Computer Vision · Assignment 2: Deep Learning for Vision

AIMS Cameroon · Structured Master's, Data Science · 2026–2027

Individual work · 100 points · Due: start of Week 3, Day 3 (set the exact date on the course page)

Submission. One PDF typeset in LaTeX plus one Jupyter notebook that reproduces every figure and number in Problem 6. Name the files `A2_<surname>.pdf` and `A2_<surname>.ipynb`. A free Colab or Kaggle GPU is sufficient. You may discuss ideas with classmates, but the write-up and code must be your own. Cite any source you use, including AI assistants, and state what you used them for.

Problem 1: Backpropagation by hand (15 points)

A network has input $x = (1, -1)$, one hidden layer of two ReLU units and one linear output:

$$h = \mathrm{ReLU}(W_1 x + b_1), \quad s = w_2^\top h + b_2, \quad \mathcal{L} = \tfrac{1}{2}(s - y)^2,$$

with $W_1 = \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix}$, $b_1 = (0,\, 0.5)$, $w_2 = (3,\, -2)$, $b_2 = 0.5$, target $y = 2$.

(a) (4) Compute the forward pass: $W_1 x + b_1$, h, s and $\mathcal{L}$.

(b) (6) Compute $\partial\mathcal{L}/\partial s$, $\partial\mathcal{L}/\partial w_2$, $\partial\mathcal{L}/\partial b_2$, $\partial\mathcal{L}/\partial W_1$ and $\partial\mathcal{L}/\partial b_1$.

(c) (3) Verify $\partial\mathcal{L}/\partial(W_1)_{11}$ with a central finite difference ($\epsilon = 0.01$).

(d) (2) Explain why the second hidden unit's incoming weights receive zero gradient, and why this can be a problem during training.

Problem 2: Shapes, parameters and receptive fields (15 points)

(a) (3) ResNet's stem applies to a $224 \times 224 \times 3$ image a 7×7 convolution with 64 filters, stride 2, padding 3, then a 3×3 max pool with stride 2, padding 1. Give both output sizes and the parameters of the convolution (with bias).

(b) (4) For a stack of layers with kernel sizes k_l and strides s_l, the receptive field obeys $r_l = r_{l-1} + (k_l - 1)\, j_{l-1}$ and $j_l = j_{l-1} s_l$, with $r_0 = j_0 = 1$. Explain each term in words, and compute the receptive field after the stem of (a) followed by one 3×3 convolution with stride 1.

(c) (4) Compute the weights (no biases) of a bottleneck block $256 \to 64 \to 64 \to 256$ (1×1, 3×3, 1×1) and of two stacked 3×3 convolutions at 256 channels. What fraction of the second does the first need?

(d) (4) Show that a depthwise separable convolution ($K \times K$ depthwise on C channels, then 1×1 to C' channels) needs a fraction $\frac{1}{C'} + \frac{1}{K^2}$ of the weights of a standard $K \times K$ convolution from C to C' channels. Evaluate for $K = 3$, $C' = 256$.

Problem 3: Softmax, cross-entropy and calibration (15 points)

Let $p_k = e^{s_k} / \sum_j e^{s_j}$ and $\mathcal{L} = -\log p_y$.

(a) (6) Prove that $\partial\mathcal{L}/\partial s_k = p_k - \mathbb{1}[k = y]$. Interpret the result.

(b) (3) Show that adding the same constant c to every score leaves p unchanged, and explain why implementations subtract $\max_j s_j$ before exponentiating.

(c) (3) Temperature scaling uses $p_k(T) \propto e^{s_k/T}$. What happens as $T \to 0$ and $T \to \infty$? Does T change the predicted class?

(d) (3) Explain what it means for a classifier to be *calibrated* and how you would check it on a validation set.

Problem 4: Evaluating a detector (15 points)

A test set contains 3 objects g_1, g_2, g_3. A detector outputs, sorted by score: d_1 (0.95, IoU with $g_1 = 0.85$), d_2 (0.90, IoU with $g_1 = 0.60$), d_3 (0.80, IoU with $g_2 = 0.70$), d_4 (0.60, no overlap), d_5 (0.50, IoU with $g_3 = 0.80$). Each ground-truth object can be matched at most once.

(a) (6) Label each detection TP or FP at IoU threshold 0.5, give precision and recall after each, and compute AP (interpolated precision summed over recall steps).

(b) (4) Repeat at IoU threshold 0.75. Explain why AP drops.

(c) (3) Apply NMS with IoU threshold 0.5 to d_1 and d_2, given $\mathrm{IoU}(d_1, d_2) = 0.65$. Which detections remain, and how does that change your answer to (a)?

(d) (2) Why does COCO average AP over IoU thresholds from 0.50 to 0.95?

Problem 5: Self-attention (10 points)

Let $X \in \mathbb{R}^{n \times d}$ hold n tokens and $\mathrm{SA}(X) = \mathrm{softmax}(XW_Q(XW_K)^\top/\sqrt{d})XW_V$, with the softmax applied row by row.

(a) (5) Let P be a permutation matrix. Show that $\mathrm{SA}(PX) = P\,\mathrm{SA}(X)$. What does this imply for images, and how do ViTs fix it?

(b) (3) Give the cost of computing the attention matrix in terms of n and d. For 224×224 images, compare patch sizes 16 and 8.

(c) (2) Name one way vision transformers reduce this cost for high-resolution images.

Problem 6: Mini-project, deep learning on your own data (30 points)

Choose *one* track and write a report section of about 3 pages with figures. Use images you collect yourself, or a public dataset from an African context (cite it); in either case include at least 30 images you photographed yourself in the *test* set.

Track A: Transfer learning for classification. At least 3 classes and 40 images per class.

- Compare: (i) HOG or colour histogram + logistic regression; (ii) frozen pre-trained features + logistic regression; (iii) fine-tuning the last stage of a pre-trained CNN or ViT; (iv) a small CNN trained from scratch.
- Plot test accuracy against the number of training images per class for (ii) and (i).
- Show a confusion matrix and Grad-CAM for correct and incorrect predictions of your best model, and discuss whether the evidence is in the right place.

Track B: Detection or segmentation. Label at least 60 images with boxes (detection) or masks (segmentation), using CVAT, Label Studio or similar (SAM-assisted labelling is allowed).

- Fine-tune a pre-trained model (e.g. a small YOLO for detection, or a U-Net with a pre-trained encoder for segmentation).

- Report AP50 and AP (detection) or IoU and Dice per class (segmentation) on a held-out test set from a different session or location.
- Compare with a simple baseline (pre-trained detector without fine-tuning, or a classical colour/threshold pipeline from Week 1) and show a page of failure cases with your explanation.

Marking (both tracks).

Data collection, labelling and splitting (documented, no leakage)	7
Correct, reproducible implementation (notebook runs top to bottom)	8
Experimental design and honest metrics (baselines, per-class results)	7
Figures, tables and failure analysis	5
Discussion linking results to lectures; LaTeX presentation	3
Total	30