ICML '24 notes and interesting posters

November 6, 2025

Montag Vormittag

Predictive attribution

- loss → what if one element changed? Leave One Out (LOO)
- can be solved in closed form for Linear Regression, Logistic Regressions, NNs,..

NN operator learning

- PDE learning → nonlinearities keys
- optimize these jointly with NN
- rewatch talk! (Physics of LLMs too)

Strategic Learning & behaviour

- behavior (human) influences ML decisions and vice-versa
- classifiers: transparent or opaque (e.g. Schufa? what are the merits)
- includes the social burden of ML system in loss (never explained how??)
- people move in classifier feature space - towards better decision rule, independent from position! $\implies$ classifiers create demand!
- strategic modification $\implies$ strategic participation in system!
  - is it worth to participate at all? - your fairness might be skewed!
  - e.g.: people might not apply in the first chance, showing a fair selection bias to "pre-selection"
  - fairness is opaque!
- possible solution: causality of change?

Montag Nachmittag

Distribution-Free UCQ

- problem of conformal splits: variability: $O(\sqrt{n})$
- no split conformal prediction
  - problem when scoring on training points → ↯ overfitting, all $s(x) = 0$
  - go back to classical CP - leave-one-out train for all → ↯$O(n^2)$
  - jackknife+ → problem if node unstable
- adaptability of CP using running adaption γ as basis (AgACI)

GNNs

- node-level tasks:
  - node embeddings (optimized based on similarity+random walks)
  - Problems: incorporate structure, adding data
- GNN - aggregate information ~ CNN
- message passing to neighbors
- final output layer: based on task!
- GCN: passing adjacency and diagonal matrix iteratively
- GraphSage
- GAT: Graph Transofromer: attend on neighbors
-
  - instead of future predict node
  - problem: convey position to transformer (usually sine embeddings)
  - output: node embeddings
  - regular sinusoidal embeddings with graph laplacian + learnable embeddings
  - problem : $O(n^2)$
- GraphGPS: message passing + transformer

Dienstag Vormittag

Unapologetic Openness

- why openness? → ecosystem ↑, community thrives!
- not just philanthropy
- why not: time advantage, could be used in harmful ways
- LLM Open Source: human feedback ↯ meta wants end user feedback, but is missing that for training...

Genie

- train Video model with action tokens for 16 frames
- goal: agents can use and understand sim

Arrows in time

- forward/backward CE of LLM
- Related to language/information theorem of Shannon
- Forward pass has a lower loss - indicates an arrow in time!
- Across all languages!
- gap increases with model size, across multiple model types
- origin: primes p_1, p_2, $p_1 \times p_2 = n$ - multiplication easy, factorisation is not
- causality?, very data-intense, not clear if it applies to other data

Transformers for pretraining Universal Forecasts: MOIRAI

- challenges: cross-frequency
- patch-based forecasting +masked
- multivariate: flattened, different encoding
- Future Work: combine with text?

Potential of Transformers for Timeseries prediction: SAMFormer

- robustness against time shift
- custom training routine: SAM
- very simple, better than MOIRAI-zero-Shot
- The same architecture works well for many systems

Mittwoch Vormittag

African Language Datasets

- translations missing, important to bring policy decisions to citizens
- no clear text available - only as PDFs or similar, only 10 % is translated!
- alignment issues
- voice dataset being built
- translate scientific content at scale
- code mixing problems (NLP)
- Lelapa (home): communicating in African languages
  - community, from scratch: 45% women!
  - legal aspects of AI largely unknown, a lot of workshops

Position: Measure Diversity, don't just claim it!

- collect geographically diverse dataset, diversity definition matters - which level of diversity, ...
- diversity can never be objective → values encode information (e.g. political)
- measurement still fundamental for ML
- measurement theory (social science), e.g. socioeconomic status based on many factors, only indirect measure possible
  - conceptualize
  - operationalize
  - evaluate
  - ?
- → scale $\neq$ diversity $\neq$ unbiased
- not much quality reported
- evaluation usually only on newer models
- measure diversity *within* dataset → problem: level of diversity, unknown definition!

Mittwoch Nachmittag

SceneCraft: Text2Scene

- challenge: semantic relationship not controllable
- solution: LLM agents repeat generative approach+function generation to build skills automatically
  1. asset list → CLIP search for similar assets
  2. scene decomposition using LLM
  3. layout checked for each object → semantics/relationships!
  4. critique & adopt functions
- extended to movie generation → movie poet, a bit weak

ChatGPT moderation at scale

- downsides to ChatGPT: learning hindered, factually incorrect
- indicator adjectives show that GPT use is on the rise
- indistinguishable from human?
- corpus-level detection (percentage)
- ~10% to 17% usage, Nature almost 0!
- Multimodal α estimation using known distributions
- ground truth generated by LLM generated reviews for papers before 2020, temporal split!
- modeling TF of on adjectives for probabilities
- common GPT detectors worse!
- BERT-based detectors weak
- deadline effect: more usage!
- more replies: less usage (more involvement!)
- only works globally, not necessarily bad - can be used as an indicator, not individual blame!

Stealing part of a production LLM

- finding single values of LLM responses
- singular value decomposition: after a certain number of stops steep falloff of values - indicates the limit of the last layer!
- indicates output subspace - consequently, last layer size!
- final layers can be learned too:

$$Q = U\Sigma V^T \tag{1}$$

- can be learned using SVD
- is worth stealing, as ML can be used to generate profit now!

MagicLens: Self-Supervised Image Retrieval

- usually in image retrieval: most *identical* image
- here: guide image + search intent - retrieve semantically relevant image!
- problem: training data:
  - websites with 2+ images as adjacent images, with nearby text
  - filter out ads (Google cannot disable their ads??)
- contrastive loss, good results
- outperforms SOTA image retrieval
- extremely good semantic retrieval

Donnerstag Vormittag

Position: Opportunities exist for ML+Fusion

- high energy output, tritium production, economics
- disruption prediction
- simulation & dynamics modeling - physics are incomplete!
- partial observability (related to our HO problem)
- controls problems, experiment design
- material design

HEPT: High Energy Particle Transformer

- Particle cloud embeddings for transformers

Donnerstag Nachmittag

Uncertainties for LLM

- perturb inputs instead of ensemble LLM
- disentangle → epistemic/aleatoric
- prompting/finetuning diversity

AlphaFlow Meets Flow Model Matching

- distribution of structures in protein folding
- generative modeling!
- AlphaFlow denoises 3D structure from template + protein

Freitag

ML4ESM: Towards improved cloud modelling

ML4ESM: Climate Set

- Climate models: future emissions → how does the climate react?
- Multiple socio-economic pathways
- ~ 390 days/simulation!
- problem: resolution scales $O(r^3)$
- ML: can help downsampling, parametrization, *emulation*
- Problems: distribution shift, data-based, high uncertainty in models (5 K)

ML4ESM: ML and Climate Change

- ML not problem/application driven!
- problem: limited resources, sparsely labeled data
- domain knowledge required - reduces compute significantly!
- Climate Simulation
  - reduce the resolution of simulation, scale up using super-resolution
  - keep physical constraints in mind
  - mapping to continuous functions: related to neural operator learning

ML4ESM: PDE+phys. Constraints+Spectral

ML4ESM: DDPM: Deep Denoising Physical Models

- PDE model using diffusion process → enables uncertainty modeling!
- constraint diffusion process!

Samstag

GRaM: Platonic Representation Hypothesis

- models learn same "representation"
- converges to same clues in feature spaces (e.g. dogs detector to ears, ...)
- "Rosetta neurons" - same representation accross many models → is there convergence?
  - H1: different representation
  - H2: or same representation? (good models ⇔ similar representation)
- Language+Visualisation: do models converge - some indications:
  - Use kernel to map similarity between models, map different concepts of e.g. GPT, ImageNet
  - result: language represents similar concepts as vision!
  - a lot of limitations, currently only 0.2/1, does not converge to reality

Sociotechnical Evaluation of AI

- layers: capabilities, human interactions, systemic impacts
- problem: only technical aspects of AI considered & mostly textual evaluation
- e.g. textual evaluation:
  - replica users, mental health impact
  - stackoverflow activity drop after ChatGPT release
  - homogenization of creative writing: least create get uplift, most creative reduce creativity - narrowing of the spectrum!
- studies: synthetic simulation?

AI safety institute (UK)

- evaluation of AI: misuse, societal impacts (long term!), autonomous systems (loss of control, safeguards for agents and tools!)

Future of video generation - beyond data and scale

- currently: imperfect control over semantics
- research: single video model, instead of foundational model → can be used to split background/-foreground, alpha & recombine

Adverserial Perturbations cannot Reliably protect artists from generative AI

- existing adversarial perturbation can easily be bypassed using:
  - Gaussian Filters
  - One Diffusion step
  - ...

CopyCat

- Remove copyrighted characters
- Using: negative prompting (post hoc - open models can easily circumvent that!)

Posters

— **Poster** — **Title** — **Information** — — — — — — — — — — ![poster](out$_r$educed/IMG$_1$560.jpeg)|Block−lev
GaneshBannur, BharadwajAmrutur∗, [SpottingLLMsWithBinoculars : Zero−ShotDetectionofMachine−GeneratedText
//arxiv.org/abs/2406.13208v1)| |![poster](out$_r$educed/IMG$_1$537.jpeg)|ScalingRectifiedFlowTransformersforHigh−R
PatrickEsser, SumithKulal, AndreasBlattmann, RahimEntezari, JonasMüller, HarrySaini, YamLevi, DominikLorenz, Axe
//arxiv.org/abs/2403.03206v1)| |![poster](out$_r$educed/IMG$_1$599.jpeg)|Black−Boxvs.Gray−Box : ACaseStudyonLearni
JanAchterhold, PhilipTobuschat, HaoMa, DieterBuechler, MichaelMuehlebach, JoergStueckler∗, [SCENE−NetV2isag
//arxiv.org/abs/2305.15189v2)| |![poster](out$_r$educed/IMG$_1$540.jpeg)|WhisperingExperts : NeuralInterventionsforTo
averageof[PIQA, SIQA, TriviaQA, TruthfuGA, Hellaswag)∗, [WhisperingExperts : NeuralInterventionsforToxicityMi
ChrisM.Chambers, WilliamA.Hiscock, BrettTaylor∗, [HumanTOMATO : Text−alignedWhole−bodyMotionGeneration](
//dx.doi.org/10.1103/PhysRevLett.78.3249)| |![poster](out$_r$educed/IMG$_1$605.jpeg)|TheeffectsofGribovcopiesin2Dga
D.Dudal, S.P.Sorella, N.Vandersickel, H.Verschelde∗, [Vid3D : SynthesisofDynamic3D](http : //dx.doi.org/10.1016/j.
LeonidasLefakis, OleksandrZadorozhnyi, GillesBlanchard∗, [PiecewiseConstantandLinearRegressionTrees](http :
//arxiv.org/abs/1907.00275v1)| |![poster](out$_r$educed/IMG$_1$598.jpeg)|fine−tuningusingtheinputrestorationshapecanin
[3]AngelX.Chang, ThomasA.Funkhouser, LeonidasJ.Guibas, PatHanrahan, Qi−XingHuang, ZimoLi, Silvio∗, [fine−tunin
AnLLMAgentforSynthesizing3DSceneasBlenderCode|∗ZiniuHu, AhmetIscen, AashiJain, ThomasKipf, YisongYue, Davi
//arxiv.org/abs/2403.01248v1)| |![poster](out$_r$educed/IMG$_1$536.jpeg)|Position : KeyClaimsinLLMResearchHaveaLon
AnnaRogers, AlexandraSashaLuccioni∗, [KeyClaimsinLLMResearchHaveaLongTailofFootnotes∗](http :
//arxiv.org/abs/2308.07120v2)| |![poster](out$_r$educed/IMG$_1$561.jpeg)|LargeLanguageModelsareGeographicallyBiase
RohinManvi, SamarKhanna, MarshallBurke, DavidLobell, StefanoErmon∗, [LargeLanguageModelsareGeographically
//arxiv.org/abs/2402.02680v2)| |![poster](out$_r$educed/IMG$_1$507.jpeg)|MISGENDERED : LimitsofLargeLanguageMo
TamannaHossain, SunipaDev, SameerSingh∗, [intheweightsofGPT−Neo−1.3Busing](http : //arxiv.org/abs/2306.03950
partners(Backpack, Book, Bottle, Opener, Candle, Sandal).∗, [layer6]()| |![poster](out$_r$educed/IMG$_1$546.jpeg)|MMOn
RepresentingMultipleModalitiesinOneScene|∗ZhifengGu, BingWang∗, [POSITION : MISSIONCRITICAL−SATELLIT
//arxiv.org/abs/2507.11129v2)| |![poster](out$_r$educed/IMG$_1$511.jpeg)|Humanvs.GenerativeAIinContentCreationComp
SymbiosisorConflict?|∗FanYao, ChuanhaoLi, DenisNekipelov, HongningWang, HaifengXu∗, [Humanvs.GenerativeAIin
](http : //arxiv.org/abs/2402.15467v1)| |![poster](out$_r$educed/IMG$_1$602.jpeg)|MultimodalCropTypeClassificationFus
ValentinBarriere, MartinClaverie∗, [QUANTiFpiNGLiKENESS](http : //arxiv.org/abs/2208.10838v1)| |![poster](out$_r$
StealingDatawithCorruptedPretrainedModels| ∗ ShanglunFeng, FlorianTramèr∗, [PrivacyBackdoors :
StealingDatawithCorruptedPretrainedModels](http : //arxiv.org/abs/2404.00473v1)| |![poster](out$_r$educed/IMG$_1$531.
|∗Objective : givensomesub−sequence(x, "..., x(∗)predictthe∗, [byLinearRegression :]()| |![poster](out$_r$educed/IMG$_1$56
Self−ReferentialSelf−ImprovementViaPromptEvolution|∗ChrisanthaFernando, DylanBanarse, HenrykMichalewski,
Self−ReferentialSelf−ImprovementviaPromptEvolution](http : //arxiv.org/abs/2309.16797v1)| |![poster](out$_r$educed
HongChulNam'", JuliusBerner?", AnimaAnandkumar2∗, [SolvingPoissonEquationsusingNeuralWalk−on−Spheres]()| |!
Next−GenerationImageRetrievalModels|∗KaiZhang, YiLuan, HexiangHu, KentonLee, SiyuanQiao, WenhuChen, YuSu, an
Next−GenerationImageRetrievalModels]()| |![poster](out$_r$educed/IMG$_1$551.jpeg)|Localvs.GlobalInterpretability :
AComputationalComplexityPerspective|∗ShahafBassan, GuyAmir, GuyKatz∗, [Localvs.GlobalInterpretability :
AComputationalComplexityPerspective](http : //arxiv.org/abs/2406.02981v2)| |![poster](out$_r$educed/IMG$_1$506.jpeg)|
nn.Convld(25, 25, 3, stride = 2, padding = 1), ∗, [9$8000089]()| |![poster](out$_r$educed/IMG$_1$529.jpeg)|IntegratedInforma
KobiKremnizer, AndréRanchin∗, [Mean−fieldChaosDiffusionModels](http : //dx.doi.org/10.1007/s10701−015−9905−
RobertHönig, JavierRando, NicholasCarlini, FlorianTramèr∗, [AdversarialPerturbationsCannotReliably](http :
//arxiv.org/abs/2406.12027v2)| |![poster](out$_r$educed/IMG$_1$568.jpeg)|AGeometricExplanationoftheLikelihoodOODD
HamidrezaKamkari, BrendanLeighRoss, JesseC.Cresswell, AnthonyL.Caterini, RahulG.Krishnan, GabrielLoaiza−Gane
//arxiv.org/abs/2403.18910v2)| |![poster](out$_r$educed/IMG$_1$552.jpeg)|ScalablePre−trainingofLargeAutoregressiveIm
AlaaeldinEl−Nouby, MichalKlein, ShuangfeiZhai, MiguelAngelBautista, AlexanderToshev, VaishaalShankar, JoshuaM
//arxiv.org/abs/2401.08541v1)| |![poster](out$_r$educed/IMG$_1$505.jpeg)|ACircuitDomainGeneralizationFrameworkforE
ZhihaiWang, LeiChen, JieWang, XingLi, YinqiBai, XijunLi, MingxuanYuan, JianyeHao, YongdongZhang, FengWu∗, [ACi
//arxiv.org/abs/2309.03208v1)| |![poster](out$_r$educed/IMG$_1$533.jpeg)|StrategiesToEvaluateTheRiemannZetaFunction
AloisPichler∗, [20202200227012021](http : //arxiv.org/abs/1201.6545v1)| |![poster](out$_r$educed/IMG$_1$564.jpeg)|Quant
AritraSarkar, ZaidAl−Ars, KoenBertels∗, [SuperpositionPrompting : ImprovingandAcceleratingRetrieval−AugmentedG
//arxiv.org/abs/2006.00987v2)| |![poster](out$_r$educed/IMG$_1$509.jpeg)|AudioFlamingo : ANovelAudioLanguageModelw
[8]Gong, Yuan, etal.Listen, Think, andUnderstand.ICLR2024.∗, [AudioFlamingo : ANovelAudioLanguageModelwith]()|
LauraR.Marusich, JonathanZ.Bakdash, YanZhou, MuratKantarcioglu∗, [USINGAIUNCERTAINTYINFORMATIONTO
//arxiv.org/abs/2309.10852v2)| |![poster](out$_r$educed/IMG$_1$549.jpeg)|AboutGeometryandInitialPhaseofCloud−to−Gr
AlešBerkopec∗, [Craftox](http : //arxiv.org/abs/1602.02496v1)| |![poster](out$_r$educed/IMG$_1$553.jpeg)|Data−freeDis
Fo(21, t, c) = fó(xt, t, n)+w.(foxt, t, c)−foxt, t, n))∗, [Data−freeDistillationofDiffusionModelswithBootstrapping]()| |!
MaximilianoUjevic, AlirezaRashti, HenriqueGieg, WolfgangTichy, TimDietrich∗, [Flow : DifferentiableSimulationswi

//dx.doi.org/10.1103/PhysRevD.106.023029)| |![poster](out$_r$educed/IMG$_1$512.jpeg)|FlowRounding| *
DongguKang, JamesPayor*, [MusicFlow : CascadedFlowMatchingfor](http : //arxiv.org/abs/1507.08139v1)| |![poster
WebLINX : Real–World[...]*, [WebLINX]()| |![poster](out$_r$educed/IMG$_1$601.jpeg)|EvaluatingCopyrightTakedownM
BoyiWei*), WeljiaShi*z, YangsiboHuang*1, NoahA.Smithz, ChiyuanZhang, LukeZettlemoyer2, KaiLit, PeterHenderson1
XindiWang, RobertE.Mercer, FrankRudzicz*, [OAK : EnrichingDocumentRepresentationsusingAuxiliaryKnowledgefor
//arxiv.org/abs/2405.19084v1)| |![poster](out$_r$educed/IMG$_1$562.jpeg)|TopologicalUncertaintyforAnomalyDetectionin
KenjiFukushima, SyoKamata*, [SECOND–ORDERUNCERTAINTYQUANTIFICATION :](http : //arxiv.org/abs/250
MauritsBleeker, MariyaHendriksen, AndrewYates, MaartendeRijke*, [RevisitingtheRoleofLanguagePriorsinVision–La
//arxiv.org/abs/2402.17510v2)| |![poster](out$_r$educed/IMG$_1$597.jpeg)|TacticsandTallies : AStudyofthe2016U.S.Presid
YuWang, XiyangZhang, JieboLuo*, [ATaxonomyofTacticsandInsightsfromMediaData](http : //arxiv.org/abs/1704.02
Gradient–basedVisualandTextualExplanationsforCLIP|*ChenyangZhao, KunWang, JanetH.Hsiao, AntoniB.Chan*, [G
//arxiv.org/abs/2502.18816v1)| |![poster](out$_r$educed/IMG$_1$515.jpeg)|Bringingmotiontaxonomiestocontinuousdomains
NoémieJaquier, LeonelRozo, MiguelGonzález–Duque, ViacheslavBorovitskiy, TamimAsfour*, [BringingMotionTaxono
//arxiv.org/abs/2210.01672v5)| |![poster](out$_r$educed/IMG$_1$542.jpeg)|Better&FasterLargeLanguageModelsviaMulti–
FabianGloeckle, BadrYoubiIdrissi, BaptisteRozière, DavidLopez–Paz, GabrielSynnaeve*, [Better&FasterLargeLangua
//arxiv.org/abs/2404.19737v1)| |![poster](out$_r$educed/IMG$_1$543.jpeg)|TJ–FlyingFish : DesignandImplementationofa
XuchenLiu, MinghaoDou, DongyueHuang, BiaoWang, JinqiangCui, QinyuanRen, LihuaDou, ZhiGao, JieChen, BenM.Che
//arxiv.org/abs/2301.12344v2)| |![poster](out$_r$educed/IMG$_1$514.jpeg)|Position : TensorNetworksareaValuableAssetf
EvaMemmel, ClaraMenzen, JetzeSchuurmans, FrederiekWesel, KimBatselier*, [Position : TensorNetworksareaValuab
//arxiv.org/abs/2205.12961v2)| |![poster](out$_r$educed/IMG$_1$596.jpeg)|Spectrallensenablesaminimalistframeworkfor
ZhouZhou, YihengZhang, YingxinXie, TianHuang, ZileLi, PengChen, YanqingLu, ShaohuaYu, ShuangZhang, GuoxingZhe
//dx.doi.org/10.1038/s41377–024–01608–w)| |![poster](out$_r$educed/IMG$_1$559.jpeg)|DualConvexifiedConvolutionalNe
SiteBai, ChuyangKe, JeanHonorio*, [ParameterRecoveryofNeuralNetworks](http : //arxiv.org/abs/2205.14056v2)| |![
BuildingOpen–EndedCreativeAgentsviaAutonomousEmbodiedVerification|*YuxuanGuo, ShaohuiPeng, JiamingGuo, Di
//arxiv.org/abs/2405.15414v1)| |![poster](out$_r$educed/IMG$_1$534.jpeg)|TRUSTWORTHYACTIONABLEPERTURBATIO
DepartmentofECE, UniversityofArizona, Tucson, AZ, USA*, [TRUSTWORTHYACTIONABLEPERTURBATIONS–"]()
ACaseStudyontheImpactofChatGPTonAIConferencePeerReviews|*WeixinLiang, ZacharyIzzo, YaohuiZhang, HaleyLe
](http : //arxiv.org/abs/2403.07183v2)|