Лекция «Состязательные модели в задачах информационной безопасности»

Инструкции для преподавателей.

**Слайд (Программа модуля).**

В модуле рассказывается про различные технологии применения генеративных (то есть что-то создающих) моделей машинного обучения, в том числе созданных с использованием состязательных подходов, к решению задач информационной безопасности. Отличие от типовых дискриминационных (то есть что-то описывающих) задач: классификация, регрессия.

**Слайд (Обучение без учителя).**

Типовое применение.

* Обработка данных
  + Понижение размерности
  + Сжатие информации
  + Очистка изображений
* Анализ данных
  + Поиск аномалий
  + Кластеризация
  + Обнаружение паттернов
* Генерация данных
  + Создание объектов

**Слайд (Понижение размерности).**

Одной из проблем в машинном обучении является избыточность признаков. Данные с большим количеством признаков сложно визуализировать, для их обработки требуется больше вычислительных ресурсов.

Практические задачи, которые решаются с использованием методов понижения размерности:

* сжатие информации,
* визуализация (например, для кластеризации)
* обнаружение паттернов.

Наиболее популярные методы для решения этой задачи методы: анализ главных компонент (Principal Component Analize, PCA) и TSNE. В результате получаем скрытый (латентный) вектор для каждого экземпляра в датасете, который содержит в себе сжатую информацию об этом экземпляре.

**Слайд (Автокодировщики).**

Автокодировщик это нейронная сеть, имеющая специальную архитектуру, состоит из двух частей: кодировщик и декодировщик.

Сначала модель обучают кодировать и декодировать сигналы из имеющейся выборки, при этом закодированный сигнал имеет гораздо меньшую размерность, чем исходный, поэтому модели приходится учиться «сжимать» информацию.

Автокодировщик, внутренний слой которого больше, чем входной, используется для удаления шумов из сигнала (denoising autoencoder)

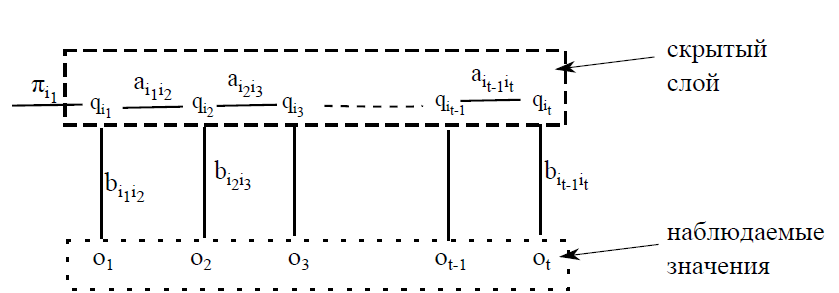
**Слайд (Машина Больцмана).**

Изобретатели: Джеффри Хинтон, Терри Сейновски, 1985 год. Представляет собой стохастическую рекуррентную сеть. Для обучения сети используется алгоритм имитации отжига.

Машина Больцмана с ограничениями (Restricted Boltzman Machine, RBM), используется в сетях глубокого доверия

«Интуитивное основание»: скрытая марковская модель

**Слайд (Скрытая марковская модель).**

****

* наблюдаемый (известный) и скрытый (неизвестный) слои
* a и b – вероятности
* Алгоритм Витерби позволяет установить наиболее вероятную цепочку
* Обратный алгоритм позволяет уточнить значения матриц A (вероятности перехода состояний в скрытом слое) и B (вероятности того, что мы будем наблюдать на внешнем слое)

**Слайд (Генеративные состязательные сети).**

Генеративно-состязательная сеть (Generative adversarial network, GAN) это модель машинного обучения, умеющая создавать искусственные объекты (тексты, изображения, видео) таким образом, что они максимально похожи на их естественные аналоги. Подход GAN был изобретен Яном Гудфеллоу и описан в статье: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>

Основные элементы GAN: генератор и дискриминатор. Процесс обучения:

* Генератор создает фейковые изображения
* Обучаем генератор распознавать фейк (подаем на вход размеченные реальные и фейковые изображения)
* Замораживаем веса дискриминатора
* Учим генератор обманывать дискриминатор (размечаем фейк как реальные изображения)
* Повторяем (долго ☺)
* Закончиться все должно незначительной победой генератора (чем позднее, тем лучше)

**Слайд (CycleGAN).**

Перенос стилей ([CycleGAN](https://github.com/junyanz/CycleGAN/" \t "_blank)) – преобразование одного изображения в соответствии со стилем других изображений (например, картин известного художника);

<https://www.tensorflow.org/tutorials/generative/cyclegan>

**Слайд (StyleGAN).**

Генерация человеческих лиц ([StyleGAN](https://en.wikipedia.org/wiki/StyleGAN)), реалистичные примеры доступны на сайте “This Person Does Not Exist”, https://www.thispersondoesnotexist.com/

**Слайд (Лучшие практики в обработке последовательностей).**

Для обработки последовательностей (текстов, изображений) наилучшие результаты (State-Of-the-Art, SOTA) показали модели трансформеров с использованием механизма внимания (Attention). Отдельные части этой архитектуры получили самостоятельное существование. OpenAI разработал сеть под названием Generative Pre-trained Transformer (GPT), которая использовала модифицированый декодер трансформера. Затем Google создал Bidirectional Encoder Representations from Transformers (BERT), используя энкодер трансформера. Помимо трансформера, новые модели объединяет стратегия обучения на большом корпусе неразмеченных текстов.  
GPT предсказывает очередное слово текста, а BERT - "закрытые" слова внутри предложения. В результате такого обучения формируется языковая модель, включающая в себя грамматику, семантику и даже определённые знания. После предварительного обучения производится тонкая настройка параметров модели под конкретную задачу уже на размеченных данных.

<https://medium.com/techvariable/transformers-and-gpt-3-f7bbc6df2c29>

<https://habr.com/ru/post/341240/>

**Слайд (Применение GAN).**

* Создание наборов данных
* Аватары
* Видеоклипы
* Создание произведений искусства
* Стилизация (нарисовать фотографию в стиле Ван Гога)
* DeepFake
* …
* До конца возможности не изучены (новая технология) ☺

**Слайд (DeepFake).**

Deepfake (дипфейк): «глубинное обучение» ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) deep learning) и «подделка» ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) fake), методика синтеза изображения, основанная на [искусственном интеллекте](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82). Она используется для соединения и наложения существующих изображений и видео на исходные изображения или видеоролики. Для DeepFake часто используют генеративно-состязательные нейросети (GAN).

Deepfake может быть использован для замены определённых элементов изображения на желаемые образы, в частности для создания фальшивых [видео](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D1%80%D0%BD%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F) со знаменитостями. Deepfake может быть использован для создания поддельных новостей и вредоносных обманов.

Deepfake-ролики можно легко найти на популярных сайтах потокового видео, таких как [YouTube](https://ru.wikipedia.org/wiki/YouTube" \o "YouTube) или [Vimeo](https://ru.wikipedia.org/wiki/Vimeo" \o "Vimeo).

Методы выявления движения и превращения в целевое видео, которое похоже на целевой образ, были представлены в 2016 году и позволяют создавать поддельные мимические изображения в существующем 2D-видео в режиме реального времени. 1 сентября 2020 года компания Microsoft анонсировала свою новую разработку – Microsoft Video Authenticator, программное обеспечение, которое позволяет определить вмешательство в видео.

**Слайд (Применение генеративно-состязательных сетей в кибербезопасности).**

Как и любой инструмент, модели машинного обучения, и GAN в частности, могут применяться как со стороны атаки, так и со стороны обороны.

**Слайд (Генерация фейковых сообщений).**

* Текстовые сообщения, максимально приближенные по стилю к имитируемому автору
* Использование для фишинга
* Цель: спровоцировать объект атаки (собеседника) на нарушение правил информационной безопасности (клик на ссылку, вредоносный файл, неправомерные или ошибочные действия)
* Сочетание с методами социальной инженерии

**Слайд (DeepFake).**

* Deepfake (дипфейк): «deep learning» + fake
* Синтез изображения, используется ИИ
* Часто используют генеративно-состязательные нейросети (GAN).
* Поддельные видео, новости и вредоносные обманы.
* Ролики на YouTube или Vimeo.
* Есть ПО для распознавания факта вмешательства в видео, например, Microsoft Video Authenticator

**Слайд (Атаки на модели машинного обучения).**

* Извлечение: кража и обратный инжиниринг модели
* Уклонение: незначительное изменение значений признаков для обмана модели
* Отравление данных: манипуляция входными данными для изменения логики работы модели

**Слайд (Генерация фишинга).**

* Используется персонифицированная информация об объекте атаки, позволяющая сделать фишинговое письмо эффективным
* Контент и характеристики фишингового письма модифицируются для сокрытия природы фишинга
* Внешний вид фейкового сайта максимально похож на оригинал внешне (для человеческого глаза), но имеет иную структуру, чем оригинал (невозможно обнаружить похожесть оригинала и фейка автоматизированными методами)