Etude comparative des méthodes de prédiction de l'échelle de débit basées sur l'apprentissage pour le streaming vidéo adaptatif

Ahmed Telili[1] Wassim Hamidouche[1] Sid Ahmed Fezza[2] Luce Morin[1]

[1] Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
[2] National Higher School of Telecommunications and ICT, Oran, Algeria

Résumé

Le streaming adaptative HTTP (HAS) est de plus en plus utilisé dans les services de streaming vidéo, permettant une transition dynamique entre différentes qualités de flux grâce à l'échelle du débit, qui fournit une table de correspondance entre le débit cible et les différentes résolutions vidéo possibles. Plusieurs approches existent pour construire cette échelle, allant de la méthode simple et statique "one-size-fits-all" à celle basée sur un codage exhaustif pour chaque résolution sur une large plage de débit. Des méthodes d'apprentissage automatique ont été proposées pour prédire l'échelle sans nécessiter un codage exhaustif et coûteux. Cet article étudie différentes approches de prédiction de l'échelle en fonction du contenu. Les résultats, basés sur 200 vidéos compressées en HEVC, montrent que la méthode la plus efficace prédit l'échelle avec une perte minimale de débit (BD-BR de 1,43 %) comparée à l'approche exhaustive.

Mots clefs

Compression video, échelle de débit, diffusion vidéo adaptative, HEVC.

1 Introduction

Le contenu vidéo est le media dominant sur internet, représentant 82% du trafic mondial [1].Le marché de la vidéo à la demande (VoD) devant connaître une croissance importante, les fournisseurs de vidéo investissent des ressources dans l'optimisation du processus d'encodage afin de garantir une meilleure qualité et une plus faible consommation d'énergie pour un streaming durable. Cependant, la qualité du contenu peut varier en fonction de facteurs tels que la bande passante du réseau, la résolution de l'écran et les conditions de visionnage.

Pour diffuser des vidéos de haute qualité à des débits aussi faibles que possible, les fournisseurs de services de streaming s'appuient sur des technologies et des normes de l'état de l'art telles que le streaming adaptative HTTP (HAS). Dans le cadre de la norme HAS, le contenu vidéo est divisé en courts segments et pré-encodé à différentes résolutions et niveaux de qualité avant d'être transmis. Les deux principales spécifications HAS, HLS [2] et DASH [3], dépendent de méthodes permettant de calculer des échelles de débit adéquates. Ces méthodes sont essentielles pour garantir un équilibre optimal entre la qualité vidéo et l'utilisation de la bande passante, ce qui améliore en fin de compte l'expérience de l'utilisateur.

Les méthodes traditionnelles utilisent des échelles de débit statiques, c'est-à-dire identiques pour tous les contenus, et qui ne s'adaptent donc pas aux complexités variables du contenu vidéo, ce qui conduit à des résultats sous-optimaux en termes de qualité et de consommation d'énergie. En revanche, l'optimisation du codage par titre proposé par Netflix [4] utilise des échelles de débit tenant compte du contenu et surpasse l'approche statique. Cette méthode implique un processus d'encodage exhaustif, à la recherche de l'échelle de débit optimale en testant systématiquement plusieurs paramètres d'encodage pour chaque titre. Cependant, elle nécessite une complexité de calcul importante et ne tient pas compte de la complexité visuelle au niveau de la scène.

Pour relever ces défis, des approches récentes basées sur l'apprentissage automatique (ML) ont été développées pour prédire une échelle de débit par scène sans qu'il soit nécessaire de procéder à un encodage exhaustif. Cette étude comparative vise à étudier la problématique de la prédiction de l'échelle de débit en fonction du contenu et à comparer les performances de diverses méthodes de prédiction basées sur l'apprentissage.

2 Données et méthodes étudiées

Cette étude est motivée par le succès remarquable des techniques basées sur l'apprentissage pour l'optimisation du codage des images et des vidéos. Bien que plusieurs approches utilisant des caractéristiques ad-hoc aient été explorées, peu de méthodes basées sur l'apprentissage profond ont été proposées en raison du manque de base de données à grande échelle. Dans cette étude, nous créons une base de données étiquetées à partir de 200 vidéo encodeés et menons une étude comparative pour évaluer les performances et la fiabilité des méthodes basées sur l'apprentissage automatique, en comparant à la fois les modèles basés sur des caractéristiques ad-hoc et ceux basés sur l'apprentissage profond.

Ce travail a été réalisé dans le cadre du projet DEEPTEC financé par la Région Bretagne.

2.1 Construction de la base de données

Collection de vidéos. Pour cette étude, il est essentiel de disposer d'une grande base de données vidéo avec des scènes variées. Nous avons utilisé la base proposé par [5] contenant 100 séquences vidéo UHD provenant de diverses sources, chaque séquence étant composée de 64 images à 60 fps. Bien qu'il soit suffisant pour les modèles basées sur les caractéristiques ad-hoc, il n'est pas adapté à l'entraînement des réseaux neuronaux profonds. Nous avons donc collecté 100 séquences UHD supplémentaires, dont 20 vidéos provenant de la base de données Waterloo IVC 4K [6] et 80 vidéos de haute qualité provenant de YouTube-UGC [7]. Les séquences ont été divisées en scènes à l'aide de l'outil PySceneDetect pour assurer que chaque séquence ne contient qu'une seule scène.

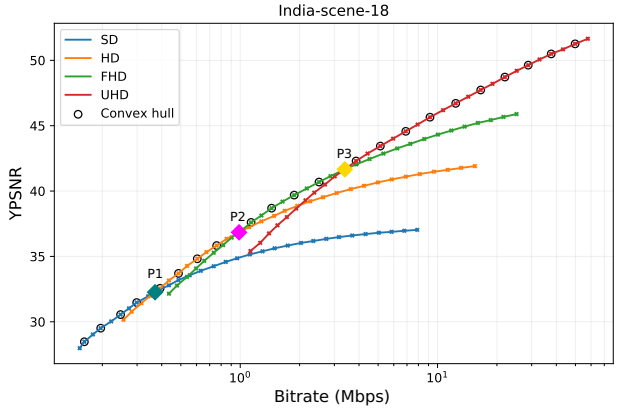

FIGURE 1 – *Exemple de courbes de débit-distorsion et de construction de la l'enveloppe convexe. P1, P2 et P3 désignent les points d'intersection entre les résolutions SD-HD, HD-FHD et FHD-UHD, respectivement.*

Construction de la vérité terrain. Pour créer une vérité terrain et identifier les débits croisés pour différentes résolutions, les séquences ont été réduites à l'échelle FHD, HD et SD à l'aide du filtre Lanczos-3 dans FFmpeg. Elles ont ensuite été encodées avec l'encodeur HEVC, générant en totale 24800 séquences vidéo encodés. Les flux binaires encodés ont été décodés, suréchantillonnés à la résolution native et évalués à l'aide des mesures de qualité YPSNR et VMAF. L'enveloppe convexe des courbes débit-distorsion obtenues pour une séquence détermine l'échelle de débit optimale pour cette séquence. Elle est définie par trois points SD-HD, HD-FHD, et FHD-UHD notés P1, P2, et P3, respectivement, comme illustré sur la figure 1.

2.2 Méthodes basées sur les caractéristiques ad-hoc

L'étape principale pour construire un modèle de prédiction de l'échelle de débit basé sur des caractéristiques est de sélectionner les caractéristiques les plus pertinents. Un ensemble initial de caractéristiques utilisées avec succès en compression vidéo a d'abord été utilisé. Les caractéristiques spatiales comprennent la GLCM, CF, SI et le bruit estimé (Noise), tandis que les caractéristiques temporelles comprennent TC, TI, NCC et les débits croisés prédits, (tableau 1). Pour réduire le nombre de caractéristiques et ne garder que les plus significatives, nous avons utilisé deux types d'algorithmes de sélection : les sélecteurs de caractéristiques basés sur des modèles ML, tels que RFR [8] et SVR [9], et l'élimination récursive des features (RFE) avec ExtraTrees [10] comme régresseur cible. Ces méthodes ont été appliquées sur 10 itérations de test-entraînement en utilisant un échantillonnage stratifié.

TABLEAU 1 – *Liste des caractéristiques et de leurs statistiques.*

Caractéristiques	Statistiques
Grey-Level Co-occurrence Matrix (GLCM)	F1.meanGLCM$_{con}$, F2.stdGLCM$_{con}$, F3.meanGLCM$_{cor}$, F4stdGLCM$_{cor}$, F5.meanGLCM$_{hom}$, F6.stdGLCM$_{hom}$, F7.meanGLCM$_{enr}$, F8.stdGLCM$_{enr}$, F9.meanGLCM$_{ent}$, F10.stdGLCM$_{ent}$
Temporal Coherence (TC)	F11.meanTC$_{mean}$, F12.meanTC$_{std}$, F13.stdTC$_{mean}$, F14.stdTC$_{std}$, F15.meanTC$_{skw}$, F16.stdTC$_{skw}$, F17.meanTC$_{kur}$, F18.stdTC$_{kur}$, F19.meanTC$_{entr}$, F20.stdTC$_{entr}$
Spatial Information (SI)	F21.meanSI, F22.stdSI
Temporal Information (TI)	F23.meanSI, F24.stdTI
Colorfulness (CF)	F25.meanCF, F26.stdCF
Noise	F27.meanNoise, F28.std Noise
Normalized Cross Correlation (NCC)	F29.meanNCC, F30.stdNCC
Predicted cross-over bitrates	F31.P3, F32.P2

2.3 Méthodes basées sur l'apprentissage profond

Récemment, les réseaux de neurones convolutifs (CNNs) profonds ont montré d'excellentes performances dans diverses tâches de vision par ordinateur. Cependant, en raison de leur complexité, une grande quantité de données est nécessaire pour les entraîner. Ainsi, nous avons utilisé des modèles pré-entraînés sur ImageNet [11] comme comme base pour extraire des descripteurs profonds.

Pour chaque séquence video, 16 images sont extraites avant d'appliquer une fenêtre glissante pour obtenir 156 patchs de taille 224×224. Chaque patch est utilisé comme entrée pour le CNN pour extraire les descripteurs discriminants. Après l'extraction, les descripteurs sont fournis en entrée de deux modèles LSTM pour capturer les dépendances à long terme entre les patchs et les images. L'entraînement est effectué pendant 200 époques, avec un taux d'apprentissage initial de $1e-4$ et MSE comme fonction de perte.

TABLEAU 2 – *Comparaison des performances des modèles évalués sur la base de données proposé. Le meilleur résultat est mis en évidence en* ***gras****.*

Métrique de la qualité	YPSNR / VMAF						
Méthode \Métrique	R2 ↑	SROCC ↑	PLCC ↑	Précision ↑	BD-BR vs GT ↓	BD-BR vs AL ↓	BD-BR vs RL ↓
ExtraTrees Regressor♦	**0.7635 / 0.6420**	**0.8174 / 0.6635**	**0.9000 / 0.8277**	**0.8779 / 0.8400**	**1.433% / 2.704%**	**-18.427%** / -18.827%	**-9.025% / -8.798%**
XGBoost♦	0.6165 / 0.5533	0.7560 / 0.6470	0.8278 / 0.7997	0.8578 / 0.8347	2.320% / 3.444%	-18.099% / -18.650%	-8.706% / -8.608%
Gaussian Process♦	0.6390 / 0.4292	0.7620 / 0.4918	0.8473 / 0.6983	0.8566 / 0.8012	1.740% / 5.254%	-18.244% / -18.328%	-6.286% / -7.688%
Random Forest Regressor♦	0.6758 / 0.5899	0.7993 / 0.6564	0.8440 / 0.8059	0.8671 / 0.8300	1.535% / 3.052%	-18.324% / **-18.887%**	-8.879% / -8.616%
Densenet 169★	0.4725 / 0.4216	0.6423 / 0.6167	0.7756 / 0.6433	0.8166 / 0.7901	3.380% / 3.820%	-15.669% / -15.892%	-8.169% / -7.851%
VGG16★	0.5172 / 0.4992	0.5236 / 0.5112	0.7652 / 0.7601	0.8223 / 0.8052	3.083% / 4.125%	-15.536% / -15.812%	-8.088% / -7.593%
ResNet-50★	0.4564 / 0.4045	0.5680 / 0.5367	0.7457 / 0.6962	0.8483 / 0.8278	2.424% / 2.969%	-15.806% / -15.941%	-8.300% / -7.810%
EfficientNet B7★	0.4237 / 0.3920	0.5649 / 0.5612	0.7159 / 0.6905	0.8004 / 0.7781	3.396% / 4.742%	-15.506% / -15.771%	-8.012% / -7.607%

★ Méthodes basées sur l'apprentissage profond . ♦ Méthodes basées sur handcrafted features.

3 Résultats

3.1 Protocole expérimental

Nous avons utilisé les échelles de débit spécifiques à chaque séquence générées par l'approche exhaustive pour fournir un point de référence à nos mesures de performance, appelé GT (Ground Truth). échelle de débit statique, appelée RL (Reference Ladder), obtenue en faisant la moyenne des échelles de débit GT sur toutes les séquences de la base d'apprentissage. Nous avons également utilisé l'échelle de débit statique proposée par Apple (AL) [12].
La performance est évaluée avec trois scores de corrélation standard : R2, PLCC et SROCC. La précision est également utilisée pour évaluer l'efficacité de chaque méthode pour prédire la résolution optimale pour les débits testés. Enfin, nous avons calculé le score BD-BR.

3.2 Résultats et analyses

Dans cette étude, nous avons entraîné et testé quatre modèles de régression ML classiques et des modèles CNN profonds connus (tableau 2). Les méthodes basées sur les caractéristiques ad-hoc surpassent les méthodes basées sur les réseaux de neurones profonds, probablement en raison de la quantité insuffisante de données nécessaire pour les entraîner. Tous les modèles basés sur l'apprentissage dépassent les approches statiques, permettant un gain moyen de BD-BR de 15,68% et une précision moyenne de 82% pour la prédiction des débits croisés. L'ExtraTrees Regressor obtient les meilleurs résultats, avec des gains de 18,42%/18,82% et 9,02%/8,79% par rapport aux approches AL et RL, respectivement. Tous les modèles d'apprentissage étudiés prédisent l'échelle de débit sans effectuer de processus d'encodage et peuvent donc être utilisés dans des applications de streaming vidéo.

4 Conclusion

Dans cet article, nous avons réalisé une étude comparative des méthodes d'apprentissage pour la prédiction d'échelles de débit pour le streaming video adaptatif. Nous avons créé une nouvelle base de données et testé plusieurs modèles d'apprentissage automatique pour la prédiction des débits croisés. Les résultats montrent que l'ExtraTrees Regressor surpasse les autres méthodes, avec un gain de débit de 18% par rapport à l'échelle statique. De plus, cette méthode réduit considérablement la complexité par rapport à la méthode exhaustive avec une perte minime de 1,4% en BD-BR. Pour les travaux futurs, nous prévoyons d'élargir la base de données en incluant d'autre séquences et d'autres types de codecs pour améliorer la performance des modèles basés sur l'apprentissage profond.

Annexe

Une version en anglais de 5 pages de cet article a été acceptée et publiée dans la conférence Picture Coding Symposium (PCS) 2022 [13].

Références

[1] U Cisco. Cisco annual internet report (2018–2023) white paper. *Cisco : San Jose, CA, USA*, 2020.

[2] Apple. Http live streaming. `https://developer.apple.com/streaming/`. Accessed on 2023-03-30.

[3] Iraj Sodagar. The mpeg-dash standard for multimedia streaming over the internet. *IEEE MultiMedia*, 18(4) :62–67, 2011.

[4] Per-Title Encode Optimization. `https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2`. Accessed on 2023-03-30.

[5] Angeliki V Katsenou, Joel Sole, et David R Bull. Content-gnostic bitrate ladder prediction for adaptive video streaming. Dans *2019 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2019.

[6] Zhuoran Li, Zhengfang Duanmu, Wentao Liu, et Zhou Wang. Avc, hevc, vp9, avs2 or av1 ?—a comparative study of state-of-the-art video encoders on 4k videos. Dans *International Conference on Image Analysis and Recognition*, pages 162–173. Springer, 2019.

[7] Yilin Wang, Sasi Inguva, et Balu Adsumilli. Youtube ugc dataset for video compression research. Dans *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019.

[8] Tin Kam Ho. Random decision forests. Dans *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[9] Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, et Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9, 1996.

[10] Pierre Geurts, Damien Ernst, et Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63 :3–42, 2006.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, et Li Fei-Fei. Imagenet : A large-scale hierarchical image database. Dans *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Best practices for creating and deploying HTTP live streaming media for the iPhone and iPad. `https://developer.apple.com/documentation/http_live_streaming/http_live_streaming_hls_authoring_specification_for_apple_devices`. Accessed on 2023-03-30.

[13] Ahmed Telili, Wassim Hamidouche, Sid Ahmed Fezza, et Luce Morin. Benchmarking learning-based bitrate ladder prediction methods for adaptive video streaming. Dans *2022 Picture Coding Symposium (PCS)*, pages 325–329, 2022.