Deepfakes II

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. What makes deepfakes difficult to detect, even with advanced AI models?

   Deepfakes are generated to mimic real-world media with high visual fidelity. They often lack global consistency (e.g., mismatched eye colors, lighting inconsistencies) but appear locally realistic. Traditional CNNs focus on local features, missing subtle, distributed artefacts. Vision Transformers (ViTs), with their self-attention mechanisms, can capture these global anomalies, but even they struggle when artefacts are minimal or cleverly disguised.

2. Why is it said that deepfakes cannot be prevented, only detected?

   Like cryptographic attacks, deepfakes exploit generative models to produce synthetic media. Prevention is nearly impossible because the tools (GANs, autoencoders) are publicly available and improving. Instead, detection mechanisms like watermarking, cryptographic signatures, and AI classifiers are used to verify authenticity post-creation.

3. How do artefacts from GAN-generated images differ from those captured by real-world sensors?

   Real images are shaped by physical laws—optical lens effects, spectral responses, and geometric consistency. GAN-generated images, however, are synthesized from random noise and often show artefacts like:

   - Blocking patterns from convolutional upsampling
   - Illumination inconsistencies
   - Geometric distortions

4. A hospital receives a video of a patient requesting a change in medication. The video was flagged by MesoNet but passed EnsembleNet. What should the verification protocol include before acting on the request?

   In such a case, the hospital must proceed cautiously. First, it should verify whether the video contains any embedded watermark or cryptographic signature that confirms its origin. If the video lacks such authentication, it should not be trusted at face value. The hospital should then escalate the video to a more advanced detection system like Vision Transformer (ViT), which can analyze global inconsistencies such as unnatural lighting, facial artefacts, or temporal glitches across frames.

   If the AI models still produce conflicting results or low confidence scores, a human-in-the-loop approach becomes essential. Medical professionals or digital forensic experts should manually review the video, considering contextual factors like the patient's known behavior, medical history, and communication style. This layered verification ensures that decisions affecting patient health are not based on potentially manipulated media.

5. A political video goes viral on social media. It was generated using a GAN and shows inconsistent eye color and lighting. How should a platform respond using the ideal anti-deepfake framework?

   In this scenario, the platform hosting the video must act responsibly. First, it should run the video through a real-time detection layer using models like MesoNet to quickly flag suspicious content. Once flagged, the video should be escalated to advanced analysis using EnsembleNet and ViT. These models can detect deeper inconsistencies, such as lack of temporal coherence or unnatural facial expressions.

If the video lacks a cryptographic signature or watermark from a trusted source (e.g., a news agency), it should be labeled as "potentially manipulated" and temporarily restricted from further sharing. The platform should also notify users that the content is under review. In high-impact cases like political misinformation, external experts or fact-checking organizations should be consulted to validate the content before any conclusions are drawn.

6. A startup wants to build a deepfake detection tool for mobile devices. Which model architecture would you recommend and why?

   A startup aims to develop a mobile app that detects deepfakes in real time. Given the constraints of mobile devices—limited processing power and battery life, the choice of model architecture is critical.

   The most suitable model for this use case is MesoNet, which is designed to be lightweight and fast. It uses a compact convolutional neural network (CNN) architecture that focuses on mesoscopic features—those between microscopic noise and high-level semantics. This makes it ideal for quick detection without heavy computation.

   However, MesoNet may miss subtle global inconsistencies. To address this, the app could offer a cloud-based option where flagged content is sent to a server for deeper analysis using ViT or EnsembleNet. This hybrid approach balances speed and accuracy, allowing users to get instant feedback while still enabling high-confidence verification when needed.