ZUSAMMENFASSUNG

Das Verständnis der subzellulären Lokalisation von Proteinen ist zentral, um zelluläre Organisation und Funktion zu erfassen. In dieser Arbeit untersuchen wir selbstüberwachtes Lernen zur Bestimmung von Proteinlokalisationen mithilfe von self-distillation with no labels (DINO) auf dem OpenCell-Datensatz und betrachten sowohl Field of View- (FOV-) als auch Single-Cell-Klassifikation. Aufbauend auf dem DINO4cells-Framework analysieren wir systematisch, wie Modellarchitektur, Pretraining-Domäne und Kanal-Embedding-Strategien die Qualität der gelernten Repräsentationen beeinflussen.

Wir vergleichen Residual Networks (ResNets) und Vision Transformers (ViTs), die entweder von Grund auf trainiert, auf ImageNet-1k oder Human Protein Atlas (HPA) vortrainiert oder zusätzlich auf OpenCell finetuned wurden. Unsere Ergebnisse zeigen, dass ViTs die ResNets deutlich übertreffen und dass domänenspezifisches Pretraining auf HPA FOV in Kombination mit Channel Mapping zu den besten Ergebnissen in der FOV-Klassifikation führt. Finetuning verbessert die Leistung weiter, solange Channel Mapping verwendet wird, während Channel Replication die Performance meist verschlechtert. Analysen der Attention Heads sowie UMAP-Projektionen zeigen zudem, dass die Modelle mit der besten Leistung biologisch sinnvolle räumliche Strukturen erkennen.

Für die Single-Cell-Repräsentationen erzielen die besten k-Nearest-Neighbor-Ergebnisse die DINO-Modelle, die auf dem HPA Single-Cell-Datensatz vortrainiert wurden, was den Wert domänenspezifischen und umfangreichen Pretrainings unterstreicht. Trotz deutlicher Klassenunbalance bilden die Embeddings gut getrennte Cluster über die meisten Kompartimente hinweg. Insgesamt zeigen unsere Ergebnisse, dass self-supervised ViTs in Kombination mit geeignetem Pretraining und passenden Kanalstrategien leistungsstarke Werkzeuge zur Analyse von Proteinlokalisationen sind.

ABSTRACT

Understanding where proteins reside within cells is essential for studying cellular organization and function. In this work, we investigate self-supervised learning for subcellular protein localization using self-distillation with no labels (DINO) on the OpenCell dataset, covering both Field of View (FOV) and single-cell classification tasks. Building on the DINO4cells framework, we systematically analyze how model architecture, pretraining domain and channel-embedding strategies influence the quality of learned representations.

We compare Residual Networks (ResNets) and Vision Transformers (ViTs) trained from scratch, pretrained on ImageNet-1k or Human Protein Atlas (HPA) datasets and further finetuned on OpenCell. Our results show that ViTs greatly outperform ResNets and that domain-specific pretraining on HPA FOV, combined with channel mapping, leads to the strongest improvements in FOV-level classification. Finetuning provides additional gains when using channel mapping, while channel replication generally reduces performance. Attention-head analysis and UMAP projections reveal that the best performing models learn biologically meaningful spatial features.

For single-cell representations, the best k-Nearest Neighbor results stem from DINO pretrained on the HPA single-cell dataset, demonstrating the value of domain-matched pretraining. Despite strong class imbalance, the embeddings form well-separated clusters across most compartments. Overall, our findings highlight that self-supervised ViTs, when paired with appropriate pretraining and channel strategies, serve as powerful and label-efficient tools for learning protein localization patterns.