Premissas -

IA verde = considerar consumo de energia, sustentabilidade ambiental e desempenho do modelo.

IA centrada em dados = Uso eficiente de dados de treinamento p/ aumentar eficiência dos algoritmos de IA

Single-Task Learning = aprende o modelo do zero

Técnicas IA centrada em dados -

Aprendizado ativo (AL) = melhorar o desempenho do modelo com rotulagem mínima dos dados

Transferência de conhecimento (KT) = reutilizar modelos treinados previamente no treino corrente

Distilação de dataset (DD) = redução do dataset de treinamento com a síntese de poucas amostras que representam o total

Coreset (CS) = redução do dataset de treinamento a um subconjunto com o mesmo desempenho de trainamento que o original

Incremento de dados (DA) = criar novos dados através de pertubações nos originais para acelear convergência

Aprendizado curricular (CL) = treinar o modelo segundo uma sequência de dificuldade, dos mais simples aos mais difíceis.

Análise Técnicas IA/CD -

AL = Descobrir que amostras tem maior impacto na evolução do modelo, rotulá-las e usá-las no treinamento. **Utilizar menos amostras com pouco impacto e por menos épocas de treinamento.** Também pode apresentar técnicas úteis de detecção de significância de amostras.

CS = Similar ao AL, para dados já rotulados. **Redução de dados de treinamento acelera o treino e diminui o custo.**

KT = Não aplicável devido à restrição do cenário em ser single-task.

DD = A síntese de dados, além de reduzir a quantidade de dados de treinamento, pode ser útil em sintetizar os dados de um histórico e reduzir o esquecimento catastrófico. Sincronizar os datasets entre clientes ao invés de modelos.

DA = Útil no aumento da diversidade do dataset de treinamento. **Reduz tempo de treinamento por acelaração da convergência.**

CL = Melhora tempo de convergência, constestado em alguns artigos, menos dados de treinamento e iterações.

FL/CD -

FedCS: Coreset Selection for Federated Learning = Seleciona um subconjunto de dados para treinamento de forma federada.

Observações:

- Subconjunto não me parece balanceado em cada cliente. DA ou DD poderiam aumentar o subconjunto para balancear o treinamento? Datasets mais IID convergem mais rapidamente.

- Ao invés de enviar o modelo agregado para os clientes, não seria interessante selecionar dentre os clientes aqueles que melhor complementam o histórico de treinamento do modelo, tornando-o mais iid? Ou seja, se K=10, teríamos 10 modelos no servidor que seriam enviados a 10 clientes selecionados que tornam a distribuição do histórico de amostras usadas mais IID. A agregação do modelo global seria sobre esses modelos.

- A seleção de coreset é realizada uma única vez, baseando-se em um modelo pré-treinado, ou seja, com base na transferência de conhecimento. Se não foi viável isso, deveria ocorrer rodada a rodada, ou em janelas de rodadas. Talvez o método usado seja custoso, e um método de AL ou CL defina melhor o modelo.

- Se houver mudança de domínio não há mais para quê usar o dataset CS definido, precisando engatilhar novamente a seleção.

FedTGP: Trainable Global Prototypes with Adaptive-Margin-Enhanced Contrastive Learning for Data and Model Heterogeneity in Federated Learning = treina modelos diferentes para o mesmo problema no mesmo aprendizado. Utiliza os protótipos das classes dos clientes para regularizar o treinamento.

Observações:

- Como esse cenário funciona com o coreset proposto FedCS?

Perguntas -

Longo prazo:

Como as técnicas são utilizadas em aprendizado distribuído? (Survey)

Qual a aplicabilidade delas em aprendizado distribuído, prós e contras? (Artigo sobre análise de desempenho)

Elas se complementam em ambiente distribuído? (Artigo de framework de CD p/ ambiente distribuído)

Curto prazo (SBRC):

Acredito que os dois marcados em verde. Validar se há ganho em mesclar CS com DA ou DD. E para sigle-task learning, uma estratégia de atualização do CS por rodadas.

Referências -

[1] Data-Centric Green Artificial Intelligence: A Survey

[2] FedCS: Coreset Selection for Federated Learning

[3] FedTGP: Trainable Global Prototypes with Adaptive-Margin-Enhanced Contrastive Learning for Data and Model Heterogeneity in Federated Learning