Generative Diffusion Models

IMPORTANT NOTES:
Study lecture materials at least 1 hour and prepare the questions prior to the tutorial session.
The questions will be discussed in the tutorial session.

1. Why are diffusion models considered more robust than GANs in generative tasks?

   - Training stability: No adversarial min-max game, less mode collapse.
   - Probabilistic modeling: Learn a tractable likelihood approximation.
   - Scalability: Perform well across text, image, audio, and multimodal synthesis.

2. Give two real-world applications of multimodal generative AI and explain why multimodality improves results.

   Image Captioning (e.g., CLIP, Flamingo) which uses both text and image input, leading to more context-aware captions. Video Synthesis (e.g., text-to-video) which combines computer vision and NLP enables richer storytelling than unimodal systems.

   Advantage: Multimodal fusion leverages complementary information streams, yielding richer and more accurate representations than unimodal systems.

3. How can generative AI make phishing attacks more dangerous compared to traditional methods?

   Generative AI enables automated creation of grammatically correct, context-aware, and highly personalized phishing content. It reduces detectable linguistic anomalies, incorporates real-time contextual data, and scales social engineering campaigns, thereby increasing success rates in spear phishing and business email compromise.

4. Explain why FGSM white-box adversarial perturbations are particularly concerning from an AI safety perspective.

   FGSM leverages gradient information to generate perturbations aligned with the model's decision boundary. These perturbations are imperceptible to humans yet induce misclassification with high reliability, exposing the vulnerabilities of deep learning systems and undermining their trustworthiness in safety-critical applications.

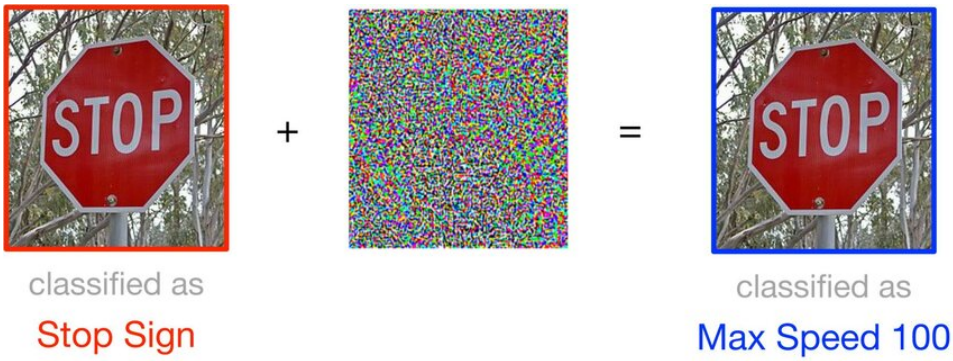
5. A financial institution intends to fine-tune an LLM on proprietary documents.

   (a) What are the risks of data leakage?
   (b) Suggest a mitigation strategy grounded in AI safety principles.

   (a) Risks include model memorization of sensitive data, vulnerability to prompt injection, and potential extraction of proprietary information through adversarial querying.
   (b) Mitigation can be achieved through Differential Privacy (DP-SGD), ensuring that model outputs are statistically invariant to the inclusion or exclusion of specific training records, thereby bounding information leakage.

6. In fraud detection, would a unimodal or multimodal generative AI model be more appropriate? Justify.

   A multimodal approach is preferable, as fraud signals manifest across heterogeneous modalities (e.g., structured transaction logs, textual communications, audio from call centers). Cross-modal fusion enhances robustness, improves detection precision, and reduces false positives relative to unimodal baselines.



7. Autonomous Driving Case: Attackers add adversarial perturbation to a stop sign, causing the self-driving car's vision model to misclassify it as a speed-limit sign.

   (a) Why is this attack effective?

   (b) Suggest two countermeasures.

   (a) Small physical changes shift the model's decision boundary while appearing normal to humans.

   (b) Countermeasures:

   - Data augmentation with perturbed/occluded traffic signs.
   - Robust architectures and redundancy (e.g., fusing camera with LiDAR and GPS signals).

8. Medical Imaging Case: An adversary applies small perturbations to CT scans so a tumor is missed by the AI detection system.

   (a) What risks does this pose?

   (b) Which defense methods could improve robustness?

   (a) Risks include misdiagnosis, delayed treatment, and patient harm. Trust in AI-assisted healthcare is also undermined.

   (b) Defenses include adversarial training with perturbed scans, input preprocessing (denoising), and ensemble models to reduce vulnerability.