Titre : Apprentissage autodidacte multi-vues pour la recherche d'information musicale

Mots clés : Apprentissage autodidacte, apprentissage equivariant, apprentissage contrastive, la recherche d'information musicale

Résumé : La recherche d'information musicale (MIR) s'est traditionnellement appuyée sur l'apprentissage supervisé, nécessitant de grandes quantités de données annotées, coûteuses à obtenir et souvent subjectives. L'apprentissage autodidacte (SSL) offre une méthode alternative en exploitant des données musicales non annotées pour apprendre des représentations transférables. Dans cette thèse, nous étudions le SSL pour l'apprentissage de représentations musicales à travers le prisme du multi-vues, où plusieurs transformations d'un même signal audio sont utilisées pour définir des objectifs d'apprentissage. Nous présentons d'abord trois grandes familles de méthodes de SSL appliquées à la MIR : joint-embedding architecture (JEA), l'apprentissage autodidacte équivariant et la modélisation par masquage. Nous étudions ensuite ces paradigmes. Premièrement, nous proposons un cadre de SSL équivariant pour l'estimation de la tonalité, introduisant les modèles STONE et S-KEY. En exploitant la transposition de la tonalité comme transformation structurée, ces méthodes apprennent des représentations tonales directement à partir de données non annotées et atteignent des performances compétitives par rapport aux approches supervisées. Deuxièmement, nous analysons l'apprentissage contrastif au sein de modèles JEA basés sur des transformors. Bien que l'objectif d'apprentissage ne soit appliqué qu'aux représentations globales, nous montrons que les caractéristiques au niveau des séquences présentent des propriétés locales émergentes, encodant des informations pertinentes pour des tâches temporellement localisées telles que la détection du tempo et l'estimation d'accords. Troisièmement, nous introduisons MT2, une architecture unifiée multi-tâches à multiples class-token, combinant des objectifs contrastif et équivariant au sein d'un même transformor. Cette approche produit des représentations complémentaires capturant à la fois des informations sémantiques globales et des structures tonales, tout en restant efficace en termes de calcul. Dans l'ensemble, cette thèse montre que le SSL multi-vues constitue un cadre puissant et flexible pour apprendre des représentations musicales riches et transférables à partir de données audio non annotées. Elle souligne l'importance de combiner des paradigmes d'apprentissage complémentaires et de concevoir des objectifs adaptés à la musique afin de mieux capturer la nature multidimensionnelle du signal musical.

Title : Multi-view self-supervised learning for music information retrieval

Keywords : self-supervised learning, equivariant learning, contrastive learning, music information retrieval

Abstract : Music Information Retrieval (MIR) has traditionally relied on supervised learning, requiring large amounts of annotated data that are costly to obtain and often subjective.
Self-supervised learning (SSL) offers a compelling alternative by leveraging unlabeled audio to learn transferable representations. In this thesis, we investigate SSL for music representation learning through the lens of multi-view, where multiple transformations of the same audio signal are used to define learning objectives. We first review three main families of SSL methods applied to MIR: joint embedding architectures (JEA), equivariant self-supervised learning, and masked modeling. We then study these paradigms. First, we propose an equivariant SSL framework for tonality estimation, introducing the STONE and S-KEY models. By leveraging pitch transposition as a structured transformation, these methods learn tonal representations directly from unlabeled data and achieve performance competitive with supervised approaches. Second, we investigate contrastive learning within transformer-based models within JEA family. Despite applying the training objective only to global representations, we show that sequence-level features exhibit emergent local properties, encoding information relevant to temporally localized tasks such as beat tracking and chord estimation. Third, we introduce MT2, a unified multi-task multi-class-token architecture that combines contrastive and equivariant objectives within a single transformer using multiple class tokens. This approach produces complementary representations that capture both global semantics and structured tonal information, while remaining computationally efficient. Overall, this thesis demonstrates that multi-view SSL provides a powerful and flexible framework for learning rich and transferable musical representations from unlabeled audio. It highlights the importance of combining complementary learning paradigms and designing music-aware objectives to better capture the multi-dimensional nature of music.

COLLEGE ED SCIENCES
DOCTORAL DE L'INGENIERIE
PAYS DE LA LOIRE ET DES SYSTEMES

CENTRALE NANTES