Demonstration of a Simple Transformer Running on the NPU of an STM32N6

ACA Project – A.Y. 2024/2025

Authors
Acquadro Patrizio
Drugman Tito Nicola

Supervisors
Prof. Silvano Cristina
Dott. Ronzani Marco

POLITECNICO
MILANO 1863

Objectives

TO BE ACHIEVED

1 • Deploy Model Zoo CNN

2 • Build and Deploy Custom CNN

3 • Full-Stack Transformer Deployment

4 • Metric Collection & Architecture Comparison

Hardware

STM32N6570-DK
THE BOARD

LOCAL HARDWARE
2 NVIDIA GEFORCE RTX 5060 TI

ST Edge AI Developer Cloud
REMOTE BOARD

Toolchain

STM32CubeIDE
GENERATE CODE AND VALIDATE

STM32CubeProgrammer
FLASHING BINs

ANACONDA PROMPT
RUN PYTHON SCRIPTS

Task & Dataset

Supervised Image Classification: broad support in Model Zoo.

Flowers: 5 classes dataset

- 3.7K RGB images.
- 20% for quantization.
- 80/20 train-validation split.

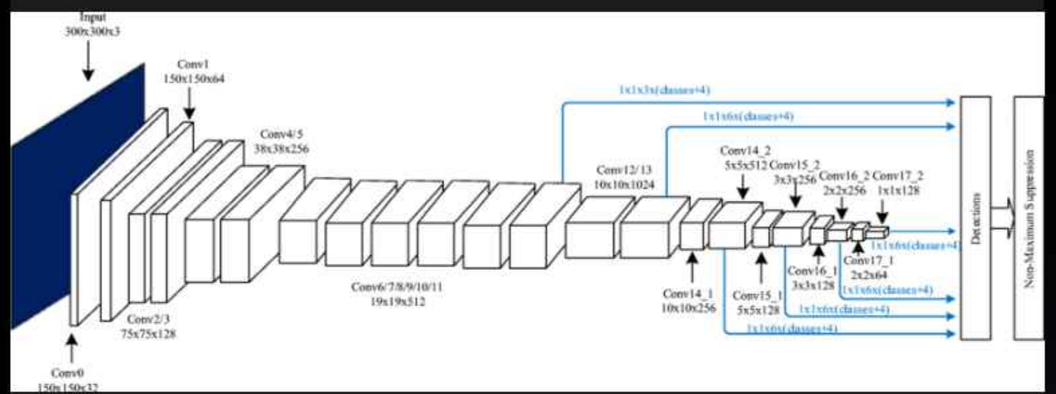
Preprocessing: RGB conversion, resizing and normalization.

Augmentation: flip, contrast, brightness, translation, rotation, zoom.

FLOWERS DATASET

Kaggle

CNNs

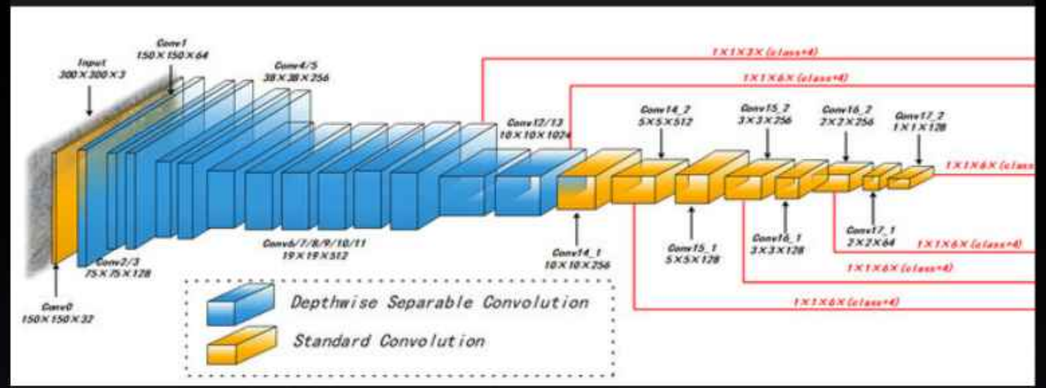


MOBILENET V2

PRETRAINED ON IMAGENET
α = 0.35 → ~16–1280 CHANNELS
17 INVERTED RESIDUAL BLOCKS (t=6)

TRAIN: DROPOUT (0.3), 100 EPOCHS, ADAM (lr=0.001)

QUANTIZATION: INT8 (POST-TRAIN)



CUSTOM CNNS

TRAINED FROM SCRATCH
α = 1 → ~32–128 CHANNELS
4 INVERTED RESIDUAL BLOCKS (t=2-3)

TRAIN: DROPOUT (***0.25***), 100 EPOCHS, ADAM (lr=0.001)

QUANTIZATION: INT8 (POST-TRAIN)

Transformer Architecture

SPECS & INPUT

~4.56M parameters.
Input: 30 tokens.
Vocab size: 20K.
128-dim embeddings + positional encodings.

CORE STRUCTURE

6 encoder blocks.
4-head MHA.
FFN: Linear with ReLU.
Residual connections + LayerNorm.

OUTPUT LAYER

Dense: 128 → 20K.
20K logits → softmax.

GENERATED FILES

Model weights: learned during training.
Tokenizer: maps each word to an index.
Embedding matrix: 20k x 128 size.

Tokenizer

CHARACTER-LEVEL

Vocab size: 66 (A–Z, a–z, digits, punctuation).
Embedding size: 66 × 128 → very compact.
Generates 1 character per step → slow & verbose.
No PAD/OOV tokens → full coverage.
Low Flash/SRAM usage.
Poor semantic/syntactic learning.

Tokenizer

CHARACTER-LEVEL

Vocab size: 66 (A–Z, a–z, digits, punctuation).
Embedding size: 66 × 128 → very compact.
Generates 1 character per step → slow & verbose.
No PAD/OOV tokens → full coverage.
Low Flash/SRAM usage.
Poor semantic/syntactic learning.

Tokenizer

WORD-LEVEL

Vocab size: 20'000 most frequent words.
Embedding size: 20K × 128 → large footprint.
Generates full words → efficient context usage.
OOV mapped to ID=1 → limited generalization.
High memory cost.
Better meaning retention, despite sparsity.

Datasets

Training Data

- Source: 4 classic English literary works.
- Merged using a custom Python script.
- Final size: 7.68 MB, ~1.4 million words.

Evaluation Data

- *The Adventures of Sherlock Holmes* by *Arthur Conan Doyle*.

Reasoning

- Diverse, high-quality English.
- No overlap training-evaluation.
- Textual richness.

PROJECT GUTENBERG

75k EBooks

FULL STACK

TRAINING

90/10 TRAIN-VAL
100 EPOCHS

QUANTIZATION

INT-8 POST TRAIN &
CALIBRATION SET

life.augmented

DEPLOYMENT

STM32CubeIDE
STM32Programmer

INFERENCE

AUTOREGRESSION
WITH NO MASKS

Now, let’s admire this masterpiece of modern poetry.

Anaconda Prompt - conda ac

```
(st_zoo_validate) C:\Users\drugm\stm32ai-modelzoo-services\dd2_text_gen_stm32n6v2\text_gen_stm32n6>python serial_demo.py
```

Model Metrics

Accuracy (14.8%) = 3000×
Perplexity (653) = 1/30

Embeddings

Cosine similarity between pairs and single tokens

Query: You		
Rank	Token	Score
1	pause	0.7655
2	eat	0.7647
3	enough	0.7572
4	reality	0.7495
5	rumour	0.7474
6	spoke	0.7465
7	ginger	0.7443
8	brown	0.7433
9	examples	0.7370

On-Board

Model size & Tot Flash
Peak SRAM & Inference

Metric	Value	Unit / Notes
Flash footprint (model + embeddings)	7.159	MB (quantized, external Octo-SPI)
Model weights	4.590	MB
Int8 model parameters	4'562'979	values
Embedding file size	2.560	MB
Int8 embedding entries	20'000 · 128	values
Peak SRAM (activations + buffers)	1.208	MB
MACC	172,531,800	ops
Epochs executed	407	epochs
Epochs on NPU (hardware)	257	runs
Epochs on CPU (software)	150	runs
Tokenizer file size (host PC)	745	KB (JSON)
Generation length (reported run)	128	tokens
Time for 128-token generation	9.361	s (single representative run)
Average latency per generated token	73.14	ms/token (same run)
Throughput	13.67	tokens/s (same run)

Embeddings

Cosine similarity between **pairs** and **single** tokens

Query: You		
Rank	Token	Score
1	pause	0.7655
2	cat	0.7647
3	enough	0.7572
4	reality	0.7495
5	rumour	0.7474
6	spoke	0.7465
7	ginger	0.7443
8	brown	0.7433
9	examples	0.7370

Model Metrics

Accuracy (14.8%) = 3000×
Perplexity (653) = 1/30

On-Board

Model size & Tot Flash
Peak SRAM & Inference

Metric	Value	Unit / Notes
Flash footprint (model + embeddings)	7.159	MB (quantized, external Octo-SPI)
Model weights	4.590	MB
Int8 model parameters	4'562'979	values
Embedding file size	2.560	MB
Int8 embedding entries	20'000 · 128	values
Peak SRAM (activations + buffers)	1.208	MB
MACC	172,531,800	ops
Epochs executed	407	epochs
Epochs on NPU (hardware)	257	runs
Epochs on CPU (software)	150	runs
Tokenizer file size (host PC)	745	KB (JSON)
Generation length (reported run)	128	tokens
Time for 128-token generation	9.361	s (single representative run)
Average latency per generated token	73.14	ms/token (same run)
Throughput	13.67	tokens/s (same run)

On-Board

Model size & Tot Flash
Peak SRAM & Inference

Metric	Value	Unit / Notes
Flash footprint (model + embeddings)	7.159	MB (quantized, external Octo-SPI)
Model weights	4.590	MB
Int8 model parameters	4'562'979	values
Embedding file size	2.560	MB
Int8 embedding entries	$20'000 \cdot 128$	values
Peak SRAM (activations + buffers)	1.208	MB
MACC	172,531,800	ops
Epochs executed	407	epochs
Epochs on NPU (hardware)	257	runs
Epochs on CPU (software)	150	runs
Tokenizer file size (host PC)	745	KB (JSON)
Generation length (reported run)	128	tokens
Time for 128-token generation	9.361	s (single representative run)
Average latency per generated token	73.14	ms/token (same run)
Throughput	13.67	tokens/s (same run)

Model Metrics

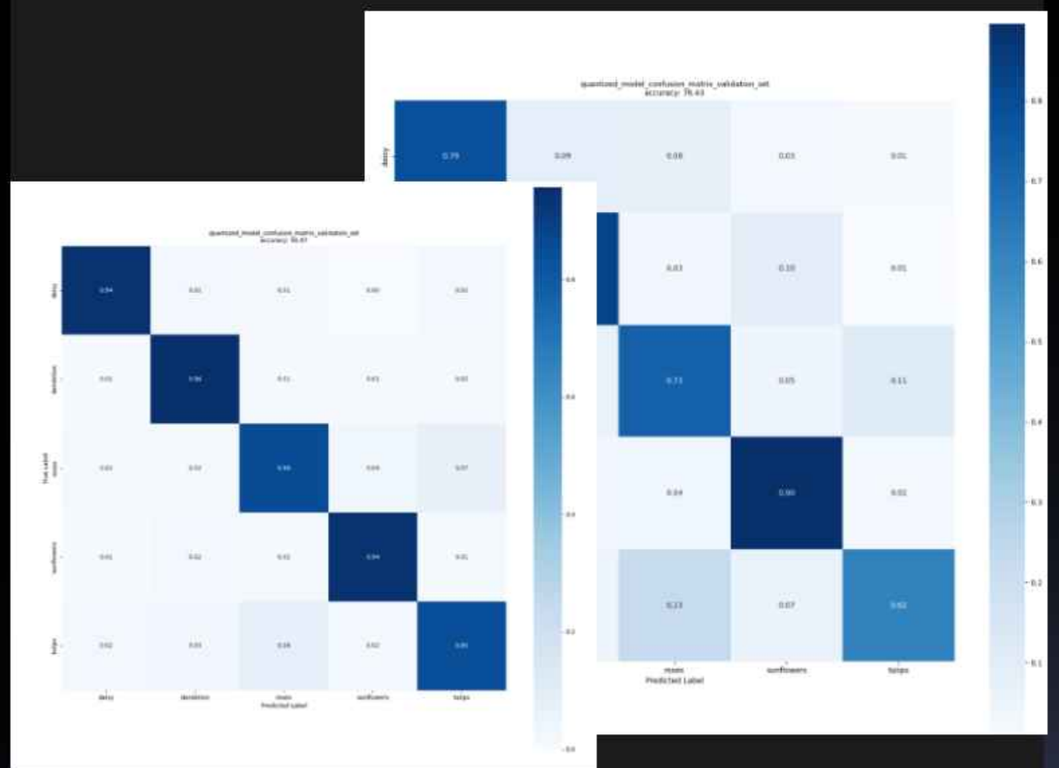
Accuracy (14.8%) = 3000×
Perplexity (653) = 1/30

Embeddings

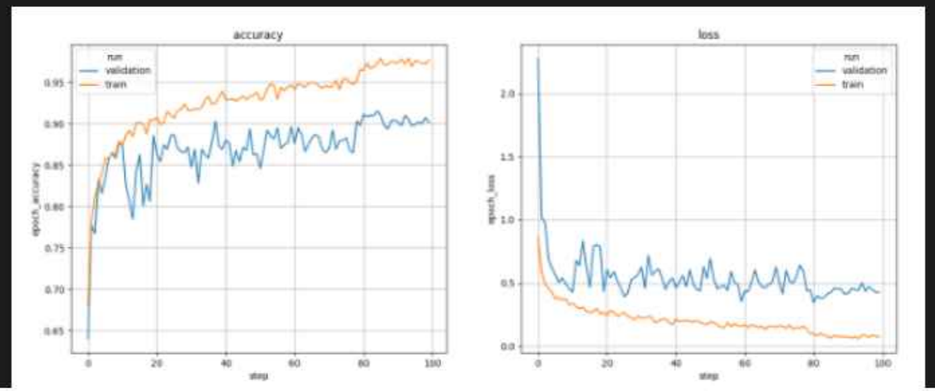
Cosine similarity between pairs and single tokens

Query: You		
Rank	Token	Score
1	pause	0.7655
2	cat	0.7647
3	enough	0.7572
4	reality	0.7495
5	rumour	0.7474
6	spoke	0.7465
7	ginger	0.7443
8	brown	0.7433
9	examples	0.7370

CNNs



MODELZOO VS CUSTOM CNNs



Metric	Model 1	Model 2	Rel. Change
Parameters (count)	395 493	235 493	−40.5 %
Weight memory (KiB)	594.3	305.8	−48.6 %
Activation memory (KiB)	240.0	576.0	140.0 %
Total memory (KiB)	834.3	881.8	5.7 %
Total number of epochs	58	20	
Training accuracy (%)	97.5	100.0	2.6 pp
Validation accuracy (%)	90.87	76.43	−15.9 %
Gen. gap (train–val, pp)	6.6	23.6	258 %

MODELZOO VS CUSTOM CNNs

...

5. Conclusions & Future Directions

Conclusions

- End-to-end INT8 Transformer (4.6 M params) successfully deployed on STM32N6570-DK.
- Fits comfortably on-board: ≈7 MB flash, ≈1.2 MB SRAM.
- 73 ms per token → 13.7 tokens/s; NPU handles 63% of epochs.
- 14.8% accuracy & 653 perplexity on unseen text. ~3 000× above random uniform baseline.
- Embeddings show some meaningful word-level semantics, validating architecture choice.
- Besides educational purposes, there are only a few practical applications for an embedded Transformer for text generation.

Future Work

- Train on larger and cleaner datasets.
- Explore language adaptation to Italian.
- Expand embedding size (128 → 256).
- Fine-tune larger models (deeper/larger blocks) using spare flash/SRAM.

Thank you

QUESTIONS?

Patrizio Acquadro, Tito Nicola Drugman
patrizio.acquadro@mail.polimi.it
titonicola.drugman@mail.polimi.it

Query: Romeo		
Rank	Token	Score
1	thee	0.7894
2	withdrew	0.7832
3	dice	0.7825
4	interview	0.7820
5	consists	0.7819
6	stricken	0.7808
7	roared,	0.7792
8	suspicious	0.7784
9	yield,	0.7768
10	witness	0.7746

Query: You		
Rank	Token	Score
1	pause	0.7655
2	eat	0.7647
3	enough	0.7572
4	reality	0.7495
5	rumour	0.7474
6	spoke	0.7465
7	ginger	0.7443
8	brown	0.7433
9	examples	0.7370
10	commence	0.7355

Query: King		
Rank	Token	Score
1	Ephesus.	0.7676
2	conversed	0.7667
3	well.	0.7612
4	death's	0.7572
5	marry	0.7566
6	gravel	0.7480
7	men's	0.7480
8	sooner	0.7460
9	Master,	0.7458
10	Either	0.7450

Cosine(Romeo, Juliet)	0.6538

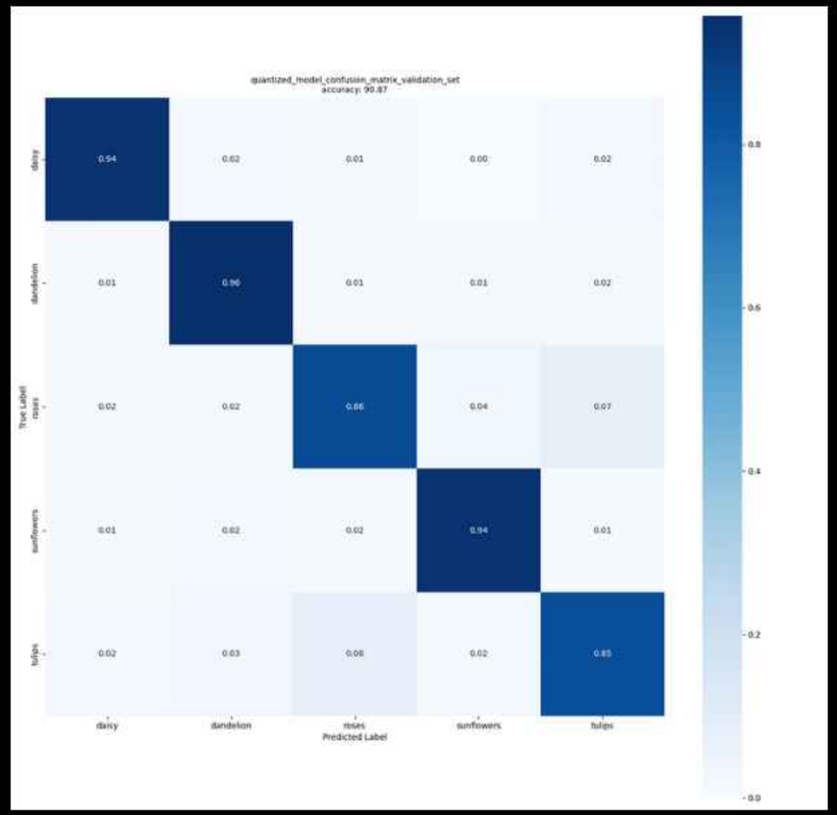
Table 4: Top-10 most similar tokens for each query, plus the cosine similarity between *Romeo* and *Juliet*.

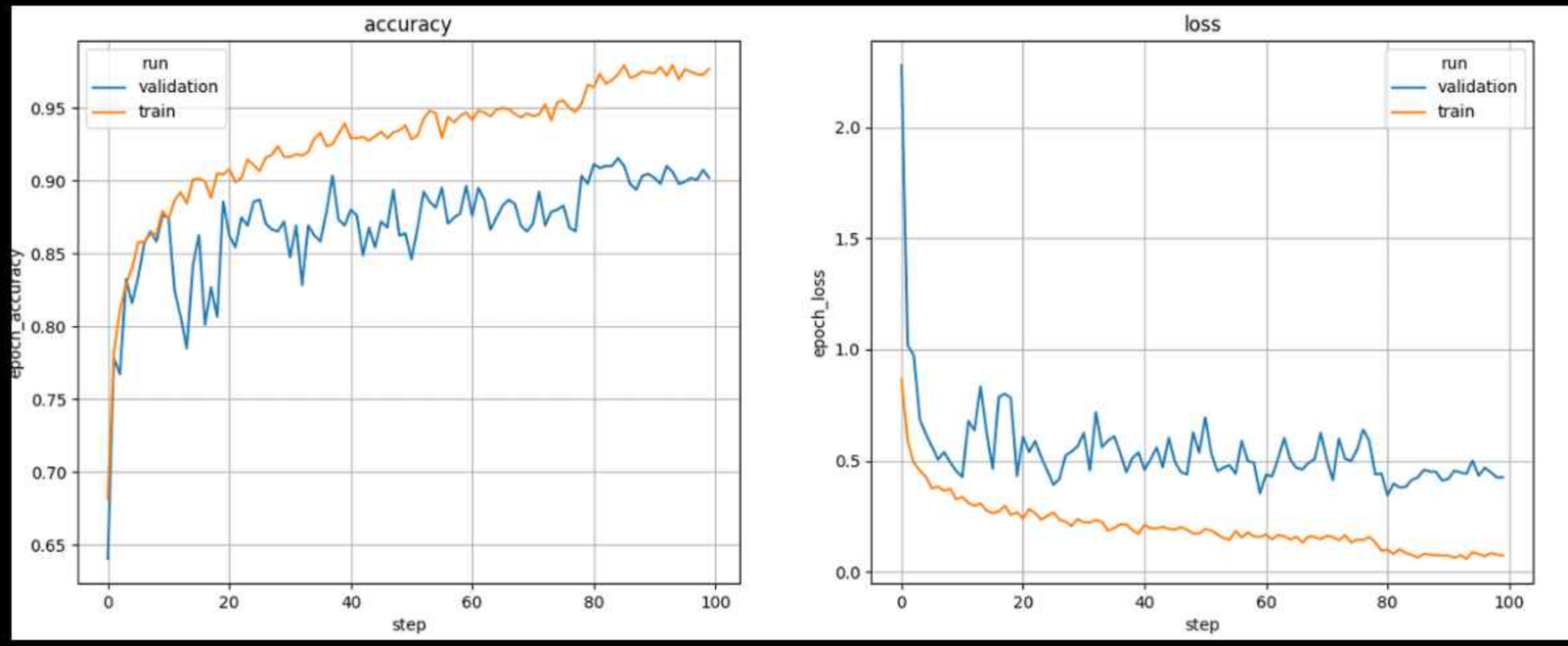
Metric	Value	Unit / Notes
Flash footprint (model + embeddings)	7.159	MB (quantized, external Octo-SPI)
Model weights	4.590	MB
Int8 model parameters	4'562'979	values
Embedding file size	2.560	MB
Int8 embedding entries	$20'000 \cdot 128$	values
Peak SRAM (activations + buffers)	1.208	MB
MACC	172,531,800	ops
Epochs executed	407	epochs
Epochs on NPU (hardware)	257	runs
Epochs on CPU (software)	150	runs
Tokenizer file size (host PC)	745	KB (JSON)
Generation length (reported run)	128	tokens
Time for 128-token generation	9.361	s (single representative run)
Average latency per generated token	73.14	ms/token (same run)
Throughput	13.67	tokens/s (same run)

Region	Address range	Usage			Weights	Activations
		Used	Total	% Used		
flexMEM	0x34000000–0x34000000	0 B	0 B	0.00%	0 B	0 B
cpuRAM1	0x34064000–0x34064000	0 B	0 B	0.00%	0 B	0 B
cpuRAM2	0x34100000–0x34200000	1.000 MB	1.000 MB	100.00%	0 B	1.000 MB
npuRAM3	0x34200000–0x34270000	147.875 kB	448.000 kB	33.01%	0 B	147.875 kB
npuRAM4	0x34270000–0x342E0000	0 B	448.000 kB	0.00%	0 B	0 B
npuRAM5	0x342E0000–0x34350000	64.688 kB	448.000 kB	14.44%	0 B	64.688 kB
npuRAM6	0x34350000–0x343C0000	0 B	448.000 kB	0.00%	0 B	0 B
octoFlash	0x71000000–0x78000000	7.159 MB	112.000 MB	6.40%	7.159 MB	0 B
hyperRAM	0x90000000–0x92000000	0 B	32.000 MB	0.00%	0 B	0 B
Total		8.367 MB			**7.159 MB**	**1.208 MB**

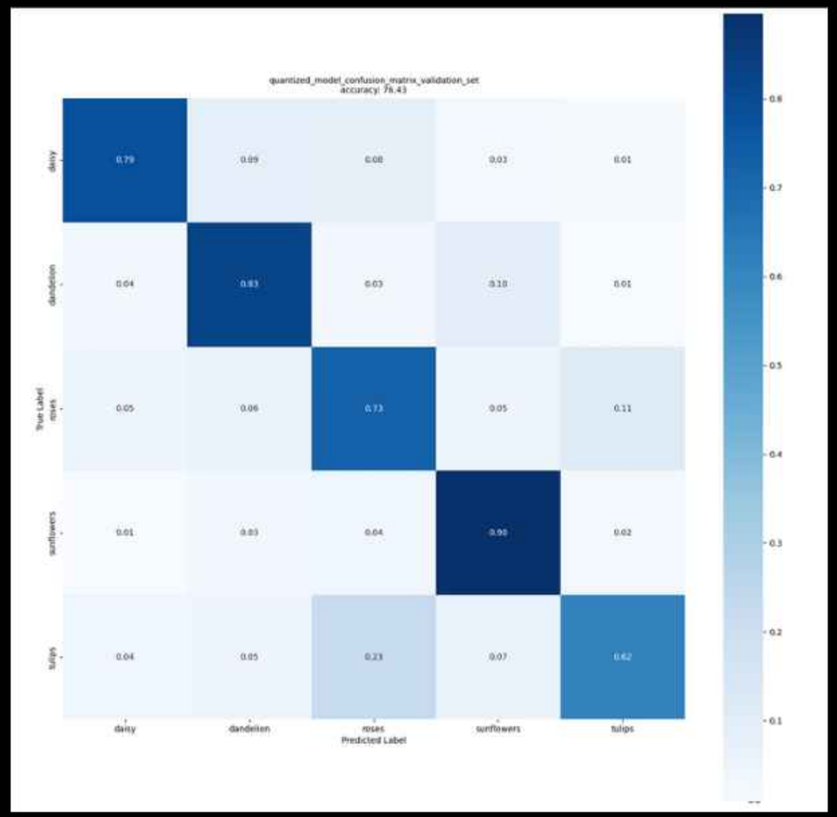
Table 2: Memory usage information (including embedding file; I/O buffers are included in activations)

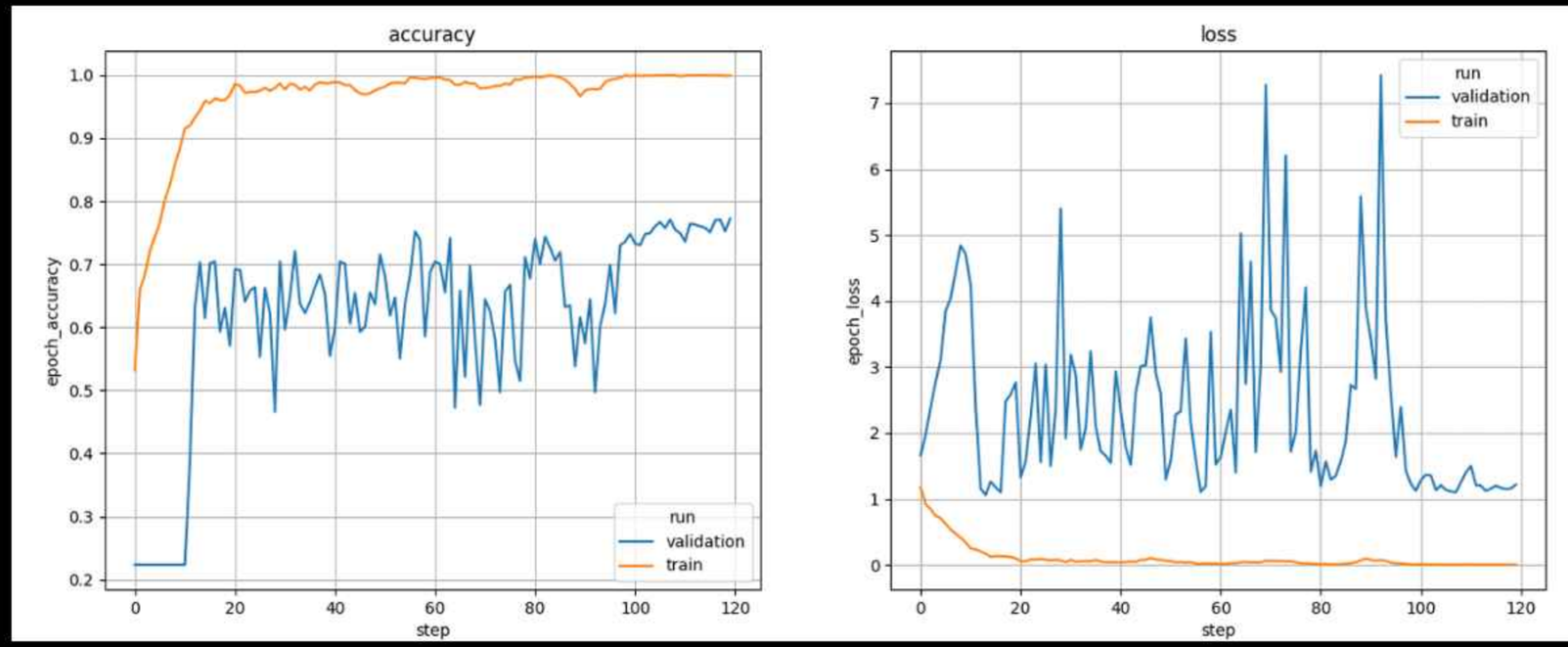
ModelZoo CNN





Custom CNN


quantized_model_confusion_matrix_validation_set
accuracy: 79.43
daisy
dandelion
roses
sunflowers
tulips
True Label
Predicted Label
0.79 0.09 0.08 0.03 0.01
0.04 0.83 0.03 0.10 0.01
0.05 0.06 0.73 0.05 0.11
0.01 0.03 0.04 0.90 0.02
0.04 0.05 0.23 0.07 0.62


accuracy
epoch_accuracy
step
run
validation
train
loss
epoch_loss
step
run
validation
train

Metric	Model 1	Model 2	Rel. Change
Parameters (count)	395 493	235 493	−40.5 %
Weight memory (KiB)	594.3	305.8	−48.6 %
Activation memory (KiB)	240.0	576.0	140.0 %
Total memory (KiB)	834.3	881.8	5.7 %
Total number of epochs	58	20	
Training accuracy (%)	97.5	100.0	2.6 pp
Validation accuracy (%)	90.87	76.43	−15.9 %
Gen. gap (train–val, pp)	6.6	23.6	258 %