

الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 08 ماي 1945 بسيدي بلعباس
École Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbès

In collaboration with the iCube Laboratory (UMR 7357 CNRS), Université de Strasbourg,
France



THESIS

To obtain the Master's degree

Field: **Computer Science**

Specialty: **Artificial Intelligence and Data Science (IASD)**

Theme

Optimal Transport-Regularised Knowledge Distillation and Synthetic Data Generation for Student–Teacher Model Compression

Presented by:
OMRANI Rabah

Supervised by:

Dr. Nassima Dif (ESI-SBA)

Supervisor

Pr. Cédric Wemmert (iCube, U. de Strasbourg)

Co-supervisor

Dr. Ali Ayadi (iCube, U. de Strasbourg)

Co-supervisor

Academic Year: 2025/2026

Abstract

The deployment of large neural networks on resource-constrained hardware has made knowledge distillation, the transfer of knowledge from a large teacher model to a compact student model, a central practical concern. Classical distillation methods remain fragile under two conditions that have become routine in modern practice: large capacity gaps between teacher and student, and architectural heterogeneity that prevents direct feature correspondence. This thesis addresses both shortcomings by reformulating distillation as a problem in the geometry of probability distributions. The central methodological contribution is a unified training objective combining cross-entropy, temperature-scaled response distillation, and an entropic optimal-transport loss acting on the empirical distributions of intermediate features; dimensional mismatch is resolved by learnable linear projections, and structural mismatch by an unbalanced transport extension. A complementary data-free pipeline synthesises training data directly from the teacher’s parameters, preserving the distributional and topological structure of its implicit data manifold through persistent homology. On the theoretical side, we derive a Lipschitz bound relating the Wasserstein feature distance to the Fisher-Rao output distance, together with a generalisation bound on the student’s risk. Empirically, optimal-transport-regularised distillation improves over vanilla knowledge distillation with high statistical significance, with the largest gains on cross-architecture and extreme-compression scenarios, and the data-free pipeline recovers over ninety percent of real-data distillation performance on standard benchmarks.

Keywords: optimal transport, knowledge distillation, model compression, student–teacher framework, synthetic data generation, information geometry.

Résumé

Le déploiement de réseaux de neurones volumineux sur du matériel à ressources limitées a fait de la distillation des connaissances, le transfert de connaissances d'un modèle enseignant volumineux vers un modèle étudiant compact, une préoccupation pratique centrale. Les méthodes classiques restent toutefois fragiles dans deux situations devenues courantes : un écart de capacité important entre l'enseignant et l'étudiant, et une hétérogénéité architecturale qui empêche toute correspondance directe des caractéristiques. Cette thèse aborde ces deux limites en reformulant la distillation comme un problème de géométrie des distributions de probabilité. La contribution méthodologique principale est un objectif d'entraînement unifié combinant l'entropie croisée, la distillation des sorties mise à l'échelle par la température et une perte de transport optimal entropique appliquée aux distributions empiriques des caractéristiques intermédiaires ; les différences de dimension sont résolues par des projections linéaires apprises, et l'hétérogénéité structurelle par une extension de transport déséquilibré. Un pipeline complémentaire, sans données, synthétise les données d'entraînement directement à partir des paramètres de l'enseignant, en préservant la structure géométrique et topologique de sa variété de données implicite grâce à l'homologie persistante. Sur le plan théorique, nous établissons une borne de Lipschitz reliant la distance de Wasserstein des caractéristiques à la distance de Fisher-Rao des sorties, ainsi qu'une borne de généralisation sur le risque de l'étudiant. Empiriquement, la distillation régularisée par transport optimal surpasse la distillation classique avec une forte significativité statistique, les gains les plus importants étant observés dans les scénarios inter-architectures et de compression extrême, et le pipeline sans données récupère plus de quatre-vingt-dix pour cent des performances obtenues avec les données réelles.

Mots-clés : transport optimal, distillation des connaissances, compression de modèles, cadre enseignant-étudiant, génération de données synthétiques, géométrie de l'information.

ملخص

أدى نشر الشبكات العصبية الكبيرة على أجهزة محدودة الموارد إلى جعل تقطير المعرفة، أي نقل المعرفة من نموذج معلم ضخم إلى نموذج طالب مُدجج، مسألةً عمليةً محورية. غير أنّ الطرائق التقليدية للتقطير تبقى هشةً في حالتين أصبحتا شائعتين في الممارسة الحديثة: اتّساع الفجوة في السعة بين المعلم والطالب، وتباين البنى المعمارية الذي يحول دون المطابقة المباشرة للسّمات. تعالج هذه الأطروحة هاتين النقطتين بإعادة صياغة التقطير بوصفه مسألةً في هندسة توزيعات الاحتمال. وتكمن المساهمة المنهجية الأساسية في دالة هدفٍ موحدة تجمع بين الإنتروبيا المتقاطعة، وتقطير المخرجات المعايير بدرجة الحرارة، وخسارة النقل الأمثل الإنتروبي المطبقة على التوزيعات التجريبية للسّمات الوسيطة؛ ويُعالج اختلاف الأبعاد عبر إسقاطات خطية قابلة للتعلّم، ويُعالج التباين البنيوي عبر امتدادٍ غير متوازن للنقل. كما نطوّر مساراً مكّماً لتوليد بيانات التدريب اصطناعياً انطلاقاً من معاملات المعلم مباشرةً، مع الحفاظ على البنية التوزيعية والطوبولوجية لمتعدّد طبّاته الضمني عبر الهومولوجيا المستمرة. وعلى الصعيد النظري، نشقّ حدّاً من نوع ليبشيتز يربط مسافة فاسرشتاين للسّمات بمسافة فيشر-راو للمخرجات، إضافةً إلى حدٍّ للتعميم على مخاطرة الطالب. وتُظهر التجارب أنّ التقطير المنظّم بالنقل الأمثل يتفوّق على التقطير التقليدي بدلالةٍ إحصائيةٍ عالية، مع تحقيق أكبر المكاسب في سيناريوهات اختلاف البنى والضغط الشديد، وأنّ المسار الخالي من البيانات يستعيد أكثر من تسعين بالمئة من أداء التقطير المعتمد على البيانات الحقيقية على معايير القياس المرجعية.

الكلمات المفتاحية: النقل الأمثل، تقطير المعرفة، ضغط النماذج، إطار المعلم-الطالب، توليد البيانات الاصطناعية، هندسة المعلومات.