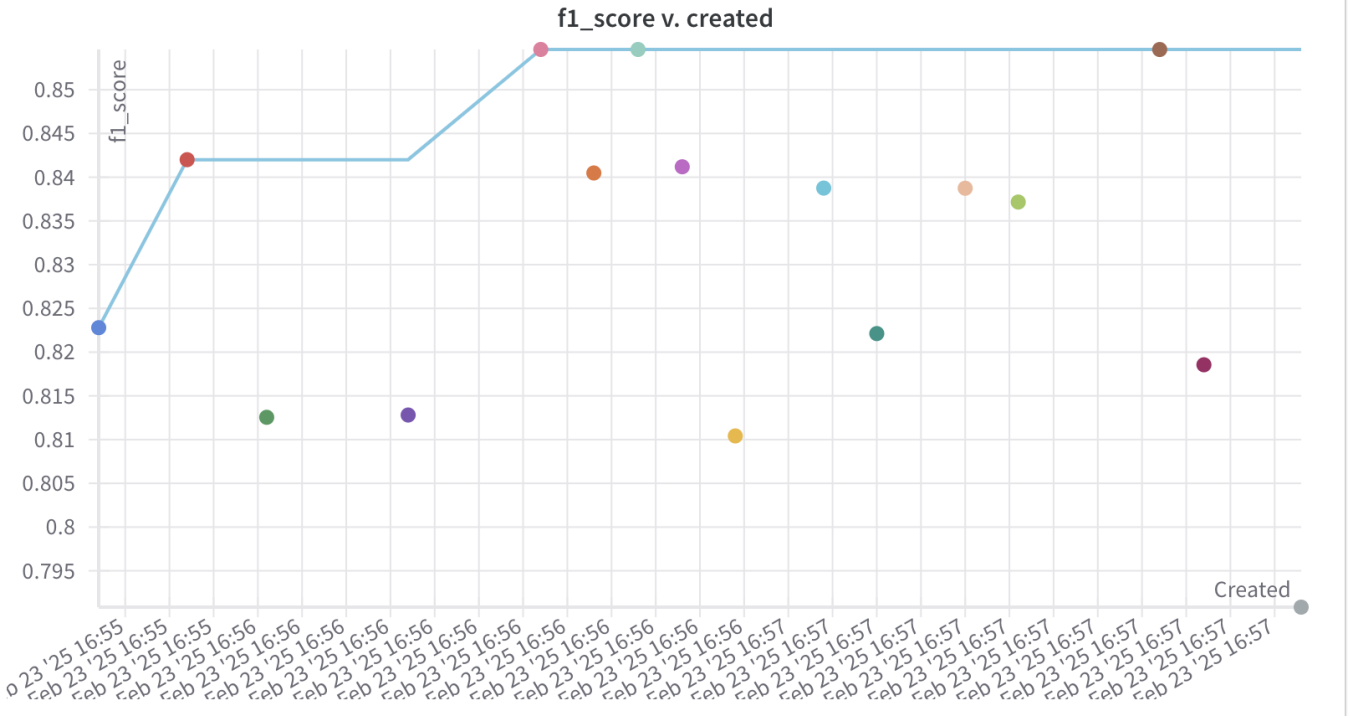
Results

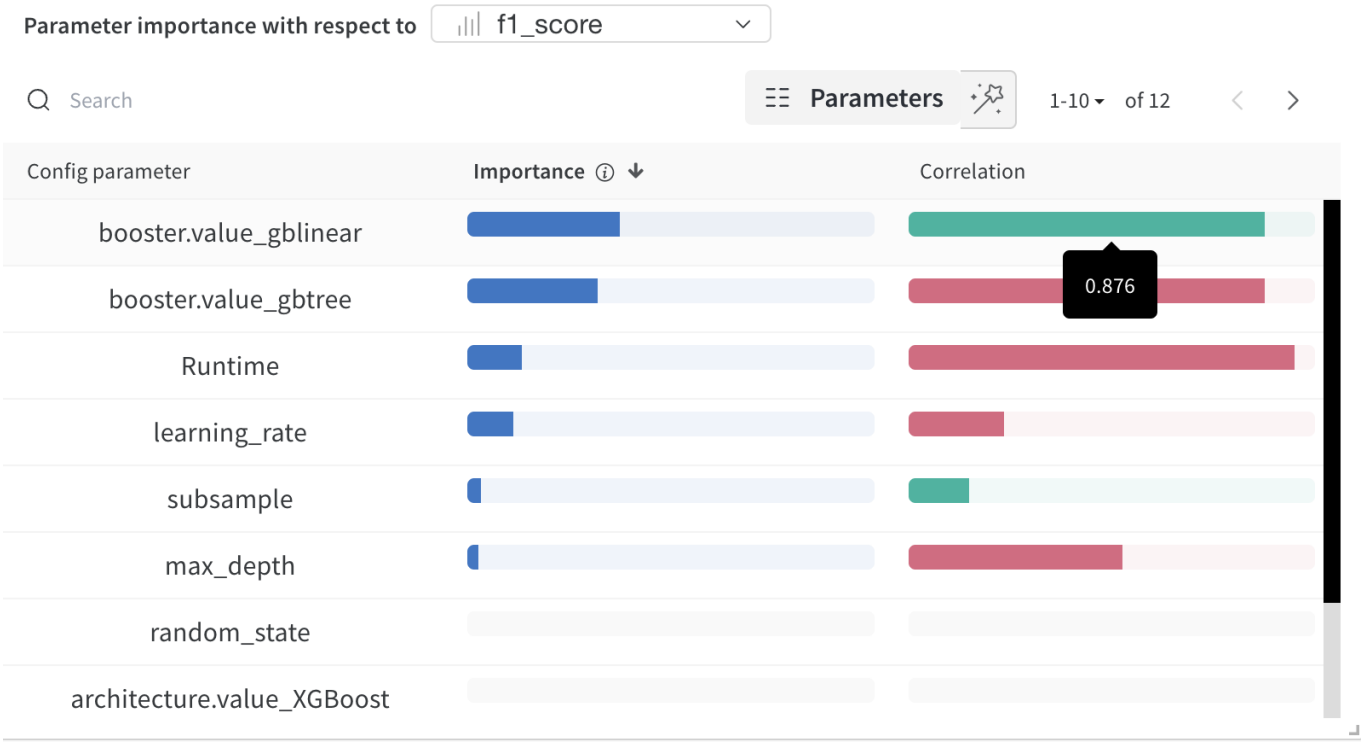
Metric	Value
Accuracy	0.84659
F1 Score	0.8474
Precision	0.84967
Recall	0.84659

SigLIP2 + XGBoost

Start Date	2025-02-20
End Date	2025-02-23
dataset	DesktopUI


booster
learning_rate
max_depth
subsample
f1_score
gbtree
gblinear
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
12
11
10
9
8
7
6
5
4
3
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.855
0.850
0.820
0.815
0.810
0.805
0.800
0.795
0.790
Name: dark-sweep-7
booster: gblinear
learning_rate: 0.05
max_depth: 6
subsample: 0.3
f1_score: 0.8546





Metric	Value
Accuracy	0.85462
F1 Score	0.8546
Precision	0.85477
Recall	0.85462

Summary

Among the four approaches evaluated, the Vision Transformer (ViT) demonstrated the best performance, achieving an F1 score of up to 0.89 when trained for 10 epochs using FP16 precision. However, this improvement came at a significant computational cost—training ViT took approximately two hours on a GPU. In contrast, SigLIP2 embeddings were extracted within just 15 minutes on a CPU, making it a far more efficient alternative.

This observation leads to the conclusion that fine-tuning Visual Transformers can be an effective strategy for maximizing performance metrics. However, the process is both computationally expensive and time-consuming, making it less practical for scenarios with resource constraints. For future work, I would recommend exploring alternative approaches that strike a balance between efficiency and accuracy, such as lightweight transformer models, optimized embedding techniques, or self-supervised learning methods.