Image Captioning in Personal Photo Management

BHT

Berliner Hochschule
für Technik

Studiere Zukunft

Personal Photo Management

Need to manage a lot of photos from different devices

Common sort criteria are when, where, who and what and this is mostly automated thanks to machine learning and metadata

Image captions in personal photo management is on the other hand rarely looked at

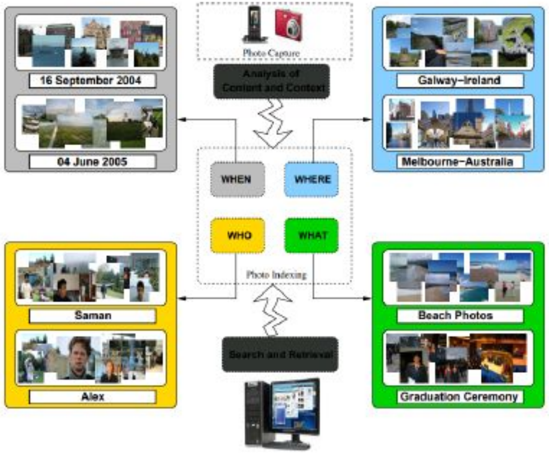


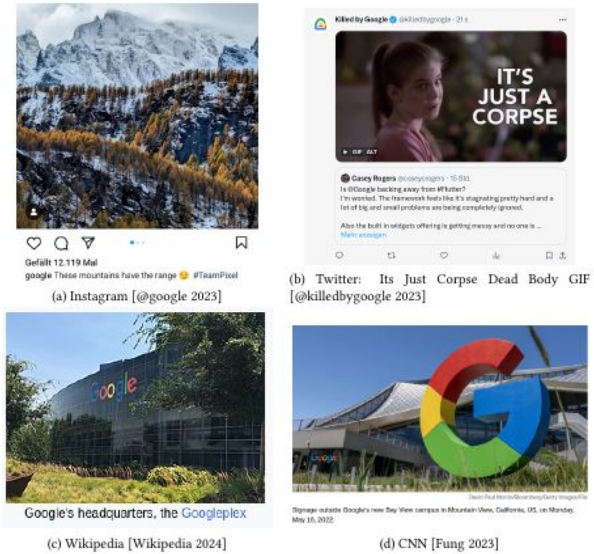
Figure 1.2: W4 entries: when, where, what and who [Cooray 2008]

Image Captions

Large differences between platforms

Different goals from searchability to reacting to other people

Different mechanisms like hashtags or attribution



Gefällt 12.119 Mal

google These mountains have the range 😉 #TeamPixel

(a) Instagram [@google 2023]

Killed by Google @killedbygoogle · 21 s

Casey Rogers @caseycrogers · 15 Std.

Is @Google backing away from #Flutter?
I'm worried. The framework feels like it's stagnating pretty hard and a lot of big and small problems are being completely ignored.

Also the built in widgets offering is getting messy and no one is …
Mehr anzeigen

(b) Twitter: Its Just Corpse Dead Body GIF [@killedbygoogle 2023]

Google's headquarters, the Googleplex

(c) Wikipedia [Wikipedia 2024]

David Paul Morris/Bloomberg/Getty Images/File

Signage outside Google's new Bay View campus in Mountain View, California, US, on Monday, May 16, 2022.

(d) CNN [Fung 2023]

Figure 1.1: Different captions on platforms

Understanding the needs and limitations of self-hosted systems

Questionnaire of 20 LibrePhotos Users

Technical Survey, which collected server stats

Needs:

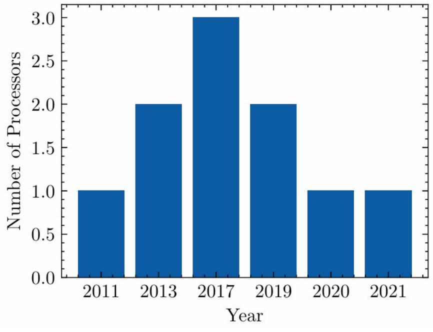
Self-hosted, private, accurate, optional, hardware accelerated, customizable, usage of previous attained knowledge like persons

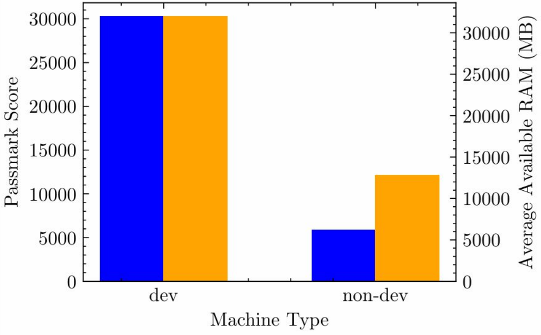
Wants:

Multiple languages, different styles of captions

Limitations:

Under 8GB of RAM and not use SSE4, AVX and AVX2





im2txt, BLIP and ONNX

im2txt is a CNN LSTM encoder decoder image to text model

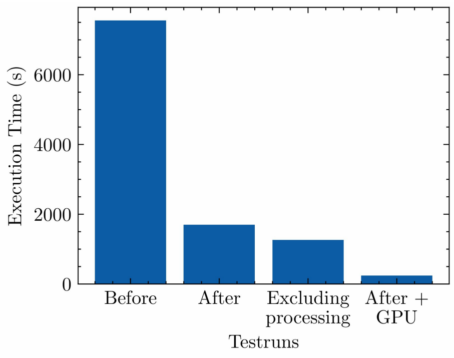
Looked into common benchmarks and established a baseline

Evaluate optimization including converting to ONNX

Contrasted this with BLIP, which is based on a transformer based architecture

(always use finetunes and not base models!)

(quantizing and converting models is difficult!)



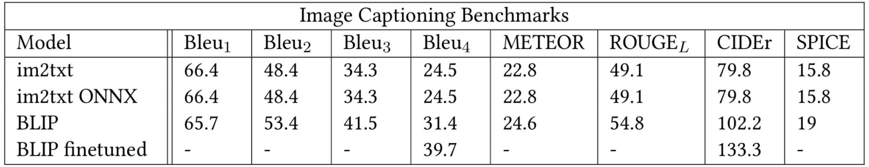


Image Captioning Benchmarks

Model	Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGE_L	CIDEr	SPICE
im2txt	66.4	48.4	34.3	24.5	22.8	49.1	79.8	15.8
im2txt ONNX	66.4	48.4	34.3	24.5	22.8	49.1	79.8	15.8
BLIP	65.7	53.4	41.5	31.4	24.6	54.8	102.2	19
BLIP finetuned	-	-	-	39.7	-	-	133.3	-

Adding an LLM

Semantic text tasks can be solved by LLMs like adding previous attained knowledge or translating captions

Criteria:

Should be self-hostable, e.g. <8GB of RAM

Needed fitting license

Should perform well enough for the task

Final model: Mistral 7B v0.1 by theBloke
(switched by now to an instruct model)

Study LLMs + Transformer based methods

Randomized captions and image order

Users should rate the captions, from 1-5

20 images with 62 responses

Used LimeSurvey as platform

CLIPScore is a reference-free Evaluation Metric

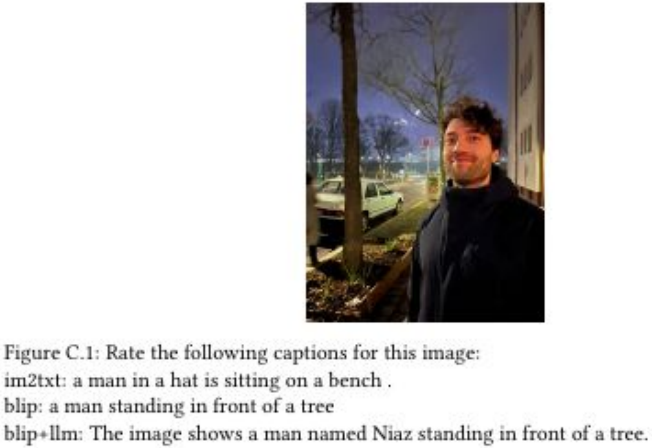


Figure C.1: Rate the following captions for this image:
im2txt: a man in a hat is sitting on a bench .
blip: a man standing in front of a tree
blip+llm: The image shows a man named Niaz standing in front of a tree.

Results and Learnings

CLIPScore can identify bad hallucination as the score decreases and increased in general

Using without prior knowledge bad idea, same with names, but users liked places

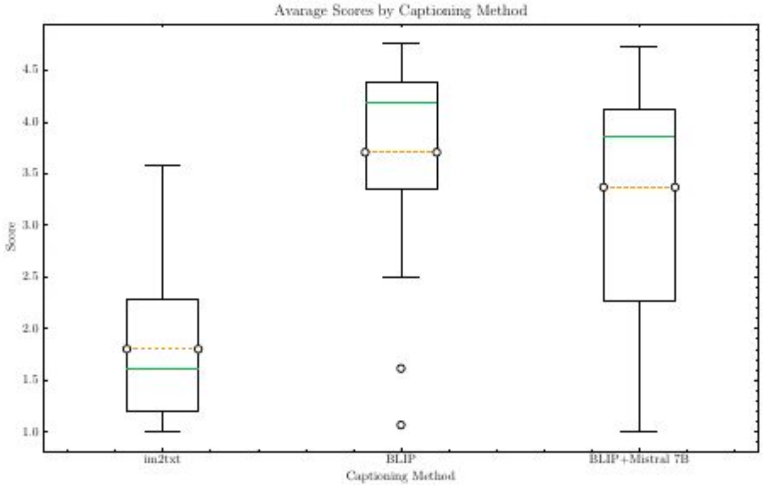


Figure 6.1: Human Ratings comparing im2txt, BLIP and BLIP + Mistral 7B

Future Work and Challenges

Trying out SOTA models like ChatGPT and multimodal models

Benchmark for user preferences with captions is missing

Adding an OCR system to solve TextCaps

Understanding multilanguage support (will look into mistral-nemo-12b)

Evaluating existing fine-tunes and creating new ones

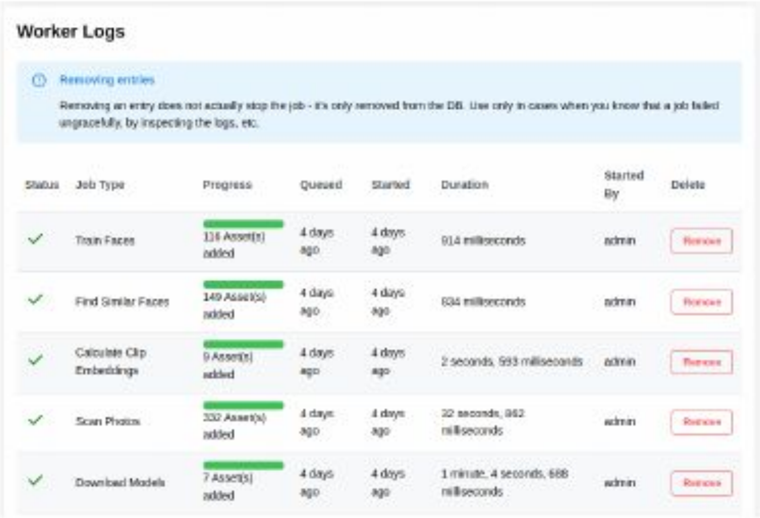
Management of machine learning models

Issue: How to ship models with image captioning service

Shipping the model in docker image is also possible, but does not scale well

Implemented a job to download models from known sources together with settings fields

Checks every time a job gets executed



Worker Logs

Removing entries

Removing an entry does not actually stop the job - it's only removed from the DB. Use only in cases when you know that a job failed ungracefully, by inspecting the logs, etc.

Status	Job Type	Progress	Queued	Started	Duration	Started By	Delete
✓	Train Faces	116 Asset(s) added	4 days ago	4 days ago	914 milliseconds	admin	Remove
✓	Find Similar Faces	149 Asset(s) added	4 days ago	4 days ago	634 milliseconds	admin	Remove
✓	Calculate Clip Embeddings	9 Asset(s) added	4 days ago	4 days ago	2 seconds, 593 milliseconds	admin	Remove
✓	Scan Photos	332 Asset(s) added	4 days ago	4 days ago	32 seconds, 862 milliseconds	admin	Remove
✓	Download Models	7 Asset(s) added	4 days ago	4 days ago	1 minute, 4 seconds, 688 milliseconds	admin	Remove

Figure 7.8: Worker logs with an entry for download models

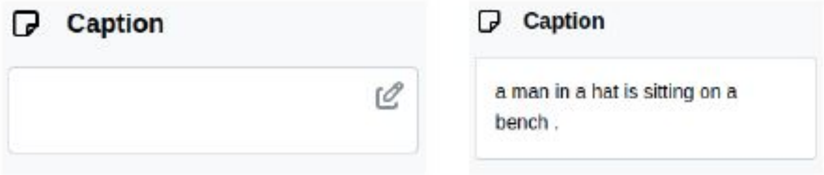
User Experience of generating captions

Issue: User may already have captions and only want to generate new ones sometimes

I added a suggestion box with a suggestion

User has to click to generate a caption on the magic wand icon

Possible different UX when using in a context of sharing images or when improving accessibility

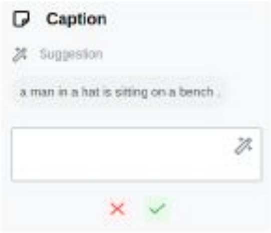


a man in a hat is sitting on a bench .

(a) Owner view of caption

(b) Shared with another user

Figure 7.3: Different caption views



a man in a hat is sitting on a bench .

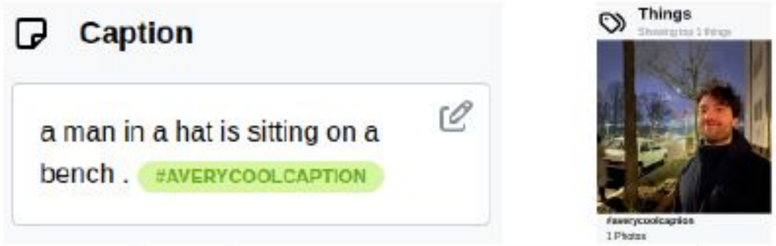
Figure 7.4: Edit Mode with Suggestion

Tags in Caption

Issue: User may prefer caption with hashtags

Implemented automated albums based on tags

This could also be something social, like tagging another user of an instance, where you maybe need an overview



Caption

a man in a hat is sitting on a bench . #AVERYCOOLCAPTION

(a) Tags within captions

(b) Tag Album

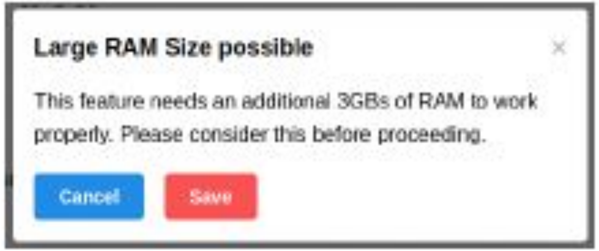
Figure 7.5: Tags implemented with Tiptap

RAM usage in python

Issue: User do not have enough RAM or want more efficient RAM usages after caption creation

Added a warning when selecting large models

Implemented restarting of flask service after a time of non usage to ensure model is unloaded after a time (hard issue!)



Large RAM Size possible

This feature needs an additional 3GBs of RAM to work properly. Please consider this before proceeding.

Figure 7.2: Warning of High RAM usage

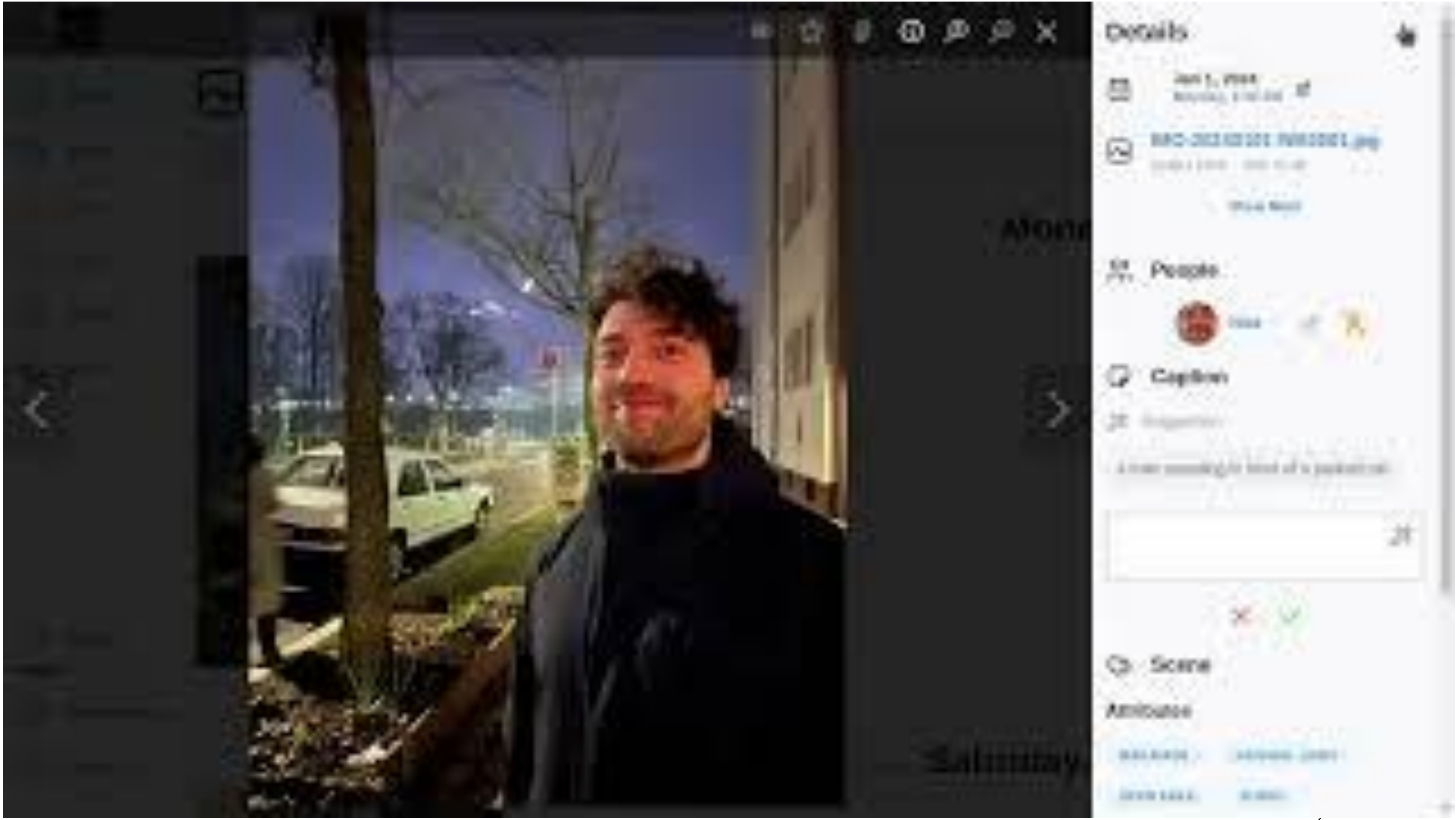
Hardware acceleration

Issue: User want faster image captioning performance

Docker image with Nvidia support used

Issues with supporting users as the GPU does not get recognized sometimes

Issues with unloading models, sometimes the driver reports having no VRAM



Details

Jan 1, 2024
[illegible]

[illegible].jpg
[illegible]

[illegible]

People

[illegible]

Caption

[illegible]

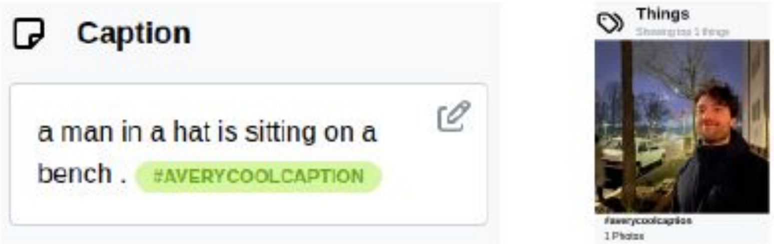
A man standing in front of a parked car.

Scene

Attributes

[illegible] [illegible]
[illegible] [illegible]

Demo: Integrating it into LibrePhotos

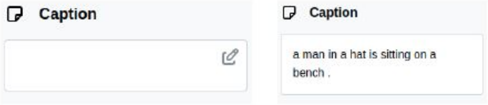


a man in a hat is sitting on a bench . #AVERYCOOLCAPTION

(a) Tags within captions

(b) Tag Album

Figure 7.5: Tags implemented with Tiptap

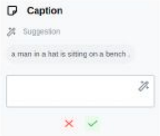


a man in a hat is sitting on a bench .

(a) Owner view of caption

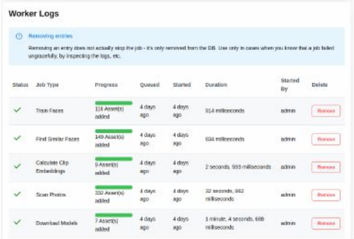
(b) Shared with another user

Figure 7.3: Different caption views



a man in a hat is sitting on a bench .

Figure 7.4: Edit Mode with Suggestion



Worker Logs

Removing entries

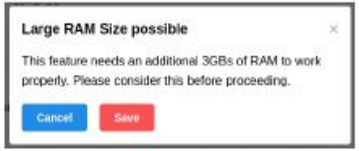
Removing an entry does not actually stop the job - it's only removed from the DB. Use only in cases when you know that a job failed ungracefully, by inspecting the logs, etc.

Status	Job Type	Progress	Queued	Started	Duration	Started By	Delete
✓	Train Faces	116 Asset(s) added	4 days ago	4 days ago	914 milliseconds	admin	Remove
✓	Find Similar Faces	149 Asset(s) added	4 days ago	4 days ago	634 milliseconds	admin	Remove
✓	Calculate Clip Embeddings	9 Asset(s) added	4 days ago	4 days ago	2 seconds, 593 milliseconds	admin	Remove
✓	Scan Photos	332 Asset(s) added	4 days ago	4 days ago	32 seconds, 862 milliseconds	admin	Remove
✓	Download Models	7 Asset(s) added	4 days ago	4 days ago	1 minute, 4 seconds, 688 milliseconds	admin	Remove

Figure 7.8: Worker logs with an entry for download models

Version: dev

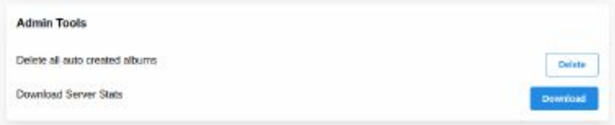
Figure 7.7: Storage and version



Large RAM Size possible

This feature needs an additional 3GBs of RAM to work properly. Please consider this before proceeding.

Figure 7.2: Warning of High RAM usage



Admin Tools

Delete all auto created albums

Download Server Stats

Figure 7.6: Download server stats button