Preferred Networks インターンシップ 2022 テーマ別課題

この課題ではインターンのテーマに沿った専門知識の理解度を問います。 あなたが選んだ第一希望テーマに相当する設問を以下の中から選び、その解答を提出して下さい。

設問が記述形式のものである場合は、`survey.pdf` というファイル名で A4 サイズの PDF にまとめて下さい。 ページ制限、様式については各設問で指定された形式に従って下さい。 特に指定がない場合はページ数は2枚以内（参考文献の節を含む）、様式は自由であるとします。

設問が記述形式以外である場合は、各設問で指定された形式で提出してください。 `zip` 形式での提出が指定された場合は、`survey.zip` というファイル名で提出してください。

以下のテーマについてはテーマ別課題を課していません。コーディング課題のみ受験してください。

- JE03. 汎用原子レベルシミュレータ Matlantis における Web システムの開発、運用
- JE05. 創薬研究を加速するライブラリ・UI開発
- JP11. 小売店向け画像認識システムに関連するソフトウェア・アプリケーション開発
- JP12. 小売業における業務課題解決・最適化

諸注意

課題には自分だけで取り組んでください。この課題を他の応募者を含めた他人と共有・相談することを禁止します。 **GitHub の公開リポジトリや SNS 等に解答や問題をアップロードする行為も禁止します。** (選考期間終了後、インターンシップ運営が問題を公開する可能性があります。インターンシップ運営が問題を公開した後は、解答を公開したりしていただいても構いません。) 漏洩の証拠が見つかった場合、その応募者は失格となります。ある応募者が別の応募者に回答をコピーさせた場合、双方の応募者が失格となります。

想定所要時間はコーディング課題とテーマ別課題の両方を合わせて最大2日です。全課題が解けていなくても提出は可能ですので、学業に支障の無い範囲で取り組んで下さい。

変更履歴

2022年4月27日: 初版

JE01. 深層学習モデルを社会実装するためのフレームワーク、ライブラリ開発

`je01` フォルダ内の pdf (`je01/README-ja.pdf`) を参照してください。

JE02. 深層学習およびシミュレーション向けのストレージ技術の開発と最適化

Linuxにおいて、以下のどちらかまたは両方の過程を性能上問題になりうる点を踏まえてなるべく詳しく解説してください。ファイルシステムやストレージデバイスは説明しやすいものを適宜選択してください。

- read(2)において、システムコールが始まってストレージメディア上のバイト列がユーザー空間にコピーされて制御が戻るまで
- write(2) によってユーザー空間のメモリ上のデータがストレージメディアに記録されて制御が戻るまで

JE03. 汎用原子レベルシミュレータ Matlantis における Web システムの開発、運用

（本テーマはテーマ別課題を設けておりません。）

JE04. Optuna 開発

（非公開）

JE05. 創薬研究を加速するライブラリ・UI開発

（本テーマはテーマ別課題を設けておりません。）

JE06. Matlantis向けの物性値計算アルゴリズムの開発

以下の最適化問題を考えてもらうことが課題になります。また、最適化すべき関数は配布するjupyter notebook内 (`je06/JE06_task.ipynb`) に記載されていますので、そちらを参照して解答をしてください。

なお、もし本課題よりも"JE19. 材料に関する機械学習や原子シミュレーションの開発・応用研究"の課題のほうが取り組みやすいと感じた場合は、そちらの課題への解答で代替することも認めます。

問題

3次元空間中に64個のラベルづけされた点を配置します。点の位置は重なることはないものとします。ラベルは整数で、0から7までの値を持つ点がそれぞれ8個ずつあります。この点の集合に対してスカラー値を返すよ

うな関数を考えます。このとき、返り値がなるべく小さくなるような点の配置を考えてください。 この課題は、原子構造を推定する問題を模しています。点を原子、スカラー値をエネルギーとみなすことで、エネルギーが低く安定となるような原子配置を考える問題とみなすことができます。

解答方法

結果(入力の組とそのときのスコア)とどのような考え方をしたかを記したレポート(PDFでA4用紙1枚以下程度、1枚に書き切れない場合は多少オーバーしても構いません)と実行に使ったプログラムのソースコードの2点を提出してください。プログラムは配布のJupyter Notebookをそのまま使っても問題ありませんが、別の形式でも構いません。 この課題では、長い時間をかけてスコアをぎりぎりまで詰める必要はありません。その代わり、見通しのよいプログラムを心がけてください。また、レポートではどのようなことを考えて最適化を行ったのかを書くようにしてください。

JE07. クリエイティブツールの開発

（非公開）

JE08. 学習環境と推論環境の違いに対応するための研究

あなたは、とある機械学習系の会議の査読を引き受けました。以下の論文の中から、あなたが面白いと思うものを任意に1つ選び、擬似的な査読コメントを書いてください。

Step1. 論文を選ぶ

以下の国際会議で出版済みのもののなかから、論文をひとつ選んでください。

- ICML 2021 https://proceedings.mlr.press/v139/

Step2. 模擬査読コメントを書く

選んだ論文に対する査読コメントを書いてください。査読コメントでは、下記の設問に答えてください。

(1) 選んだ論文のタイトル、著者、URL (例: https://proceedings.mlr.press/v139/abdolshah21a.html)
(2) この論文の貢献内容およびそのインパクトについて、サマリーを記述してください。
(3) この論文の強みを3つ挙げてください。
(4) この論文の弱みを3つ挙げてください。
(5) 詳細なコメントを自由に記述してください。

これまでの設問をふまえつつ、論文に対する中立的な批評を行ってください。 accept/rejectの形式での結論や、著者に対する質問を書いたりする必要はありません。 A4用紙で1ページまたは2ページのPDFファイルを提出してください。フォントサイズは10pt以上としてください。

JE09. 欠損を含む表データに対する深層学習

The task is given in the English language but applicants can answer either in Japanese or English. The maximum page length is **two pages** in A4 format (the minimum font size is 10 pts.). We only accept the submission in .pdf format. It is allowed to add figures and tables.

The purpose of this thematic task is to evaluate your familiarity with this topic and your overall research skill. There are 4 questions. We highly encourage you to provide some references (e.g., conference/journal papers) to support your answers.

Here, we provide a list of references that could be useful (it is allowed to also cite the papers that are not listed here):

- Van Buuren, S. (2018). Flexible imputation of missing data. CRC press. (free online access book: https://stefvanbuuren.name/fimd/)
- Yoon, J., Jordon, J., & Schaar, M. (2018, July). GAIN: Missing data imputation using generative adversarial nets. In International conference on machine learning (pp. 5689-5698). PMLR. (https://proceedings.mlr.press/v80/yoon18a.html)
- Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling incomplete heterogeneous data using vaes. Pattern Recognition, 107, 107501. (https://arxiv.org/abs/1807.03653)
- Gondara, L., & Wang, K. (2018, June). Mida: Multiple imputation using denoising autoencoders. In Pacific-Asia conference on knowledge discovery and data mining (pp. 260-272). Springer, Cham. (https://arxiv.org/abs/1705.02737)
- Tashiro, Y., Song, J., Song, Y., & Ermon, S. (2021). CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. Advances in Neural Information Processing Systems, 34. (https://arxiv.org/abs/2107.03502)
- Wang, Z., Akande, O., Poulos, J., & Li, F. (2021). Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison. arXiv preprint arXiv:2103.09316. (https://arxiv.org/abs/2103.09316)
- Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., & Acharya, A. (2019). MICE vs PPCA: Missing data imputation in healthcare. Informatics in Medicine Unlocked, 17, 100275. (https://www.sciencedirect.com/science/article/pii/S2352914819302783)

Question 1

Suppose we want to train a machine learning model to predict urinary sugar level (尿糖), which has five different levels (1-5: discrete from normal to highly unsafe) from available health information of a person. In this problem, urinary sugar level 1 is considered normal. From level 2, it can be considered as abnormal.

The dataset for training the model consists of the health information of 10,000 people collected from several sources in a tabular format: (1) 1000 elderly people from nursing homes, (2) 3000 students from universities, (3) 5000 patients from hospitals, and (4) 1000 players from soccer clubs.

For each person i, we are given the following information:

1. Feature vector: 100-dimensional feature vector x_i represents a person's health information: systolic blood pressure, diabolic blood pressure, height, weight, age, etc. Each feature can be either in a real-valued or categorical format. However, in our problem, some people may not have all 100 information we need, that is, there are missing values. For example, it is observed when collecting the data that 20% of patients in hospitals are not comfortable providing their age information. But all university students provide the age information. Another example we found is 80% of elderly

people had taken a diabetes screening test before, while none of the university students took it. Nevertheless, it is guaranteed that everyone provides height, weight, blood pressure, blood sugar level, and cholesterol level.
2. Source label: $s_i \in$ {nursing homes, universities, hospitals, soccer clubs}, which indicates where the data is collected from. It is guaranteed that s_i is not missing for any person i. Thus, it is possible to identify the data source for each person.
3. Ground truth label: urinary sugar level $y_i \in \{1,2,3,4,5, \text{NaN}\}$, where NaN indicates that the ground truth is missing. It is observed that 70% of the training data contain the target label (urinary sugar level).

Given the scenario above, please answer the following questions:

Q1.1: What do you think are the two biggest technical challenges to use machine learning for solving this problem? Please briefly describe them with a few sentences. For example, applicants can pick challenges from the following list below. It is also allowed to choose other challenges that are not listed here.

- Missing features
- Data size difference
- Domain mismatch
- Label imbalance

Q1.2: Which type of missing feature is the closest one in this scenario and why do you think it is the case?

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR) (aka., Not missing at random (NMAR))

More information about the definitions of MCAR, MAR, and MNAR can be found here:

- https://stefvanbuuren.name/fimd/sec-MCAR.html
- https://www.ncbi.nlm.nih.gov/books/NBK493614/

Question 2

Should data preprocessing by imputing missing features always improve the performance for supervised classification? Please express your thoughts with an example.

Question 3

Describe the advantages and disadvantages of the following preprocessing methods by feature imputation for the supervised learning task (e.g., classification, regression):

- Discarding data points (i.e., a row in a DataFrame of pandas) that contains missing features and use only data points that have all features available.
- Discarding features (i.e., columns) that have missing values.
- Imputing the missing features with mode values of each categorical feature.
- Imputing the missing features with median values for each real-valued feature.
- Imputing the missing features with mean values of each real-valued feature.

Question 4

Discuss the weaknesses of tree-based methods (e.g., gradient boosting, random forest, decision tree) in tabular data and how deep learning could be promising to alleviate such weaknesses.

JE10. 医用画像を対象とした機械学習手法に関する研究

以下の3つの課題のうち1つを選び、A4用紙2ページ以内のpdfで提出してください。フォントサイズは10pt以上としてください。なお、医用画像はX線画像やCT/MRI画像、病理画像など臨床で用いられる画像全般を指します。

- 深層学習モデルが推論を行った際に、推論結果をどれくらい信頼できるのか（確信度）を定量的に測れることは臨床の場において有用です。一方で対策を取らなければモデルの確信度は「自信過剰」になる傾向があるため、対策として確信度の校正 (calibration) 手法等が研究されています。モデルの推論における確信度・不確実性の計算手法に関する論文を1つ選び、内容を簡潔に要約してください。また、他の手法と比較したその論文の強みと弱みについて説明してください。論文を選ぶ際に以下のレビュー論文を参考にしていただいても構いません。 Abdar et al, “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges”, https://arxiv.org/abs/2011.06225
- 近年、一般画像においてデータ効率化を目指した生成モデル活用の研究としてDatasetGAN*などが登場し、大きな注目を集めています。医用画像解析においては、アノテーションコストが高いためにアノテーション済みのデータを収集するのが難しいという課題があり、データ効率化に向けたアプローチは重要な研究テーマの一つとなっています。本課題では医用画像解析の効率化につながるような画像生成技術を用いている論文をDatasetGAN以外で1つ選び、内容を簡潔に要約してください。また、他の手法と比較したその論文の強みと弱みについて説明してください。 *Zhang et al, “DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort”, https://arxiv.org/abs/2104.06490
- 医用画像解析において、訓練データと異なる施設・撮像条件で撮られた画像に対してモデルの精度が下がりやすいことが知られています。特に病理画像においては染色条件の違いが汎用性の高いモデルを組む上での大きな課題となっています。このような医用画像における domain adaptation の課題を解決するための手法に関する論文を1つ選び、内容を簡潔に要約してください。また、他の手法と比較したその論文の強みと弱みについて説明してください。論文を選ぶ際に以下のレビュー論文を参考にしていただいても構いません。 Guan et al, “Domain Adaptation for Medical Image Analysis: A Survey”, https://arxiv.org/abs/2102.09508

JE11. リモートセンシングデータに対する画像解析・超解像

画像解析、デノイジング、超解像などに関する2019年以降の論文を一つ選んで、以下の観点から論文についてレポート・評価してください。対象とする画像は、Synthetic Aperture Radar(SAR) imageであるのが望ましいが、それに限られない。

i) どのような背景でこの研究が必要となったか
ii) この研究が解決した問題は何か
iii) どのような方法で解決したか
iv) この論文の強み
v) この論文の弱み

JE12. 動画からの深度推定

Self-supervised learningによる深度推定に関する2019年以降の論文を一つ選んで、以下の観点から論文についてレポート・評価してください。

i) どのような背景でこの研究が必要となったか
ii) この研究が解決した問題は何か
iii) どのような方法で解決したか
iv) この論文の強み
v) この論文の弱み

JE13. 気象学の物理知識を利用した気象予測モデルの学習

2021年のNatureの論文”Skilful precipitation nowcasting using deep generative models of radar”(https://www.nature.com/articles/s41586-021-03854-z) について以下の観点からレポート・評価してください。

i) どのような背景でこの研究が必要となったか
ii) この研究が解決した問題は何か
iii) どのような方法で解決したか
iv) この論文の強み
v) この論文の弱み

JE14. センサデータに対する深層学習圧縮モデルの研究開発

以下の論文のうち、一つ、好きなものを選んでその内容をレポート・評価してください。レポートには以下の要件を含めてください。

i) どのような背景でこの研究が必要となったか
ii) この研究が解決した問題は何か
iii) どのような方法で解決したか
iv) この論文の強み
v) この論文の弱み
vi)この論文の技術を、論文で検証していない他のセンサ・ドメインデータに応用したい。このとき、適したセンサ・ドメイン、適さないセンサ・ドメインとして何があるか？なぜそう考えたか？

言語：日本語でも英語でも可
レポート分量：A4で２～３枚程度

対象となる論文

- Zamir+, “Restormer: Efficient Transformer for Hig-resolution Image Restoration”, CVPR 2022. https://arxiv.org/abs/2111.09881

- Hu+, “FVC: A New Framework towards Deep Video Compression in Feature space”, CVPR 2021. https://arxiv.org/abs/2105.09600v2
- Cheng+, “Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules”, CVPR 2020. https://openaccess.thecvf.com/content_CVPR_2020/html/Cheng_Learned_Image_Compression_With_Discretized_Gaussian_Mixture_Likelihoods_and_Attention_CVPR_2020_paper.html
- Chen+, “Simple Baselines for Image Restoration”, https://arxiv.org/abs/2204.04676
- Cohen and Welling, “Steerable CNNs”, ICLR 2017. https://arxiv.org/pdf/1612.08498.pdf
- Ulyanov+, “Deep Image Pror”, CVPR 208. https://dmitryulyanov.github.io/deep_image_prior

JE15. HCI for MLのためのユーザインタフェースの開発

課題1

これまでご自身が取り組んできた下記分野例に該当する研究・開発プロジェクトの概要を説明してください (600-1200字程度、図1,2点)

分野例: Machine Learning, Computer Vision, Human-Computer Interaction, Accessibility, Visualization, Information Retrieval

課題2

以下の1. 2.からひとつを選択して回答してください (1000-2000字程度)

1. 以下の会議で出版された論文の中からHCI for MLに関する論文を1本選択し以下のQ1-Q5に回答をしてください

会議: CHI2019-2021, UIST2019-2021, IUI2019-2022

Q1 どのような背景でこの研究が扱う課題にHCI for MLが必要となったか
Q2 この研究が解決した問題は何か
Q3 課題の解決方法とその特色は何か
Q4 この論文の方法がうまくいくケースとうまくいかないケースは何か
Q5 この論文の不十分な点は何か

2. 以下のQ1-3について回答してください

Q1 HCIはどのように機械学習に貢献できると考えますか。少なくとも3つの例を挙げてください。
Q2 学習データが少量の際の対処方法をいくつか説明してください
Q3 学習データの取得とアノテーションを効率化するための方法をいくつか説明してください

JE16. 多種多様なネットワークにおいてJust-in-Time通信を実現するためのプロトコル開発

末尾の文献リストから論文を1つ選び、以下の点について述べてください。どの文献を選んだかは評価に影響を及ぼしません。

- サービス品質 (QoS) や 体感品質 (QoE) 改善に向けた課題について、文献内で指摘されている点を参考に述べてください。
- 文献内で提案されている手法とその特徴について解説してください。
- 文献内の提案手法の改善点や実展開における課題について、他の文献等で指摘されていたり、広く知られているものがあれば、その文献等を引用して説明してください。他の文献等で述べられていない課題については含めないでください。

文献リスト
[1] M. Palmer et al., "VOXEL: Cross-layer Optimization for Video Streaming with Imperfect Transmission," ACM CoNEXT'21, pp.359--374, 2021
[2] Z. Zheng et al., "XLINK: QoE-driven multi-path QUIC transport in large-scale video services," ACM SIGCOMM'21, pp.418--432, 2021

JE17. 大規模深層学習のための効率的なデータ転送技術の研究開発

末尾の文献リストから論文を1つ選び、以下の点について述べてください。どの文献を選んだかは評価に影響を及ぼしません。

- 分散システム、Disaggregated Computing、データセンタネットワークにおける課題について、文献内で指摘されている点を参考に述べてください。
- 文献内で提案されている手法とその特徴について解説してください。
- 文献内の提案手法の改善点や実展開における課題について、他の文献等で指摘されていたり、広く知られているものがあれば、その文献等を引用して説明してください。他の文献等で述べられていない課題については含めないでください。

文献リスト
[1] J. Min et al., "Gimbal: Enabling Multi-tenant Storage Disaggregation on SmartNIC JBOFs," ACM SIGCOMM'21, pp.106--122, 2021
[2] B. Li et al., "1Pipe: Scalable Total Order Communication in Data Center Networks," ACM SIGCOMM'21, pp.78--92, 2021
[3] S. Abdous et al., "Burst-tolerant datacenter networks with Vertigo," ACM CoNEXT'21, pp.1--15, 2021
[4] K.Liu et al., "Floodgate: taming incast in datacenter networks," ACM CoNEXT'21, pp.30--40, 2021
[5] R. Segal et al., "SOAR: minimizing network utilization with bounded in-network computing," ACM CoNEXT'21, pp.16--29, 2021
[6] Q. Zhang et al., "MimicNet: Fast Performance Estimates for Data Center Networks with Machine Learning," ACM SIGCOMM'21, pp.287--304, 2021

JE18. 創薬に関する機械学習や分子シミュレーションの応用研究

2つの課題に取り組んでいただきます。

(1) あなたが本インターンで取り組もうと思っているトピックに関連した論文(雑誌あるいは学会発表)を1本挙げて、それについて以下の点をふまえてまとめてください。読者は分野に対する基本的な知識を持っていると仮定して構いません。

- 論文(あるいは論文内で紹介されている手法)が解決したい問題

- 過去の解決方法とその欠点（なぜその問題を解決できないか）
- 論文がその問題をどのように解決しているか
- 解決方法の有効性をどのように検証しているか
- 解決方法の限界と考えられる欠点

例として論文をリストアップします。リストから選択しても構いませんし、関連する論文を別個選択して頂いても構いません。特定の論文を選んだことによる評価や選考への影響はありません。

(2) 1) 分子生成モデル、2) 分子プロパティ予測、3) 分子シミュレーションに基づいた結合評価 のいずれかのトピックに関連して、今後10年以内に発展するとあなたが考える分野をあげてください。課題1の論文で解決されていない欠点に注目しても構いませんし、あなた自身の意見を中心に記載しても構いません。課題について現時点で達成されていない原因と、解決の糸口についての2点をふまえてまとめてください。

本課題はインターンシップであなたが取り組む課題を議論する出発点になりますので、特に取り組みたいと思うトピック/論文を選んでください。

レポートにおいては、10pt以上のフォントを用いてPDF形式・A4用紙1ページから2ページ程度で報告をしてください。

論文リスト

- Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design. https://arxiv.org/abs/2110.04624
- Benchmarking Peptide-Protein Docking and Interaction Prediction with AlphaFold-Multimer. https://www.biorxiv.org/content/10.1101/2021.11.16.468810v1
- EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. https://arxiv.org/abs/2202.05146
- Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking. https://arxiv.org/abs/2111.07786
- Equivariant Diffusion for Molecule Generation in 3D. https://arxiv.org/abs/2203.17003
- Sample-Efficient Optimization in the Latent Space of Deep Generative Models via Weighted Retraining. https://arxiv.org/abs/2006.09191
- SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects: https://www.nature.com/articles/s41467-021-27504-0

JE19. 材料に関する機械学習や原子シミュレーションの開発・応用研究

以下のレポート課題に取り組んでいただきます。

原子シミュレーションと機械学習の双方に関わるトピックの論文を1本挙げて、それについて以下の観点からまとめてください。読者は分野に対する基本的な知識を持っていると仮定して構いません。

- 論文の文献情報
- 論文内で引用している先行研究を1本あるいは複数本あげ、それに対する論文の優位性
- 論文でその優位性を出せた理由

レポートにおいては、10pt以上のフォントを用いてPDF形式・A4用紙1ページ程度で報告をしてください。
論文の選択は、もしあなたが本インターンで取り組もうと思っている課題がある場合には関連した論文を挙

げてください。まだトピックが固まってない場合、以下から論文を選ぶことも可能です。特定の論文を選んだからといって、評価が良くなったり悪くなったりすることはありません。

なお、もし本課題よりも ”JE06. Matlantis向けの物性値計算アルゴリズムの開発” の課題のほうが取り組みやすいと感じた場合は、そちらの課題への解答で代替することも認めます。

- D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, Ab Initio Solution of the Many-Electron Schr"odinger Equation with Deep Neural Networks, Phys. Rev. Research 2, 033429 (2020). https://journals.aps.org/prresearch/abstract/10.1103/PhysRevResearch.2.033429
- A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad, Solving the Electronic Structure Problem with Machine Learning, Npj Computational Materials 5, 22 (2019). https://www.nature.com/articles/s41524-019-0162-7
- O. T. Unke and M. Meuwly, PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges, Chem. Theory Comput., 15, 6, 3678–3693 (2019). https://pubs.acs.org/doi/10.1021/acs.jctc.9b00181
- S. Nikolov, M. A. Wood, A. Cangi, J.-B. Maillet, M.-C. Marinica, A. P. Thompson, M. P. Desjarlais, and J. Tranchida, Data-Driven Magneto-Elastic Predictions with Scalable Classical Spin-Lattice Dynamics, Npj Computational Materials 7, 1 (2021). https://www.nature.com/articles/s41524-021-00617-2
- B. Herzog, M. C. da Silva, B. Casier, M. Badawi, F. Pascale, T. Bučko, S. Lebègue, and Dario Rocca, Assessing the Accuracy of Machine Learning Thermodynamic Perturbation Theory: Density Functional Theory and Beyond, J. Chem. Theory Comput. 18 (3), 1382-1394 (2022). https://pubs.acs.org/doi/10.1021/acs.jctc.1c01034
- M. Cools-Ceuppens, J. Dambre, and T. Verstraelen, Modeling Electronic Response Properties with an Explicit-Electron Machine Learning Potential, J. Chem. Theory Comput. 18, 1672 (2022). https://pubs.acs.org/doi/abs/10.1021/acs.jctc.1c00978

JE20. ディープラーニングを使ったクリエイティブへの応用

（非公開）

JE21. 金融分野のための深層学習を用いた研究開発

課題について

以下の課題A、B、Cのいずれか1つを選んで回答してください。 興味・余力があれば、2つ以上の課題に回答しても構いません (2つ以上回答することは、選考には必須ではありません)。

共通の注意事項

- 答案はA4用紙2枚程度を目安としてください。
- 答案の作成において参照した文献やweb上の資料などは、適切に引用するようにしてください。
- プログラムを作成する必要があるときは、使用するプログラミング言語は任意のものでよく、必要に応じて既存のライブラリの機能を使用して構いません。作成したプログラムのソースコードを提出する必要はありませんが、面接等で内容について説明していただく場合があります。

課題A

A-1

以下のポートフォリオ構築手法について、「期待リターン」「共分散行列」というキーワードを使いつつ、簡潔に説明してください。(各5行以内程度が目安)

- 平均分散ポートフォリオ
- 最小分散ポートフォリオ
- リスクパリティポートフォリオ

A-2

3つのリスク資産があり、それぞれの来月の期待リターン `mu` と共分散行列 `cov` を以下のように見積もりました。

```
mu = [0.005, 0.007, 0.004]

cov = [
  [1.0, 0.1, 0.2],
  [0.1, 2.0, 0.6],
  [0.2, 0.6, 1.0]
]
```

設問 A-1 の3種類のポートフォリオのうち、2種類以上を計算するプログラムを作成してください。 定式化や計算手法が一意でない場合は、どのような手法を採用したかについても数式等を利用して説明してください。 上記の `mu`, `cov` の推定に対してポートフォリオを計算し、得られた結果についてわかりやすく説明してください。

課題B

B-1

株式のリターンの系列などでは、ボラティリティ (標準偏差) が時間とともに変動するという性質が経験的に知られています。 多変量時系列のモデルで、各変量のボラティリティおよび変量間の相関行列が時間変動するものをひとつ挙げ、その定式化について説明してください。

B-2

多変量時系列データを人工的に生成するプログラムを作成しましょう。 長さ1000の3変量時系列データ (`1000 * 3` の2次元配列) を生成し、その結果を可視化してください。 このとき、何らかの意味で、ボラティリティおよび変量間の相関が時間変動しているとみなせるようなデータを生成し、そのことがわかるような可視化を行ってください。 また、どのような手法を利用してデータを生成したかについても説明してください。利用する手法は、設問B-1で回答したモデルと同じである必要はありません。

課題C

C-1

以下のキーワードについて簡潔に説明してください。必要であれば数式等を使って構いません。

- インプライド・ボラティリティ
- ボラティリティ・スマイル

C-2

ヨーロピアン・コールオプションの価格を計算しましょう。 まず、ブラック・ショールズモデルのもとでコールオプションの価格を出力するプログラムを作成してください。 次に、以下の設定のもとでオプションの価格を計算し、結果について説明してください。

- 現在の原資産の価格は `26750` 円とする。
- オプションが満期 (限月) を迎えるまでの時間は残り1ヶ月 (`1/12`年) とする。
- 行使価格 `K` は、`[26000, 26500, 27000, 27500, 28000]` の5通りについて計算すること。
- 原資産のボラティリティは年率20%とする。
- 無リスク金利は年率0%とする。

C-3

設問C-2で考えたコールオプションの市場価格 `C` を証券取引所で調べたところ、実際には以下のようになりました。

行使価格K	オプション価格C
26000	1120
26500	760
27000	480
27500	270
28000	130

インプライド・ボラティリティを計算するプログラムを作成してください。 上の市場価格データに対してボラティリティ・スマイルを計算し、得られた結果について考察してください。

JP01. MN-Core向けのコンパイラ及び周辺ライブラリの開発

`jp01` フォルダ内の pdf (`jp01/README-ja.pdf`) を参照してください。

JP02. 次世代MN-Coreのマイクロアーキテクチャ検討

`jp02.pdf` を参照してください。

JP03. エッジとクラウドの両方を考えた実用的なネットワークおよびシステムの設計と実装

以下の2つの課題に回答してください。

課題 1

計算能力をもつカメラデバイス（監視カメラのようなもの）が各地に設置されている。それらのデバイスで得られる画像情報を用いて、各地の人の混雑度をクラウドに集約したい。（混雑度は、カメラに映る人数と、カメラに紐づいた他の情報から算出されるとし、後者は事前に与えられ、変更はされないものとする）

- 画像を受け取って混雑度を返すアルゴリズムの計算量（ただし、画像から混雑度を推定するアルゴリズムは、画像から人数カウントを行うアルゴリズムと人数から混雑度を推定するアルゴリズムに分割してもよい）
- データサイズ
- 通信路の制約
- エッジデバイスの計算能力の制約（エッジデバイスのコストの制約に含まれる）
- クラウドのコストの制約

以上の前提条件を自ら規定した上で、

1. 全ての計算をカメラデバイスで行う場合
2. 一部の計算をカメラデバイスで行い、一部の計算をクラウドで行う場合
3. 全ての計算をクラウドで行う場合

の３つのケースにおけるメリット・デメリットを記述せよ。

課題 2

IoT機器におけるセキュリティの問題点について自由に記述せよ。

JP04. Kubernetesにおけるコンテナ実行環境の改善

以下の課題1,2,3に回答してください。

課題1

インフラに関連する技術として、仮想マシン（VM）やコンテナがあります。VMとコンテナについて、それぞれはどのような技術によって実現されているのか、それぞれにはどのような違いがあるのか、どのような時にVMを使うべきなのか、どのような時にコンテナを使うべきか、の4点について説明してください。

課題2

Kubernetesについて説明してください。この時、Kubernetesを構成する以下のコンポーネントについてその役割、動作、それぞれのコンポーネントの関係性についての説明を含めてください。リストにないコンポーネントの説明を加えても構いません。募集要項に列挙しているテーマ（コンテナ起動時間の高速化、Kubernetes上からFUSEを安全に使える方法の検討、コンテナネットワーキングスタックのリプレース）の中に興味のあるテーマがある場合は、そのテーマに関係するコンポーネントについて他のコンポーネントより詳しく説明してください。

- kube-apiserver
- etcd

- kube-proxy
- kubelet
- コンテナランタイム

課題3

Kubernetes クラスタに HTTP サーバをデプロイするマニフェストファイルを作成してください。デプロイする HTTP サーバは Go 言語を用いて自身で実装してください。

- myapp という名前の Deployment を作成
- Deployment の Pod レプリカ数は 3
- HTTP サーバはポート 8080番でリッスン
- HTTPサーバは GET リクエストに対して text/plain 形式で Pod が実行されているKubernetes ノードの IP アドレスを返す。それ以外の場合は適切は HTTP エラーを返す。

下記ファイル群全体を zip で圧縮し添付ファイルとして提出してください。

- Kubernetes マニフェストファイル
- HTTP サーバのソースコード
- Dockerfile
- 動作確認を行うためのREADMEファイル

JP05. 大規模な機械学習向け計算基盤（インフラ技術）の研究開発

以下の課題1,2,3のうち1つ選んで回答してください。いずれの課題もA4 2枚程度でまとめてください。形式は自由です。図版は見やすいサイズにし、必要なら別ページで添付してください。

課題1：サーベイ課題

ノード内接続技術としてCXLの標準化が進んでいる。CXLのコンセプトを説明し、CXL がクラスタ計算機の設計に及ぼす影響について論ぜよ。

課題2：システム設計課題

クラスタ計算機のシステム設計を実施してください。ネットワーク課題とストレージシステム課題のどちらかを選択できます。

ネットワーク課題

設問1以外はオプションです。意欲のある応募者は挑戦してみてください。

(設問1）

以下のネットワーク接続を持つ256台の計算ノードからなるクラスタ計算機システムを考える。その際に条件を満たしたネットワークシステムを設計しその設計を説明せよ。

- 各計算ノードは400Gbps帯域のイーサネットが2本でネットワークへ接続される
- ネットワークに利用できるネットワークスイッチは最大ポート数32ポート(400Gbps)とする

- 計算ノード間の通信のみを考慮すればよく、外部接続との通信は仮定しない(アップリンクポートはゼロで良い）。

設計に関する説明には全体のネットワーク構成図（トポロジ図など）を含める必要がある。ネットワーク構成図は必要に応じて省略や抽象化をしても良いが、spine-leaf 構成や core-distribtion 構成などの階層構造やスイッチ間のリンク数は明記すること。

(設問2：オプション課題）

計算ノードを接続したネットワークでは下記に示す通信フローが流れるとする。

- 任意の1対1の計算ノード間が任意のタイミングで32GB/sのデータ送信が2秒行われ、その後受信側のデータ処理を待つため1秒は送信されない。
- ワーストケースでは全ノード間でほぼ同時に通信される可能性がある
- （その他、回答に必要な要件があれば自分で定義しても良い）

設問1で設計したネットワークで上記の通信フローを処理しようとした場合に発生する問題があるか考察し、問題があればその問題を説明せよ。また、問題を改善する方針があればそれを提案せよ。

(設問3：オプション課題）

上記のシステムが稼働した1年後および2年後に全計算ノード数がそれぞれ2倍（512ノード)、4倍(1024ノード）に増加した場合のネットワーク構成を提示せよ。ネットワークを拡張する際に考慮した点を説明せよ。

ストレージシステム課題

設問1,2以外はオプションです。意欲のある応募者は挑戦してみてください。

(設問1)

以下の構成のクラスタ計算機を考える。このクラスタ計算機が最大性能を実現するために必要なストレージサブシステムの性能要件を定義し説明せよ。計算ノードとストレージサブシステムは十分な容量を持つ一様な遅延特性のロスレスなネットワークで接続されていると仮定する。

- 計算ノードはそれぞれ8個のアクセラレーターを搭載しており、クラスタ計算機に含まれる計算ノード数は16ノードである。
- それぞれのアクセラレータ最大性能を提供するためには、毎分1回10GBのデータがアクセラレータの近傍に存在する必要がある。
- 各計算ノードは400Gbpsのイーサネットでネットワークに接続される。

設問1ではストレージシステムから計算ノードへの一方向のデータ移動だけを考慮するだけでよい（つまり演算結果をストレージシステムに書き戻す操作は発生しないと仮定しても良い）。

(設問2)

設問1で設定したストレージサブシステムを以下の構成要素を用いて設計し説明せよ。設計の説明には構成要素の台数を明記しその根拠の説明を含めるものとする。

- 平均で100MB/s(Read/Write)でアクセスできるHDD
- 最大で24台までのHDDを搭載できるストレージサーバー。ストレージサーバーは10Gbpsのネットワーク接続を最大2系統搭載可能であるとする。ストレージサーバーに搭載するHDD数やネットワーク接続数は制限の範囲内で自由に設定しても良い。

（設問3：オプション課題）

計算ノードからストレージノードへ10秒毎に1回1GBのデータを書き戻す操作が必要だった場合、設問1および設問2で考慮しなければならない要素を論じよ。

課題3: OS課題

次に示す「JE02. 深層学習およびシミュレーション向けのストレージ技術の開発と最適化 」の課題を解いて、本テーマ（JP05）に応募しても良い。

Linuxにおいて、以下のどちらかまたは両方の過程を性能上問題になりうる点を踏まえてなるべく詳しく解説せよ。ファイルシステムやストレージデバイスは説明しやすいものを適宜選択してよい。

- read(2)において、システムコールが始まってストレージメディア上のバイト列がユーザー空間にコピーされて制御が戻るまで
- write(2) によってユーザー空間のメモリ上のデータがストレージメディアに記録されて制御が戻るまで

JP06. 機械学習・微分可能レンダラを用いた3次元復元

末尾の文献リストの中から一つ選び、以下の点について論じてください。A4で1ページのPDFファイルを提出してください。フォントサイズは10pt以上としてください。英語か日本語のうち、どちらか得意な言語で構いません。

(1) どのような課題を解決しようとしているか
(2) どのような方法により解決しているか
(3) その論文の復元対象物（文献[1]の場合は「動く人間」、文献[2, 3]の場合は「小さめの静止物」）を三次元復元するために、実用上は他にどのような方法（機械学習の活用の有無は問わない）が考えられるか。また、それぞれの方法の典型的な長所と短所はなにか

文献リスト
[1] HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video https://arxiv.org/abs/2201.04127
[2] Modeling Indirect Illumination for Inverse Rendering https://arxiv.org/abs/2204.06837
[3] Extracting Triangular 3D Models, Materials, and Lighting From Images https://arxiv.org/abs/2111.12503

JP07. コンピュータービジョンのためのウェブアプリケーション開発

現代の機械学習はより高精度なデータを多数求められています。その一つに角度付きバウンディングボックスがあります。著名なデータセットでは DOTA が上げられます。

-- *DOTA Dataset*

本課題では、このような角度付きバウンディングボックスを画像に付与するアノテーションツールを実装し、そのコード全体を圧縮した zip を提出してください。満たすべき仕様は以下になります。

- ローカルマシンにある任意の画像に対し、角度付きバウンディングボックスを複数含むアノテーションデータを作成する UI を持つこと
- 角度付きバウンディングボックスを、以下の TypeScript 型で示す JSON で出力できること

```
{
  "bboxes": {
    // BBox 中央の X, Y 座標は画像の top-left corner からの px で表現する
    // BBox 自体の回転の原点は BBox 中央で、時計回りに度数が増加する
    cx: number; // BBox 中央の X 座標（px）
    cy: number; // BBox 中央の Y 座標（px）
    width: number; // BBox の幅（px）
    height: number; // BBox の高さ（px）
    angle: number; // BBox の向き（deg、0-360 の間）
  }[]
}
```

例えば、1024×768 の画像全体をぴったり覆う回転していない BBox は

```
{
  "bboxes": [
    {
      "cx": 512,
      "cy": 384,
      "width": 1024,
      "height": 768,
      "angle": 0
    }
  ]
}
```

と表現されます

- クライアントサイドは TypeScript で実装すること。なお、React を用いることが望ましいです
- README を含むこと

これ以外の挙動や構成に指定はありません。例えば、以下の挙動は任意です。

- 画像の読込方法に指定はありません。以下のような方法で実現してください
  - サーバを実装しアップロードする
  - File API などの仕組みでローカルファイルを読み取る
  - Electron などの仕組みでローカルアプリケーションとして実装する
  - 他
- 同様にアノテーションデータの出力方法に指定はありません。以下のような方法で実現してください
  - サーバに保存され、ダウンロードできる
  - その場で任意の名前でダウンロードできる
  - 特定の名前（画像.jpg へのアノテーションなら画像.jpg.annotation.json など）で保存される
  - 他

このような未定義な挙動のうち、ツールにおいて重要なものは README に記載してください。

課題に取り組むに当たって、以下の部分を大きく評価します。

- コードの可読性や保守性
- アノテーションデータ作成のためのオリジナリティや工夫。例えば、新規作成の挙動はいくつか考えられます
  - BBox をドラッグで作成し、回転して合わせる
  - 三点をクリックすると近い角度付き BBox が生成される
  - 画像処理アルゴリズムを用いて、自動で角度付き BBox を付与し、作業者が正しい物を選ぶ
  - 他

これ以外の部分、アプリケーションのデザインや仕様外の機能の実装はあまり大きく評価しません。機能は最小限で構いません。

JP08. 機械学習を用いた地震波解析

`jp08` フォルダ内の pdf (`jp08/README-ja.pdf`) を参照してください。

JP09. 機械学習による時系列予測モデルの高度化

この課題では、２次元空間上の単調な関数を予測する機械学習モデルの作成に取り組みます。

`jp09` フォルダ内の notebook 中の指示に従って課題に取り組み、編集した notebook (.ipynb ファイル) をレポートとして提出してください。

添付されている書類は以下のとおりです:

- problem.ipynb
    - 課題を表す notebook ファイルです。このファイルを編集して提出してください。
- train.csv, valid.csv, test.csv
    - データセットです。 notebook 中の指示に従い取り扱ってください。

JP10. 機械学習による化学工学シミュレーションの高度化

つぎの各問にすべて答えてください。それぞれの問で指定する方法でレポートを作成してください。

第1問

この問題では簡単な質点の運動とそのシミュレーションを考察します。

`jp10` フォルダ内の notebook 中の指示に従って課題に取り組み、編集した notebook (.ipynb ファイル) をレポートとして提出してください。

第1問用の添付ファイルは以下のとおりです。

- problem.ipynb: 課題を表す notebook ファイルです。このファイルを編集して提出してください。
- data.csv: データセットです。notebook 中の指示に従い取り扱ってください。

第2問

【JP09. 機械学習による時系列予測モデルの高度化】 テーマの課題を問3まで解いてください。該当 notebook 中の指示に従って課題に取り組み、編集した notebook (.ipynb ファイル) をレポートとして提出してください。なお、提出物に問4に関する解答が含められた場合、問4に関しては評価しません。

JP11. 小売店向け画像認識システムに関連するソフトウェア・アプリケーション開発

（本テーマはテーマ別課題を設けておりません。）

JP12. 小売業における業務課題解決・最適化

（本テーマはテーマ別課題を設けておりません。）

JP13. 自動運転プランナ用の評価シミュレーション開発

自動運転のプランニング・シミュレーションの文脈における、安全性検証・他車挙動・安全なプランニングに関連する論文を一本選び、以下の点についてまとめてください。

1. 論文の要約
2. あなたがその論文を選んだ理由
3. 提案手法の新規性、特筆すべき点
4. 提案手法の課題点（例えば、十分考慮されていないケース、パラメータの疑問点、現実の交通シーンとのギャップ）
5. シミュレーション上での再現に必要な工数、実装上の課題点

全体でA4サイズ2枚以内で記述してください。

以下で発表された論文から選ぶことを推奨します。下記リスト外から論文を選択することも可能です。

- ITSC 2021
- IV 2021
- IROS 2021
- その他、自動運転関連企業が出した、2020 年以降の論文

JP14. 自動運転に用いる自己位置推定手法の開発

HD mapおよびカメラ・LiDARを用いた自己位置推定手法もしくは点群地図およびLiDAR点群データを用いた自己位置推定手法に関連する論文を一本選び、以下の点についてまとめてください。 高速道路100km以上にわたる広範囲に適用可能かつ自動運転車に搭載できる実アプリケーション向け手法を選択してください。

1. 論文の要約
2. 技術的な貢献、新規性や応用可能性などに基づいて、この論文にご興味を持たれた理由
3. 提案手法の課題点（例えば、十分考慮されていないケース、実アプリケーションで利用する際に生じる問題）
4. 論文の再現実装にかかる工数見積もり

全体でA4サイズ2枚以内で記述してください(様式自由)。

JP15. 自動運転に用いる物体認識手法の開発

自動運転の文脈でディープラーニングを使った3次元物体認識（検出・トラッキング）手法の応用・高速化もしくはセンサーデータのノイズ除去（雨や霧などの環境下で発生するデータノイズ）手法に関連する論文を一本選び、以下の点についてまとめてください。 オフライン学習手法を除き、自動運転車に搭載できる実アプリケーション向け手法を選択してください。

1. 論文の要約
2. 技術的な貢献、新規性や応用可能性などに基づいて、この論文にご興味を持たれた理由

3. 提案手法の課題点（例えば、十分考慮されていないケース、実アプリケーションで利用する際に生じる問題）
4. 論文の再現実装にかかる工数見積もり

全体でA4サイズ2枚以内で記述してください(様式自由)。

JP16. 教育のための学習管理システムの開発

あなたは、小学生から中学生くらいの生徒が主に学習塾で利用する、プログラミング学習アプリの開発者です。 あなたは、ユーザー数や継続率を高めるため、生徒が普段からどの程度アプリを利用しているかを分析し、このアプリを改善したいと思いました。 どのような分析を行うと良いか、またそのためにどのようなデータが必要だと考えられるかを、A4用紙1〜2ページ程度で具体的に議論してください。 ただし、アプリケーションには自由に細工を行うことができるとします。 なお、アプリのコンテンツに関してではなく、分析に関する提案を行ってください。

このプログラミング学習アプリには、以下のような特徴があります。必要なら、ここに無い特徴を仮定して頂いても構いません。

- 学習アプリはタブレット端末上で動作し、ネットワーク上のサーバーでプレイデータが管理される
- 小学生から中学生くらいの生徒が利用する
- 生徒は学習塾での授業の他に、自宅に端末を持ち帰って自由に学習できる
- 年齢などのユーザーの属性はある程度取得可能とする
- Unityで開発されている
- 学習モードと創作モードがある
- 学習モードではたくさんのステージがあり、ビジュアルプログラミングでパズルを解きながら、順番にステージを進める
- 創作モードでは、プログラミングをしながら自由に作品を作ることができる

例えば、以下の要素は加点要素となります。

- 分析観点のアイディアをたくさん出す
- ユーザー数や継続率を改善できるプロセスが論理的に記述されている
- ユーザー数や継続率を改善する手法を仮定し、その仮定を裏付けるような分析手法を提案をしている
- 利用者のペルソナを想定し、その背景に基づく提案をしている
- 開発のみならず、マーケティング等の多面的な視点を備えている
- Unityで開発し、タブレット上で動作するという技術特性に基づいた提案をしている

JP17. 機械学習・微分可能レンダラを用いた3次元復元

末尾の文献リストの中から一つ選び、以下の点について論じてください。A4で1ページのPDFファイルを提出してください。フォントサイズは10pt以上としてください。英語か日本語のうち、どちらか得意な言語で構いません。

(1) どのような課題を解決しようとしているか
(2) どのような方法により解決しているか
(3) 提案手法に欠点や限界や新規性の欠如はあるか。ある場合には特に重要なものはなにか

(4) 実験・議論の欠点や、論文で触れられていない潜在的な懸念はあるか。ある場合には特に重要なものはなにか

文献リスト

[1] HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video https://arxiv.org/abs/2201.04127
[2] Modeling Indirect Illumination for Inverse Rendering https://arxiv.org/abs/2204.06837
[3] Extracting Triangular 3D Models, Materials, and Lighting From Images https://arxiv.org/abs/2111.12503