Sie dürfen zu diesem Übungsblatt eine Abgabe machen bis spätestens zum 16.5., damit könnten Sie eine der erforderlichen fünf Abgaben schaffen für die ULP.

Sie dürfen in Gruppen bis max. drei Personen zusammen arbeiten, aber bitte geben Sie die Zusammenarbeit auf der Abgabe an und bitte geben Sie **jede(r) individuell** ab.

Besuchen Sie gerne das Praktikum am 9.5. In einer der nächsten Übungen können wir auch gerne ein „Fachgespräch“ zu diesen Aufgaben führen.

Wir wollen untersuchen, wie stabil (oder umgekehrt „sensibel“) die Klassifikation von MNIST-Bildern mit einem unserer KNNs ist. Konkret schauen wir uns an, ob Änderungen jeweils *eines einzelnen* Bildpunkts dazu führen können, dass ein ursprünglich korrektes Ergebnis falsch wird.

Führen Sie dazu folgendes Experiment durch:

- Trainieren Sie ein Netz für MNIST, das gut funktioniert
- Speichern Sie das Netz mit

  ```
  torch.save( model.state_dict(), ... )
  ```

- Laden Sie das Netz in einem neuen Projekt mit

  ```
  model.load_state_dict( torch.load(...) )
  ```

- Laden Sie die MNIST-Trainingsdaten
- Nutzen Sie einen `DataLoader`, um alle Trainings-Bilder einzeln nacheinander verarbeiten zu können
- Für jedes Bild bestimmen Sie zunächst die Klasse, der Ihr Modell das Bild zuordnet
- Wenn das Bild korrekt klassifiziert wird: Ändern Sie nacheinander jeden einzelnen Pixel, indem Sie z. B. den Wert invertieren: $f(a) = 1 - a$
- Kontrollieren Sie nach jeder einzelnen (der 784 möglichen, unabhängigen) Veränderung, ob Ihr Modell das Bild immer noch richtig klassifiziert
- Bereiten Sie Ihre Ergebnisse auf, z. B.:
  - Wie viele der Trainings-Bilder sind „angreifbar“, welche sind „robust“ in dem Sinne, dass keine einzige der Veränderungen zu einer Falsch-Klassifikation führt?
  - Welche Vertauschungen zwischen Klassen können Sie wie häufig provozieren?
  - Auf die Veränderung welcher Bildpunkte reagiert Ihr Modell besonders häufig?
  - (Sie können diese Auswertungen auch auf eine einzige Klasse beschränken, dann müssen Sie nur ca. 1/10 der Bilder verarbeiten und alles geht schneller)