M2 internship
Self-supervised learning for link prediction

Context

Network science

Network science [1] is an interdisciplinary field of research interdisciplinary research field, within the broader field of Complex Systems. Its main focus is the study of interactions within real complex systems, such as social networks, economic networks, transportation networks, etc. The underlying idea is that for a complex system, in which interactions between components are not random, but follow a specific organization, it is essential to understand this organization in order to understand how the system works. Complex networks typically obey several types of organization, such as community structures, spatial structures, core-periphery structures, hierarchical structures, nestedness and so on. [2] In recent years, machine learning approaches, and more specifically those based on deep neural networks, have attracted considerable interest in many fields, and network science is no exception. This is because, a network is a complex, structured object, in the same way as an image or a text. Much of the work carried out in these fields has therefore naturally been transposed to networks, giving rise to the field of Graph Neural Networks (GNN) [7].

Self-supervised learning

Self-supervised learning is a type of learning consisting in defining a supervisory signal, in the form of a pretext task, from unlabeled data, for which the available databases are generally larger. The representations learned can then be used for one or more downstream tasks. Two main types of pretext task were proposed in the literature. Generative tasks, such as the auto-encoder, which consist of using an encoder that projects inputs into a latent space, followed by a decoder whose task is to reconstruct the data. The discriminative tasks, that interest us here, such as contrastive learning, were initially applied in the image domain [3, 5, 8]. The aim is to learn a latent space latent space where two transformations of the same data (generated via augmentations), have similar representations, while using different tricks to avoid collapsing to simple solutions. We thus force explicitly similar synthetic inputs to represent similar concepts and therefore must have similar representations. The obtained classification performances are generally better.

Subject

The internship will explore the use of self-supervised learning for a classic problem in network science, link prediction. The objective of this task is to predict, for a given graph, which pairs of nodes are most likely to be connected by a link, among those that aren't already. Beyond practical applications such as recommendation (e.g., user/product linkage) or health (interactions between interactions between drugs or proteins), the ability to predict a link is also a proxy to assess our ability to understand how a graph is organized: the organization of the graph is in fact entirely defined by its links, and conversely, the observed links are a consequence of the graph's organization. In classical GNN approaches to link prediction, the prediction of a link between two nodes is calculated from the pairs of vectors of the corresponding nodes in the embedding space. The objective of contrastive learning methods in this context is to be able to guide the structure of the embedding space in a way that preserves desired properties, while varying others. For example, a Stochastic Block Model (SBM) with a given community structure can generate a large number of

different graphs, with few links in common, but all corresponding to the same organization. This information could be used to generate pairs of graphs with few links in common, but the same structure, or on the contrary, with many links in common, but different structures. Some works in this direction has recently been published, e.g., [6].

The first task will be to reproduce these results, in order to evaluate them in our context of link prediction. The second task will focus on proposing original approaches. In particular, most existing approaches seek only to preserve the community structures of the graph. We will seek to preserve a set of structures, by generating contrastive pairs not only for different graphs with the same community structure, but also for pairs of graphs with the same spatial structure, core-periphery, etc. By imposing an embedding that respects as much as possible the multiple aspects of the network structure, based for example on similar work in image processing [4], we believe we will be able to improve the ability to capture the complexity of its organization, and thus to predict its future links.

Profile

The following skills are mandatory:

- master's degree in artificial intelligence / machine learning or equivalent
- good programming skills (Python, Pytorch/Tensorflow)
- autonomy
- scientific curiosity

Duration

The internship will start in February-March and last 5-6 months.

Gratification

4.05€/h, 35h/week (i.e. around 580€/month)

Localisation

LIRIS laboratory, Lyon, France.

Advisors

- Rémy Cazabet: associate professor at LIRIS (`http://cazabetremy.fr/`)
- Mathieu Lefort: associate professor at LIRIS (`https://perso.liris.cnrs.fr/mathieu.lefort/`)

Application

Please send a CV, cover letter and transcripts of your current and previous years' results to Rémy Cazabet (remy.cazabet@liris.cnrs.fr) and Mathieu Lefort (mathieu.lefort@liris.cnrs.fr).

References

[1] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

[2] Remy Cazabet, Salvatore Citraro, and Giulio Rossetti. Structify-net: Random graph generation with controlled size and customized structure. *arXiv preprint arXiv:2306.05274*, 2023.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve visual instance discrimination. In *The Eleventh International Conference on Learning Representations*, 2022.

[5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[6] Bolian Li, Baoyu Jing, and Hanghang Tong. Graph communal contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 1203–1213, 2022.

[7] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[8] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

Stage M2
Apprentissage auto-supervisé pour la prédiction de liens

Contexte

Science des réseaux

La science des réseaux [1] est un domaine de recherche interdisciplinaire, s'inscrivant plus généralement dans le champ des Systèmes Complexes. Son objet principal est l'étude des interactions au sein des systèmes complexes réels, tels que les réseaux sociaux, les réseaux économiques, réseaux de transports, etc. L'idée sous-jacente est que pour un système complexe, dans lequel les interactions entre les composantes ne sont pas aléatoires, mais obéissent à une organisation spécifique, il est indispensable de comprendre cette organisation pour comprendre le fonctionnement du système. Les réseaux complexes obéissent typiquement à plusieurs types d'organisation, telles que des structures en communautés, structures spatiales, en cœur-périphérie, hiérarchique, *nestedness*, etc. [2] Au cours des dernières années, les approches de type apprentissage automatique, et plus spécifiquement à base de réseaux de neurones profonds, ont attiré un intérêt important dans de nombreux domaines, et la science des réseaux ne fait pas exception. En effet, un réseau constitue un objet complexe et structuré, comme peut l'être une image ou un texte. De nombreux travaux effectués dans ces domaines ont donc naturellement été transposés sur les réseaux, donnant naissance au domaine des *Graph Neural Networks* (GNN) [7].

Apprentissage auto-supervisé

L'apprentissage auto-supervisé est une forme d'apprentissage consistant à définir une fonction de supervision, sous la forme d'une tâche prétexte, à partir de données non labellisées, dont les bases de données disponibles sont généralement plus grandes. Les représentations ainsi apprises peuvent ensuite être utilisées pour une ou plusieurs tâches aval. Deux grands types de tâches prétextes ont été proposées dans la littérature. Les tâches génératives comme l'auto-encodeur qui consiste en l'utilisation d'un encodeur qui projette les entrées dans un espace latent, puis d'un décodeur dont la tâche est de reconstruire les données. Les tâches discriminatives, qui nous intéressent ici, comme l'apprentissage contrastif qui a été initialement appliqué dans le domaine de l'image [3, 5, 8]. L'objectif est d'apprendre un espace latent où deux transformations d'une même donnée (générées via des augmentations), ont des représentations similaires, le tout en utilisant différentes astuces pour éviter un effondrement vers des solutions simples. On force ainsi explicitement des entrées synthétiques similaires à représenter des concepts similaires, et donc doivent avoir des représentations similaires. Les performances obtenues en classification sont ainsi généralement meilleures.

Sujet

Le stage visera à explorer l'utilisation de l'apprentissage auto-supervisé pour un problème classique en science des réseaux, la prédiction de lien. L'objectif de cette tâche est de prédire, pour un graphe donné, quelles sont les paires de nœuds ayant le plus de chance d'être connectées par un lien, parmi celles qui ne le sont pas déjà. Au-delà d'applications pratiques telles que la recommandation (e.g., lien utilisateur/produit) ou la santé (interactions entre médicaments ou entre protéines), la capacité à prédire un lien est également un proxy pour évaluer notre capacité à comprendre comment un graphe est organisé : l'organisation du graphe est en effet intégralement définie par ses liens, et réciproquement, les liens observés sont une conséquence de l'organisation du

graphe. Dans les approches classiques de prédiction de lien par GNN, la prédiction d'un lien entre deux nœuds se calcule à partir des paires de vecteurs des nœuds correspondant dans l'espace de plongement de ce graphe. L'objectif des méthodes d'apprentissage contrastif dans ce contexte est d'être capable de guider le plongement des nœuds de manière à préserver des propriétés désirées, tout en en variant d'autres. Par exemple, un modèle de graphe en bloc (*Stochastic Block Model*, SBM) ayant une structure en communauté donnée peut générer un grand nombre de graphes différents, ayant peu de liens en commun, mais correspondant tous à une même organisation. Cette information pourrait être utilisée, en générant des paires de graphes ayant peu de liens en commun, mais une même structure, ou au contraire de nombreux liens en commun, mais des structures différentes. Des travaux dans cette direction ont été publiés récemment, e.g., [6]. La première tâche consistera à reproduire ces résultats, afin de les évaluer dans le contexte qui nous intéresse, celui de la prédiction de lien. La seconde tâche s'attachera à proposer des approches originales. En particulier, la plupart des approches existantes ne cherchent qu'à conserver les structures en communauté du graphe. Nous chercherons à préserver un ensemble de structures, en générant des paires contrastives non seulement pour des graphes différents ayant la même structure en communauté, mais également des paires de graphes ayant la même structure spatiale, en cœur périphérie, etc. En imposant un plongement préservant autant d'aspects que possible de la structure du réseau, en nous appuyant par exemple sur des travaux similaires en image [4], nous pensons être en mesure d'améliorer la capacité à capturer la complexité de son organisation, et donc à prédire ses futurs liens.

Profil

Les compétences suivantes sont indispensables :
— master en intelligence artificielle / machine learning ou équivalent
— bonne capacité de programmation (Python, Pytorch/Tensorflow)
— autonomie
— curiosité scientifique

Durée

Le stage commencera en Février-Mars pour une durée de 5-6 mois.

Gratification

4.05€/h, 35h/semaine (i.e. environ 580€/mois)

Localisation

Laboratoires LIRIS, Lyon, France.

Encadrants

— Rémy Cazabet : MCF au LIRIS (`http://cazabetremy.fr/`)
— Mathieu Lefort : MCF au LIRIS (`https://perso.liris.cnrs.fr/mathieu.lefort/`)

Candidature

Merci d'envoyer un CV, une lettre de motivation et les relevés de notes de l'année en cours et précédente à Rémy Cazabet (remy.cazabet@liris.cnrs.fr) et Mathieu Lefort (mathieu.lefort@liris.cnrs.fr).

Références

[1] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 371(1987) :20120375, 2013.

[2] Remy Cazabet, Salvatore Citraro, and Giulio Rossetti. Structify-net : Random graph generation with controlled size and customized structure. *arXiv preprint arXiv :2306.05274*, 2023.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Alexandre Devillers and Mathieu Lefort. Equimod : An equivariance module to improve visual instance discrimination. In *The Eleventh International Conference on Learning Representations*, 2022.

[5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33 :21271–21284, 2020.

[6] Bolian Li, Baoyu Jing, and Hanghang Tong. Graph communal contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 1203–1213, 2022.

[7] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1) :61–80, 2008.

[8] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins : Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.