Kurzvortrag: Survey on Continual Learning

Jörg Schantz

26. November 2024

Aktueller Gliederungsentwurf

2. Framework

Tasks $t = 1, ..., T$ und Sampleset $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_T\}$

$\mathcal{D}_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t}$ mit n_t als die t-te Stichprobengröße

$\mathcal{D}_t$ folgen den Verteilungen $\mathbb{D}_t = p(X_t, Y_t)$

Model $p(\mathcal{D}_{1:k}|\theta) = \prod_{t=1}^{k} p(\mathcal{D}_t|\theta)$

2.1 Evaluation Metrics

Average Accuracy $AA_k = \frac{1}{k} \sum_{i=1}^{k} a_{k,i}$

Backward Transfer $BWT_k = \frac{1}{k-1} \sum_{i=1}^{k-1} (a_{k,i} - a_{i,i})$

Forward Transfer $FWT_k = \frac{1}{k-1} \sum_{i=2}^{k} (a_{i,i} - \tilde{a}_i)$

$\tilde{a}_{k,i}$ Accuracy des Models $\mathrm{p}(\mathcal{D}_{1:k}|\theta)$ auf Testset i
$\tilde{a}_i$ Accuracy des Models $p(\mathcal{D}_i|\theta)$

2.2 Distribution Drift and Bayes

Verteilungen alter und neuer Tasks müssen erhalten bleiben

Speichern oder generieren alter Daten

Formulierung eines Prior basierend auf dem bisher Gelernten (alter Posterior) als Laplace Approximation oder Variational Inference

3.1 Replay

Experience Replay speichert kleine Menge alter Samples
Auswahl: feste Menge, nah am Mean, Erhaltung des Gradient

Generative Replay: zusätzliches Generativ Model; X_k und $X'_{1:k-1}$ ja nach Gewichtung gemischt

3.2 Optimization

Parameter-Update durch alte Gradient Descent Richtung kontrolliert.
Orthogonal; Dynamische Priorisierung

Anpassung des "loss-landscape"

Meta-learning als Versuch Task-spezifische Bias zu lernen

3.3 Representation

Pre-Training methoden und Self-Supervised Learning
Herausforderung ist das Pre-Training flexibel einzusetzen

3.4 Architecture

Aufteilung der Parameter in "Task-spezifisch" und "geteilt"

4.1 Parameter Space

Gewichtung der Parameter basierend auf Fisher Information mit einer idR quadratischen Loss-Funktion
Über FI direkt, Total-Loss, Änderung der Prediction

Andere sind: Variational Inference, Forgetting Rate um den Prior zu regulieren

4.2 Function Space

Reguliert den Prediction Output
Teacher / Student Beziehung in Kombination mit Knowledge Distillation Unterschiede in Vergleichsgröße der KD (alte Samples,

generiert, neues Sample, unlabeled Data), deswegen oft in Kombination mit Replay

Conclusion

Großes Fragezeichen...