Приложение к ОПОП «Безопасность и этика искусственного интеллекта»

СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

МИНОБРНАУКИ РОССИИ

федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И.Ульянова (Ленина)»
(СПбГЭТУ «ЛЭТИ»)

РАБОЧАЯ ПРОГРАММА

дисциплины

«АТАКИ НА НЕЙРОННЫЕ СЕТИ»

для подготовки магистров

по направлению

09.04.01 «Информатика и вычислительная техника»

по программе

«Безопасность и этика искусственного интеллекта»

Санкт-Петербург

2023

ЛИСТ СОГЛАСОВАНИЯ

Разработчики:

доцент, к.т.н. Казимиров А.С.

Рабочая программа рассмотрена и одобрена на заседании кафедры ВТ 02.09.2021, протокол № 6

Рабочая программа рассмотрена и одобрена учебно-методической комиссией ФКТИ, 16.09.2021, протокол № 6

Согласовано в ИС ИОТ

Начальник ОМОЛА Загороднюк О.В.

1 СТРУКТУРА ДИСЦИПЛИНЫ

Обеспечивающий факультет	ФКТИ
Обеспечивающая кафедра	ИС
Общая трудоемкость (ЗЕТ)	2
Курс	2
Семестр	3
Виды занятий	
Лекции (академ. часов)	8
Практические занятия (академ. часов)	9
Иная контактная работа (академ. часов)	1
Все контактные часы (академ. часов)	18
Самостоятельная работа, включая часы на контроль (академ. часов)	54
Всего (академ. часов)	72
Вид промежуточной аттестации	
Дифф. зачет (курс)	2

2 АННОТАЦИЯ ДИСЦИПЛИНЫ

«АТАКИ НА НЕЙРОННЫЕ СЕТИ»

В дисциплине рассматриваются теоретические основы атак на нейронные сети с целью обмана моделей с помощью специально подобранных входных данных. В курсе используются библиотеки построения нейронных сетей PyTorch и TensorFlow, а также библиотека генерации атак FoolBox.

SUBJECT SUMMARY

«ADVERSARIAL ATTACKS ON NEURAL NETWORKS»

This course concerns adversarial attacks on neural networks—technique that attempts to fool models by supplying deceptive input. Course uses Python frameworks for training neural networks such as PyTorch and Tensorflow and a special library for generating adversarial attacks—Foolbox.

3 ОБЩИЕ ПОЛОЖЕНИЯ

3.1 Цели и задачи дисциплины

1. Целью дисциплины является рассмотрение различных видов атак на модели машинного обучения, основанные на нейронных сетях и приобретение навыков использования полученных знаний по предотвращению атак в профессиональной деятельности. Способность разрабатывать и модернизировать программное и аппаратное обеспечение технологий и систем искусственного интеллекта с учетом требований информационной безопасности в различных предметных областях.

2. Задачами дисциплины является изучение различных видов атак на нейронные сети. Приобретение навыков защиты от атак на нейронные сети с целью обмана моделей с помощью специально подобранных входных данных

3. Дисциплина формирует знания об основных типах атак на нейронные сети и защите от них.

4. Дисциплина формирует умения по анализу нейронных сетей на подверженность атакам и защите от них. Разрабатывать программное и аппаратное обеспечение технологий и систем искусственного интеллекта для решения профессиональных задач с учетом требований информационной безопасности в различных предметных областях.

5. Результатом освоения дисциплины является приобретение навыков работы с библиотеками FoolBox и SecML для анализа нейронных сетей на подверженность распространенным типам атак. Модернизировать программное и аппаратное обеспечение технологий и систем искусственного интеллекта для решения профессиональных задач с учетом требований информационной безопасности в различных предметных областях.

3.2 Место дисциплины в структуре ОПОП

Дисциплина изучается на основе ранее освоенных дисциплин учебного плана:

1. «Введение в нейронные сети»

2. «Машинное обучение»

и обеспечивает подготовку выпускной квалификационной работы.

3.3 Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения образовательной программы обучающийся должен достичь следующие результаты обучения по дисциплине:

Код компетенции/ индикатора компетенции	**Наименование компетенции/индикатора компетенции**
ПК-30	Способен разрабатывать и модернизировать программное и аппаратное обеспечение технологий и систем искусственного интеллекта с учетом требований информационной безопасности в различных предметных областях
ПК-30.1	*Разрабатывает программное и аппаратное обеспечение технологий и систем искусственного интеллекта для решения профессиональных задач с учетом требований информационной безопасности в различных предметных областях*
ПК-30.2	*Модернизирует программное и аппаратное обеспечение технологий и систем искусственного интеллекта для решения профессиональных задач с учетом требований информационной безопасности в различных предметных областях*

4 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Содержание разделов дисциплины

4.1.1 Наименование тем и часы на все виды нагрузки

№ п/п	Наименование темы дисциплины	Лек, ач	Пр, ач	ИКР, ач	СР, ач
1	Введение в состязательное машинное обучение	2	1		
2	Атаки с уклонением	2	3		20
3	Атаки с отравлением данных	2	3		20
4	Способы защиты от атак	2	2	1	14
	Итого, ач	8	9	1	54
	Из них ач на контроль	0	0	0	0
	Общая трудоемкость освоения, ач/зе	72/2			

4.1.2 Содержание

№ п/п	Наименование темы дисциплины	Содержание
1	Введение в состязательное машинное обучение	Рассматриваются примеры атак на нейронные сети для различных классов задач. Примеры атак для неправильной классификации изображений и детекции объектов. Примеры атак на спам-фильтры.
2	Атаки с уклонением	Примеры атак с уклонением. White-box и black-box атаки. Атаки с уклонением на ImageNet с использованием библиотек SecML и FoolBox.
3	Атаки с отравлением данных	Примеры атак с отравлением данных. Атаки, основанные на градиенте. Атаки с отравлением на глубокие нейронные сети. Атаки с отравлением данных с использованием библиотек SecML и FoolBox.
4	Способы защиты от атак	Детектирование и отклонение атак. Робастная оптимизация. Deep Neural Rejection с использованием библиотеки SecML.

4.2 Перечень лабораторных работ

Лабораторные работы не предусмотрены.

4.3 Перечень практических занятий

Наименование практических занятий	Количество ауд. часов
1. Введение в состязательное машинное обучение	1
2. Атаки с уклонением	3
3. Атаки с отравлением данных	3
4. Способы защиты от атак	2
Итого	9

4.4 Курсовое проектирование

Курсовая работа (проект) не предусмотрены.

4.5 Реферат

Реферат не предусмотрен.

4.6 Индивидуальное домашнее задание

Индивидуальное домашнее задание выполняется по 2 темам дисциплины:

1) Атаки с уклонением

2) Атаки с отравлением данных

Обучающиеся получают задание от преподавателя, выполняют его самостоятельно, затем сдают на проверку. ИДЗ оценивается по системе ”зачтено” / ”не зачтено”. Если задание выполнено правильно ИДЗ ставится оценка ”зачтено”.

4.7 Доклад

Доклад не предусмотрен.

4.8 Кейс

Кейс не предусмотрен.

4.9 Организация и учебно-методическое обеспечение самостоятельной работы

Изучение дисциплины сопровождается самостоятельной работой студентов с рекомендованными преподавателем литературными источниками и информационными ресурсами сети Интернет.

Планирование времени для изучения дисциплины осуществляется на весь период обучения, предусматривая при этом регулярное повторение пройденного материала. Обучающимся, в рамках внеаудиторной самостоятельной работы, необходимо регулярно дополнять сведениями из литературных источников материал, законспектированный на лекциях. При этом на основе изучения рекомендованной литературы целесообразно составить конспект основных положений, терминов и определений, необходимых для освоения разделов учебной дисциплины.

Особое место уделяется консультированию, как одной из форм обучения и контроля самостоятельной работы. Консультирование предполагает особым образом организованное взаимодействие между преподавателем и студентами, при этом предполагается, что консультант либо знает готовое решение, которое он может предписать консультируемому, либо он владеет способами деятельности, которые указывают путь решения проблемы.

Самостоятельное изучение студентами теоретических основ дисциплины обеспечено необходимыми учебно методическими материалами (учебники, онлайн-версия курса), выполненными в печатном или электронном виде.

По каждой теме содержания рабочей программы могут быть предусмотрены индивидуальные домашние задания (расчетнографические работы, рефераты, конспекты изученного материала, доклады и т.п.).

Изучение студентами дисциплины сопровождается проведением регулярных консультаций преподавателей, обеспечивающих практические занятия по-

дисциплине, за счет бюджета времени, отводимого на консультации (внеаудиторные занятия, относящиеся к разделу «Самостоятельные часы для изучения дисциплины»).

Текущая СРС	**Примерная трудоемкость, ач**
Работа с лекционным материалом, с учебной литературой	15
Опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	0
Самостоятельное изучение разделов дисциплины	0
Выполнение домашних заданий, домашних контрольных работ	15
Подготовка к лабораторным работам, к практическим и семинарским занятиям	14
Подготовка к контрольным работам, коллоквиумам	0
Выполнение расчетно-графических работ	0
Выполнение курсового проекта или курсовой работы	0
Поиск, изучение и презентация информации по заданной проблеме, анализ научных публикаций по заданной теме	0
Работа над междисциплинарным проектом	0
Анализ данных по заданной теме, выполнение расчетов, составление схем и моделей, на основе собранных данных	0
Подготовка к зачету, дифференцированному зачету, экзамену	10
ИТОГО СРС	54

5 Учебно-методическое обеспечение дисциплины

5.1 Перечень основной и дополнительной литературы, необходимой для освоения дисциплины

№ п/п	Название, библиографическое описание	К-во экз. в библ.
Основная литература		
1	Ян Пойнтер Программируем с PyTorch: Создание приложений глубокого обучения [Электронный ресурс] / Пойнтер Ян, 2021. -256 с.	неогр.
2	Брайан Макмахан Знакомство с PyTorch: глубокое обучение при обработке естественного языка [Электронный ресурс] / Макмахан Брайан, Рао Делип, 2021. -256 с.	неогр.
3	Шакла Нишант Машинное обучение и TensorFlow [Электронный ресурс] / Нишант Шакла, 2019. -336 с.	неогр.
Дополнительная литература		
1	Элбон Крис Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. [Электронный ресурс] / Крис Элбон, 2019. -384 с.	неогр.
2	Сет Вейдман Глубокое обучение: легкая разработка проектов на Python [Электронный ресурс] / Вейдман Сет, 2021. -272 с.	неогр.

5.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет», используемых при освоении дисциплины

№ п/п	Электронный адрес
1	Библиотека FoolBox. URL:https://foolbox.readthedocs.io/
2	Библиотека SecML. URL:https://secml.gitlab.io/

5.3 Адрес сайта курса

Адрес сайта курса: https://vec.etu.ru/moodle/course/view.php?id=7664

6 Критерии оценивания и оценочные материалы

6.1 Критерии оценивания

Для дисциплины «Атаки на нейронные сети» предусмотрены следующие формы промежуточной аттестации: зачет с оценкой.

Зачет с оценкой

Оценка	Описание
Неудовлетворительно	Курс не освоен. Студент испытывает серьезные трудности при ответе на ключевые вопросы дисциплины
Удовлетворительно	Студент в целом овладел курсом, но некоторые разделы освоены на уровне определений и формулировок теорем
Хорошо	Студент овладел курсом, но в отдельных вопросах испытывает затруднения. Умеет решать задачи
Отлично	Студент демонстрирует полное овладение курсом, способен применять полученные знания при решении конкретных задач.

Особенности допуска

Для допуска к экзамену студент должен успешно выполнить и защитить ИДЗ.

6.2 Оценочные материалы для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине

Вопросы к дифф.зачету

№ п/п	Описание
1	Основные архитектуры нейронных сетей.
2	Состязательное машинное обучение. Примеры атак.
3	Классификация атак на модели машинного обучения по целям, времени и имеющейся информации.
4	Атака с уклонением. Способы осуществления и примеры.
5	Способы защиты от атак с уклонением.
6	Атака с отравлением данных. Способы осуществления и примеры.
7	Способы защиты от атак с уклонением.

Весь комплект контрольно-измерительных материалов для проверки сформированности компетенции (индикатора компетенции) размещен в закрытой части по адресу, указанному в п. 5.3

6.3 График текущего контроля успеваемости

Неделя	Темы занятий	Вид контроля
5	Атаки с уклонением	
6		
7		
8		ИДЗ / ИДРГЗ / ИДРЗ
10	Атаки с отравлением данных	
11		
12		ИДЗ / ИДРГЗ / ИДРЗ

6.4 Методика текущего контроля

на лекционных занятиях

Текущий контроль включает в себя контроль посещаемости (не менее ***80*** % занятий).

на практических занятиях

Текущий контроль включает в себя контроль посещаемости (не менее ***80*** % занятий).

В ходе проведения практических занятий целесообразно привлечение студентов к как можно более активному участию в дискуссиях, решении задач, обсуждениях и т. д. При этом активность студентов также может учитываться преподавателем, как один из способов текущего контроля на практических занятиях.

самостоятельной работы студентов

Контроль самостоятельной работы студентов осуществляется на лекционных, лабораторных и практических занятиях студентов по методикам, описанным выше.

7 Описание информационных технологий и материально-технической базы

Тип занятий	Тип помещения	Требования к помещению	Требования к программному обеспечению
Лекция	Лекционная аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, компьютер или ноутбук, проектор, экран, маркерная доска.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Практические занятия	Аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, компьютер или ноутбук, проектор, экран, маркерная доска.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Самостоятельная работа	Помещение для самостоятельной работа	Оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше

8 Адаптация рабочей программы для лиц с ОВЗ

Адаптированная программа разрабатывается при наличии заявления со стороны обучающегося (родителей, законных представителей) и медицинских показаний (рекомендациями психолого-медико-педагогической комиссии). Для инвалидов адаптированная образовательная программа разрабатывается в соответствии с индивидуальной программой реабилитации.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Дата	Изменение	Дата и номер протокола заседания УМК	Автор	Начальник ОМОЛА
1	14.02.2023	Программа актуальна, изменения не требуются	14.02.2023, протокол заседания УМК №2	доцент, к.т.н., А.С. Казимиров	