FORSCHUNG

MASCHINELLES LERNEN FÜR PER-TITLE ENCODING

CHRISTOPH MÜLLER

Video-Inhalte unterscheiden sich in ihrer Komplexität – herkömmliche statische Encoding-Verfahren ignorieren jedoch die individuellen Videocharakteristiken und wenden für alle, teils sehr unterschiedliche Videodateien die gleichen Einstellungen an. Dies führt zu unnötig hohem Speicherbedarf und gesteigerten Übertragungskosten für Streaming-Anbieter. Der *Per-Title Encoding Ansatz* adressiert dieses Problem und hat das Potenzial, die Speicher- sowie Übertragungskosten von Video-Streams erheblich zu senken. Bisherige Lösungen erfordern in der Regel eine große Anzahl von Test-Enkodierungen, die entsprechende Rechenzeiten benötigen und daher zu erheblichen Mehrkosten führen. Dieser Artikel beschreibt eine Lösung, die den konventionellen Ansatz für Per-Title Encoding um Verfahren des maschinellen Lernens erweitert und somit in der Lage ist, vollständig auf rechenaufwändige Test-Enkodierungen zu verzichten.

● Video content differs in its complexity – and yet, conventional static encoding methods ignore the individual video characteristics and apply the same settings to all video files equally. This leads to unnecessarily high storage requirements and increased transmission costs for streaming providers. The Per-Title Encoding approach addresses this problem and has the potential to significantly reduce the storage and transmission costs of video streams. Previous solutions usually require a large number of test encodings, which require corresponding computing time and therefore lead to significant additional costs. This article describes a solution that extends the conventional approach to per-title encoding by machine learning methods and is thus able to completely dispense with computationally expensive test encodings.

Einleitung

Streaming-Portale wie beispielsweise YouTube, Netflix oder Amazon Prime Video sind verantwortlich für einen beträchtlichen Anstieg der Datenübertragung im Internet. Netflix allein verursachte in 2018 fast 15 Prozent des globalen Internet-Verkehrs.[1] Mehr als eine Milliarde Stunden Video-Inhalte werden täglich durch YouTube Nutzer konsumiert.[2] Jedes einzelne angeschaute oder hochgeladene Video erhöht den Datenverkehr. Bis 2021 wird die Übertragung von Bewegtbildinhalten schätzungsweise über 81 Prozent des globalen Internetverkehrs ausmachen. Um die bei digitalem Videomaterial anfallenden Datenmengen an die im Internet begrenzten Übertragungskapazitäten anzupassen, finden vor der Übertragung eine ganze Reihe an Optimierungen statt, um die Ausspielung des Videos möglichst effizient zu gestalten.

Das sogenannte Video-Encoding spielt hierbei eine wichtige Rolle. Video-**Encoding** ist die Umwandlung eines – in der Regel – unkomprimierten Ausgangs-Video-Signals in ein komprimiertes, digitales Format, welches mit verschiedenen Endgeräten oder Software-Video-Playern kompatibel ist und auf ihnen abgespielt werden kann – beispielsweise im Web-Browser, mit einem Smartphone oder auf einem Smart-TV. Die Kompression ist jedoch nicht verlustfrei. Der Encoding-Prozess verkleinert das Ausgangs-Signal zwar deutlich in seiner Speicher- und Übertragungsgröße, durch die Kompression verliert es aber in der Regel an Qualität. Die Herausforderung der modernen Video-Codecs, wie beispielsweise H.265, VP9 oder AV1, besteht genau darin, die Übertragungsgröße und die Bitrate des Ausgangs-Signals größtmöglich zu verkleinern und dabei gleichzeitig eine möglichst hohe Videoqualität beizubehalten. Um dies zu erreichen, müssen die optimalen Encoding-Parameter ermittelt werden.

Im Bereich des Online-Media Streaming haben sich mit der Zeit verschiedene Standards zur Übertragung von Videoinhalten etabliert, dazu zählt unter anderem das sogenannte **Adaptive Bitrate Streaming.** Hier wird die Qualität des Inhalts jeweils an die beim Zuschauer verfügbare Datenrate angepasst. Dazu werden mehrere verschiedene Qualitätsstufen (und damit verschiedene Größen und Bitraten) des Videos vorbereitet – von niedriger Auflösung und Bitrate zu hochaufgelösten Varianten – damit das Video vom Zuschauer in verschiedensten Netzwerk-Situationen und auf unterschiedlichen Endgeräten problemlos angeschaut werden kann. Der Video-Player entscheidet selbst anhand verschiedener Parameter, wie zum Beispiel der verfügbaren Bandbreite, welche Qualitätsstufe zum aktuellen Zeitpunkt abgespielt werden. Auf diese Weise verringert sich unter Umständen die Qualität des Videos bei unzureichender Bandbreite, lästige Unterbrechungen und „Ruckeln" beim Abspielen werden aber vermieden. Adaptives Streaming zeigt seine Stärken vor allem bei schlechten oder wechselnden Netzwerkbedingungen, wie beispielsweise während einer Zugfahrt, überlasteten Heimnetzwerken, oder bei gleichzeitigem Videoabruf in einem Netzwerk durch viele Zuschauer, wie etwa bei großen Sportveranstaltungen.

1) https://www.statista.com/chart/15692/distribution-of-global-downstream-traffic/
2) https://blog.youtube/news-and-events/you-know-whats-cool-billion-hours

Tabelle / Abbildung 1: Standard h.264 Encoding Ladder
Quelle: Apple Developer Handbook

Auflösung	Bitrate	Framerate
416 x 234	145	≤ 30 fps
640 x 360	365	≤ 30 fps
768 x 432	730	≤ 30 fps
768 x 432	1100	≤ 30 fps
960 x 540	2000	Wie Quellvideo
1280 x 720	3000	Wie Quellvideo
1280 x 720	4500	Wie Quellvideo
1920 x 1080	6000	Wie Quellvideo
1920 x 1080	7800	Wie Quellvideo

Um Inhalte für adaptives Streaming vorzubereiten, müssen diese in verschiedenen Bitraten und unterschiedlichen Auflösungen encodiert und anschließend in Segmente von typischerweise zwei bis zehn Sekunden Länge unterteilt werden, bevor sie an den Zuschauer ausgeliefert werden können. Das Resultat ist eine sogenannte **Encoding Ladder,** die verschiedene vordefinierte Auflösungen und die dazugehörigen Bitraten festlegt. Ein Beispiel für eine solche Encoding Ladder ist in Abbildung 1 zu sehen.

Die hier dargestellte Encoding Ladder wurde von Apple spezifiziert. Seit ihrer ersten Vorstellung im Jahr 2010 (Apple Tech Note TN2224) wurden die empfohlenen Bitraten und Auflösungspaare mehrfach aktualisiert, das Grundprinzip bleibt aber bis heute dasselbe. Nutzt ein Streaming-Anbieter diese Encoding-Ladder als Vorlage, müssen zu jedem Ausgangsvideo acht verschieden encodierte Versionen des Videos erstellt werden – von niedrig aufgelösten Varianten mit geringer Bitrate, bis hin zu Full-HD Varianten mit einer Bitrate von 7,8 Mbit/s.

Die steigende Nachfrage nach Online-Videos sowie der Trend zu hochauflösenden, adaptiv gestreamten Videos führt zu einem deutlichen Kostenanstieg für die Erzeugung, Speicherung und Übertragung der Inhalte. Um möglichst alle relevanten Zielplattformen abzudecken, müssen Inhalteanbieter die Streaming-Formate HLS und DASH für die verschiedenen Endgeräte-Plattformen (PCs, Fernseher, sowie mobile Geräte) unterstützen.

Was ist Per-Title Encoding?

Das Konzept des **Per-Title Encodings** wurde in einem Blog Post von Netflix im Jahr 2015 eingeführt[3]. Dabei handelt es sich um eine weitere Optimierung des Encoding-Prozesses. Statt alle Videos mit gleichen Einstellungen zu encodieren, werden beim Per-Title (pro Video) Encoding, speziell auf die jeweiligen Videos hin optimierte Encoding-Parameter verwendet. Dabei kommen verschiedene Strategien zum Einsatz die zum Beispiel die Einsparung der Bitrate oder die Einhaltung einer bestimmten Videoqualität zum Ziel haben.

Im Vergleich zu klassischen Encoding-Ansätzen, bei denen für alle Arten von Inhalten die gleiche, vordefinierte Encoding-Ladder verwendet wird, reduziert Per-Title Encoding die Speicher- und Übertragungskosten deutlich. Videos niedriger Komplexität sind einfach zu encodieren und können so mit deutlich niedrigeren Bitraten bei subjektiv gleich oder teils besser empfundener Qualität bereitgestellt werden.

3) https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2

Verschiedene Arten von Videoinhalten besitzen meist eine grundlegend verschiedene Komplexität und Charakteristiken. Sport und Action-Filme beispielsweise haben eine viel höhere Informationsdichte, schnelle Szenenwechsel und viel Bewegung im Bild und brauchen somit typischerweise mehr Bitrate, da weniger Möglichkeiten bestehen, redundante Teile des Videos effektiv zu komprimieren.

Naturdokus oder Animationsfilme hingegen haben wesentlich mehr Redundanz, sich wiederholende Muster, langsame Schwenk-Bewegungen und weniger komplexe Szenen im Bild, sodass der Video-Codec besser Möglichkeiten hat, das Video zu komprimieren, ohne dabei an Qualität zu verlieren. Statt also diese offensichtlich verschiedenen Inhalte mit den gleichen Encoding-Einstellungen zu komprimieren, passt man die Encoding-Einstellungen je nach zu komprimierendem Inhalt an, um so möglichst viel Bitrate und damit Speicherplatz- und Übertragungskosten zu sparen und trotzdem noch die höchstmögliche Qualität an den Zuschauer auszuliefern.

Wie funktioniert Per-Title Encoding?

Per-Title Encoding kann man in drei grundlegende Schritte unterteilen:

1. Das Erstellen von Test-Encodes

Zu Beginn werden sogenannte Test-Encodes des Ausgangsvideos erstellt – eine ganze Reihe an encodierten, also komprimierten Varianten mit verschiedenen Encoding-Einstellungen. Dabei werden vor allem die wichtigsten Parameter wie Bitrate, Auflösung oder verschiedenste Meta-Parametern des Codecs selbst variiert.

2. Berechnung der Qualitätswerte

Um die erstellten Test-Encodes objektiv miteinander vergleichen zu können, wird anschließend für jedes Video eine Qualitätsmetrik wie VMAF (Video Multi-Method Assessment Fusion) oder PSNR (Peak Signal to Noise Ratio) berechnet. Hierbei handelt es sich um Kennzahlen, welche die wahrgenommene Qualität des encodierten Videos im Vergleich zu seinem Ausgangssignal beschreiben.

3. Auswahl der optimalen Encoding-Einstellungen

Basierend auf allen berechneten Qualitätswerten der Test-Encodes werden schließlich die Videos mit einem optimalen Verhältnis zwischen Bitrate, Auflösung und Qualität gewählt. Die Auswahl erfolgt anhand der sogenannten konvexen Hülle. Hierbei handelt es sich um die kleinste Menge, die alle berechneten Test-Encodes umschließt. Die konvexe Hülle spiegelt die idealen Bitrate-Auflösungspaare wider und ermöglicht, eine optimierte Encoding Ladder für ein Ausgangs-Video zu bestimmen.

In Abbildung 2 sind verschiedene Test-Encodes eines Videos visualisiert. Die x-Achse zeigt die Bitrate der Test-Encodes, die y-Achse stellt die dazugehörige, berechnete Qualitätsmetrik (VMAF) dar. Das Ausgangs-Video wurde in sieben verschiedenen Auflösungen und jeweils zwölf verschiedenen Bitraten pro Auflösung encodiert, was insgesamt zu 84 einzelnen Test-Encodes führt. Erwartungsgemäß haben niedrige Auflösungen wie 320x240 Pixel bei geringen Bitraten unter 1 Mbit/s auch schlechtere VMAF-Werte als die Full-HD Test-Encodes mit über 6 Mbit/s. Auffällig ist, dass nahezu jede Auflösung eine Bitraten-Region hat, die andere Auflösungen qualitativ übertrifft und andersherum auch immer Regionen, bei denen andere Auflösungen bei gleicher Bitrate bessere Qualität liefern. Die optimale visuelle Qualität für ein Video mit mehreren Qualitätsstufen ist zu erwarten,

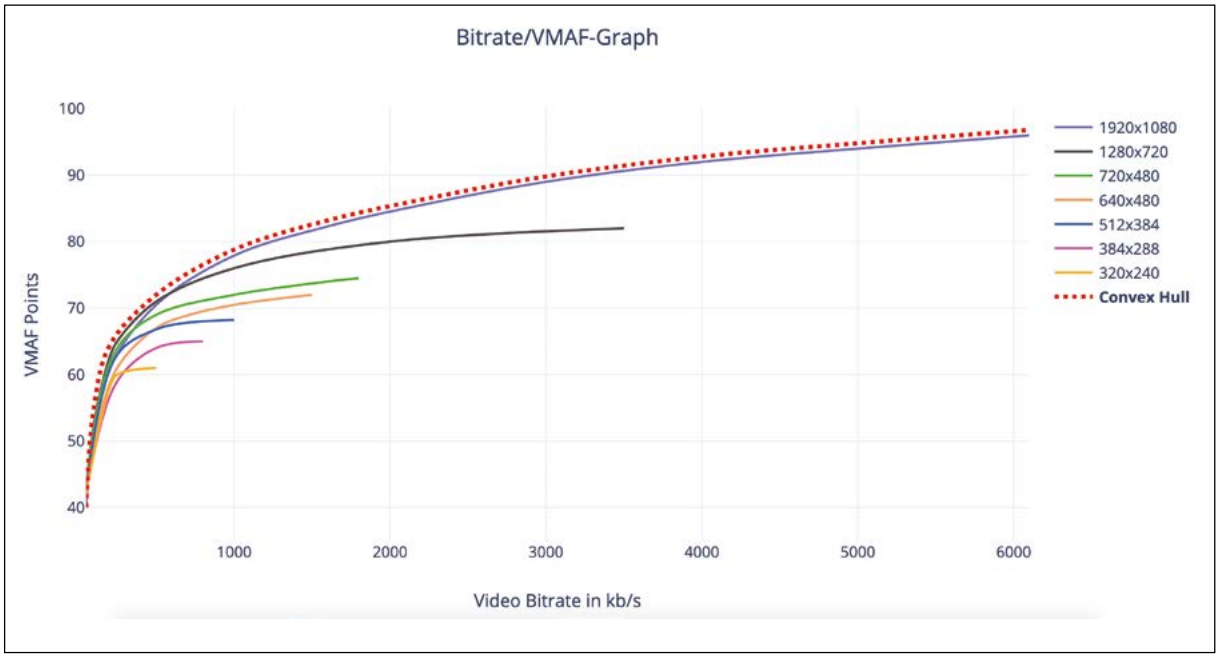

Abbildung 2: Visualisierung der konvexen Hülle am Beispiel mehrerer Test-Encodes eines Ausgangs-Videos

wenn Bitrate-Auflösungspaare identifiziert und encodiert werden, die so nah wie möglich an der konvexen Hülle (im Graphen rot dargestellt) liegen.

Während die Berechnung der optimalen Bitrate-Auflösungspaare mittels der konvexen Hülle zwar verlässliche Ergebnisse liefert, stellt die große Menge an Test-Encodes, die für die initiale Komplexitäts-Analyse notwendig ist, den größten Nachteil der konventionellen Herangehensweise dar. Um ausreichend Informationen für die Analyse zu sammeln, müssen typischerweise pro Ausgangs-Video zwischen 80 und 100 Test-Encodes berechnet werden. Die Erstellung dieser komprimierten Varianten ist äußerst rechenintensiv und zeitaufwändig.

Die anschließende Qualitätsmetrikberechnung, welche das Ausgangsvideo mit der komprimierten Variante Frame für Frame vergleicht, ist sehr zeitintensiv. Dieser Vergleich wird für jedes einzelne Test-Encode durchgeführt und benötigt entsprechende Rechenleistung.

Maschinelles Lernen für Per-Title Encoding

Um das Per-Title Encoding Verfahren weiter zu optimieren und vor allem das Problem der aufwändigen Test-Encodes zu adressieren, kommen bei der Lösung von Fraunhofer FOKUS Methoden des maschinellen Lernens zum Einsatz. Ziel dieser Herangehensweise ist es, vollständig auf Test-Encodes und vorausgehende Qualitätsmetrik-Berechnungen verzichten zu können und stattdessen statistische Vorhersagen zu nutzen, um die optimalen Encoding-Parameter für ein beliebiges Ausgangsvideo zu bestimmen.

Die Basis für diese automatisierten Vorhersagen bilden Algorithmen, die mit Hilfe von Trainingsdatensätzen statistische Modelle entwickeln, die Muster und Gesetzmäßigkeiten in Videos erkennen. Mit Hilfe dieser Modelle werden anschließend neue, bisher unbekannte Videos effizient beurteilt. Die so trainierten Modelle sind in der Lage, anhand extrahierter Charakteristika eines Videos, Aussagen über die perzeptuelle Qualität verschiedener Qualitätsstufen zu treffen. Die Qualitätsstufen entsprechen verschiedenen Bitraten/Auflösungs-Kombinationen. Die zuvor notwendigen Test-Encodes entfallen gänzlich.

Die Entwicklung eines solchen statistischen Modells lässt sich grundlegend in fünf Phasen einteilen:

1. **Sammeln der Daten:** Um ein Modell erstellen zu können, müssen zunächst Daten für das Training erzeugt und gesammelt werden.
2. **Bereinigung der Daten:** Zum Training des Modells müssen die Daten bereinigt und vorab auf Plausibilität geprüft werden (beispielsweise müssen fehlerhafte Messwerte korrigiert oder entfernt werden, welche unter anderem durch inkorrekte oder unvollständige Metadatenextraktion aus einem Video entstehen können).
3. **Trainieren des Modells:** Die nun vorliegenden Daten werden in mehreren Iterationen dazu genutzt, ein statistisches Modell zu entwickeln.
4. **Testen des Modells:** Mit einem neuen Testdatensatz wird die Güte des Modells geprüft und damit festgestellt, wie geeignet das Modell für statistische Vorhersagen ist.
5. **Verbesserung:** Sobald ein Modell trainiert ist, kann es fortlaufend anhand neuer Daten weiter angepasst und verbessert werden. Das ist beispielsweise notwendig, wenn sich Attribute der Datensätze ändern oder neue Attribute hinzukommen.

Preprocessing & Training der Modelle

Um möglichst genaue Modelle zu erstellen, werden sie auf Videos verschiedener Inhaltstypen – beispielsweise Sportinhalte, Dokumentationen, oder Nachrichtensendungen – und unterschiedlicher Enkodierungseinstellungen trainiert. Aus diesen, in der Regel unkomprimierten Videos werden 31 verschiedene Merkmale extrahiert. Darunter allgemeine Video-Metadaten wie Auflösung, Speichergröße, Video-Codec und spezifische Video-Merkmale. Letztere bestehen, unter anderem, aus der Anzahl der Szenenwechsel (basierend auf einer bestimmten Wahrscheinlichkeit), Farbhistogrammen, Helligkeitswerten, Inhaltstyp(en), Klassifikationswerten bzw. Kategorien und sogenannten Labels zur Inhaltsbestimmung, räumlichen und zeitlichen Merkmalen und mehr. All diese Informationen ermöglichen eine Beurteilung der Komplexität des Videos und somit schließlich die Vorhersage von Qualitätswerten für bestimmte Kombinationen aus Bitrate und Auflösung, woraus sich Encoding-Ladder ableiten lässt, welche auf die Charakteristika des Ausgangs-Videos optimiert ist. Als Teil des maschinellen Lernprozesses werden die gesammelten Daten bereinigt und anschließend in Trai-

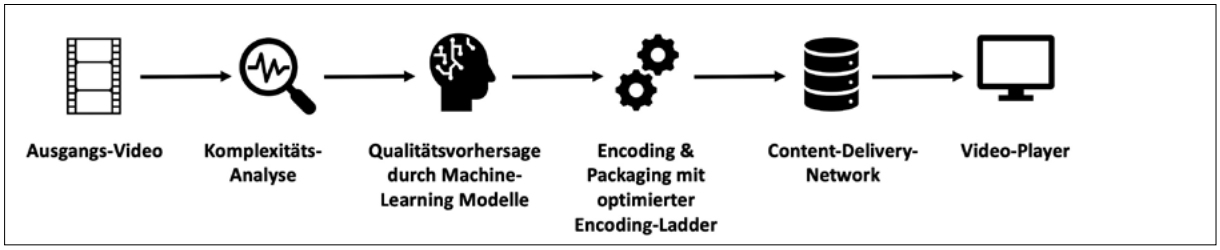

Abbildung 3: Der Per-Title Encoding Workflow von Fraunhofer FOKUS mit Unterstützung durch Machine-Learning Modelle

nings- und Testdatensätze aufgeteilt, wobei hier mit einem typischen 80-20-Split gearbeitet wird – 80 Prozent der Daten werden für das Training verwendet, 20 Prozent für die Validierung der Modelle. Die extrahierten Videoattribute werden dann merkmalsbezogen bearbeitet, so dass weitere einzigartige Videomerkmale festgelegt werden können. Dazu zählt beispielsweise die Kombination von Breite und Höhe zur Bildung der Auflösungsgröße. Dieser Prozess erfordert spezifisches Domänenwissen (zum Beispiel über Videocodierungsattribute) und kann die Leistung von Modellen des maschinellen Lernens erheblich verbessern. Da mehrere der numerischen Merkmale aus verschiedenen Werte-Bereichen bestehen, werden bestimmte Merkmale skaliert, normalisiert oder standardisiert, um sie untereinander vergleichbar zu machen.

Zur Validierung der Modelle und Bestimmung der Genauigkeit der Vorhersagen werden verschiedene Maße zur Prognosegüte eingesetzt. Hierbei kommt unter anderem der mittlere quadratische Wurzelfehler (Root Mean Squared Error, kurz RMSE) zum Einsatz. Anhand dieser Kenngröße werden regelmäßige Überprüfungen durchgeführt, um die Modelle kontinuierlich zu optimieren. Beispielsweise wird während des Trainings mit dem Ziel gearbeitet, einen RMSE-Wert von zwei zu erreichen. In diesem Fall entspricht der RMSE der Abweichung von vorhergesagten zu gemessenem VMAF Werten – ein RMSE von zwei bedeutet also, dass das Modell Vorhersagen mit einer Genauigkeit von zwei VMAF Punkte treffen kann. Da erst ab circa sechs VMAF Punkten einen spürbaren Qualitätsunterschied (Just Noticeable Difference, kurz JND) wahrgenommen wird, ist diese Toleranz hier völlig ausreichend. Qualitätsmessungen, die an neuen Inhalten automatisch durchgeführt werden, werden in einer Rückkopplungsschleife in den Trainings-Prozess zurückgespielt. Hierdurch können die Modelle wiederum kontinuierlich verbessert werden, um sich beispielsweise an neue Inhalte oder andere Ausgangsparameter anzupassen.

Nach diesem Vorgehen wurden die folgenden Modelle entwickelt:

- **Feed-forward, fully connected neural network** (FFFC): Ein universell einsetzbares neuronales Netzwerk, bei dem sich die Knoten in eine Richtung bewegen: von der Eingabeschicht über die verborgenen Schichten bis hin zur Ausgabeschicht. Dieses Modell ist robust in seiner Fähigkeit, potenziell fehlende Eingabewerte zu unterstützen (auch als „Fehlertoleranz" bekannt), und führt nach dem Trainieren zu einem geringeren Speicherplatz (einige KBs gegenüber Hunderten von MBs/GBs). Das FFFC-Modell ist jedoch anfällig für Überanpassungsprobleme („Overfitting"). Wenn akkurate Vorhersagen getroffen werden sollen, müssen die Eingabe-Daten daher zwingend normalisiert werden, was üblicherweise zu einem höheren Zeitaufwand für die Hyperparameter-Optimierung sowie zur Notwendigkeit einer größeren Datenmenge führt.
- **Convolutional Neural Network** (CNN): Ein neuronales Netz, das hauptsächlich für die Bilderkennung und Videoverarbeitung verwendet wird. Im Gegensatz zum FFFC-Modell unterstützt das CNN-Modell 3 Dimensionen (Breite, Höhe und Tiefe) und ist in Bezug auf die Verarbeitung flexibler und videofreundlicher. Ohne eine starke GPU ist die Trainingszeit jedoch langwierig und kann rechenintensiv sein.
- **XGBoost** (XGB): Ein Ensemble, das aus schwachen Vorhersagemodellen (im Allgemeinen Entscheidungsbäumen) besteht und stufenweise aufgebaut ist, die durch die differenzierbare Verlustfunktion optimiert werden können. Dieses Modell erfordert nicht so viel Normalisierung für das Trainieren von Videoattributen, jedoch können bestimmte Attributkodierungsmethoden (zum Beispiel eine One-Hot-Kodierung) die Leistung schwächen.
- **Stacked Model:** Dieses gestapelte Modell besteht aus 3 separaten Modellen, die miteinander kombiniert („gestapelt") werden: lineare Regression, Random Forest (eine verbreitete Methode zur Klassifikation und Regression von Datensätzen) und XGBoost. Diese Kombinationstechnik ist im Hinblick auf ihr Hauptkonzept flexibel, indem sie grundlegende mit fortgeschrittenen Ansätzen kombiniert. Aufgrund ihrer Komplexität erfordert jedoch jedes einzelne Modell mehrere Iterationen der Verfeinerung, um gemeinsam eine gute Leistung erzielen zu können.

Durch mehrere Iterationen des Modelltrainings wurde erkannt, dass die spezifischen Attribute wie beispielsweise Farbhistogramme und räumlich-zeitliche Merkmale einen großen Einfluss auf die Qualitätswert-Vorhersagen haben. Darüber hinaus verhielten sich Videos mit dem Scantyp *„interlaced"* (Zeilensprungverfahren) unterschiedlich im Vergleich zu Videos desselben Inhalts mit dem Scantyp *„progressiv"* (Vollbildverfahren). Modelle, die mit vorwiegend Progressive-Scan-Videos trainiert wurden, hatten niedrigere RMSE-Werte, als die auf Interlaced-Scan-Videos trainierten Modelle. Infolgedessen wurden die Modelle angepasst, um beide Scantypen zu erfassen und den Qualitätswert noch genauer vorherzusagen.

Durch die gezielte Kombination der Vorhersagen der Modelle ist die Lösung somit in der Lage, für bisher unbekannte Videos die Qualitätsmetrik VMAF auf bis zu zwei Punkte Genauigkeit abzuschätzen. Dies ermöglicht, eine für beliebige Inhalte optimierte Encoding-Ladder zu generieren, ohne dafür aufwändige Test-Encodes berechnen zu müssen. Die Modelle bieten darüber hinaus auch niedrigere Bitratenschätzungen (im Vergleich zur statischen "one-size-fits-all"-Encoding-Ladder), so dass die tatsächlichen Bitraten nicht „verschwendet" werden, in dem Qualitäten oder Auflösungen ausgeliefert werden, die dem Zuschauer keinen spürbaren Qualitätsgewinn bieten. Zusätzlich unterstützt dieser Vorhersageprozess nicht nur On-Demand Videoin-

halte, sondern kann auch für Live-Video Streaming genutzt werden. Darüber hinaus eignet sich die Lösung auch für Vorhersagen auf Szenenbasis („Shot-Based" oder „Per-Scene Encoding"), bei dem im Gegensatz zum Per-Title Encoding nicht das gesamte Video mit den gleichen Encoding-Einstellungen encodiert wird, sondern die einzelnen Szenen selbst je nach Komplexität mit individuellen Encoding-Einstellungen encodiert werden.

Zusammenfassung und Ausblick

Die Online-Media Streaming Landschaft wird mehr denn je von adaptiven Streaming Technologien dominiert. Video-Inhalte werden von den Anbietern in verschiedenen Qualitätsstufen encodiert, und vom Video-Player automatisch anhand der verfügbaren Bandbreite die passende Qualität zum Abspielen des Videos ausgewählt. Dies bedeutet für Streaming-Dienstleister nicht nur weitere zeitliche Belastung, sondern auch höhere Kosten durch zusätzlich benötigten Speicherplatz und gesteigerten Rechenaufwand. Konventionelle Encoding-Ansätze sind zwar für Streaming-Anbieter einfach umzusetzen, führen aber durchaus zu „unnötigem" Datenverkehr – wenn beispielsweise höher als notwendige Bitraten oder Auflösungen ausgespielt werden, ohne dass der Zuschauer einen qualitativen Unterschied bemerkt.

Per-Title Encoding sieht jeweils für verschiedene Arten von Videoinhalten unterschiedliche Bitraten und Encoding-Einstellungen vor, um die optimale Videoqualität bei möglichst geringer Größe des Videos bzw. minimaler Datenrate zu erreichen. Im Vergleich zu klassischen Encoding Ansätzen, bei denen für alle Arten von Inhalten die gleiche, vordefinierte Encoding-Ladder verwendet wird, reduziert Per-Title Encoding die Speicher- und Übertragungskosten von Videostreams deutlich. Gleichzeitig kann durch Minimierung der Datenraten die Internetverbindung des Zuschauers entlastet und die Stabilität der Wiedergabe verbessert werden. Durch den Einsatz von Per-Title Optimierungen können – je nach betrachtetem Inhalt – bis zu 50 Prozent der Speicherkosten und bis zu 55 Prozent an Übertragungskosten gespart werden.

Per-Title Encoding löst bereits viele Probleme des statischen Encodings, bei dem meist eine vorgefertigte Encoding-Einstellung für jeden Inhalt verwendet wird, bedarf aber der recht rechenaufwändigen und zeitintensiven Erstellung von vielen Test-Encodes pro Video mit sich.

Die von Fraunhofer FOKUS vorgestellte Lösung erweitert den Per-Title Encoding-Ansatz um Methoden des maschinellen Lernens. Durch das Training und die gezielte Kombination verschiedener statistischer Modelle ist es somit möglich, vollständig auf die zeit- und rechenintensive Erstellung von Test-Encodes zu verzichten. Stattdessen sind die Modelle in der Lage, Vorhersagen zur visuellen Qualität verschiedener Qualitätsstufen eines Videos zu treffen. Dazu bedient sich die Lösung unter anderem weiteren automatisierten Prozessen wie die automatische Erkennung von Inhalten- und Art eines Videos, die Identifikation von Szenen und Szenenwechseln, sowie der automatisierter Extraktion von Metadaten, um für beliebige Inhalte Qualitätswerte vorherzusagen. Aus den Vorhersagen lassen sich anschließend für jeden einzelnen Inhalt optimierte Encoding-Einstellungen ableiten, welche für die jeweiligen Charakteristika des Videos die optimale Qualität bei möglichst geringer Bitrate liefern. Durch das kontinuierliche Neu-Trainieren der Modelle sind diese in der Lage, dynamisch auf neue Inhalte oder Zielparameter zu reagieren. Ein entscheidender Vorteil hierbei ist, dass durch den Wegfall der zeitaufwändigen Test-Encodes und dadurch schnelleren Analyse diese Lösung auch für das Live-Streaming einsetzbar ist.

Obwohl Per-Title Encoding im Bereich des Online-Media Streaming noch ein recht neues Konzept ist, wurden in den letzten Jahren schon die nächsten Evolutionen vorgestellt. „Shot-Based" oder **Per-Scene Encoding** wurde beispielsweise von Netflix[4] vorgestellt. Statt ein ganzes Video von Anfang bis Ende mit einer passenden Encoding-Ladder optimal zu enkodieren, wird hierbei das Video in kleinere Teilstücke heruntergebrochen. Um das Video dabei nicht in Segmente willkürlicher, fester Länge zu unterteilen, analysiert der Encoder den Videoinhalt und erkennt einzelne Szenen, denen dann je nach Komplexität ebenfalls eine eigene, optimierte Encoding-Ladder zugewiesen wird.

Die von Fraunhofer FOKUS vorgestellte Lösung stützt sich ebenfalls auf die statistische Erkennung von Szenen sowie Szenenwechseln im Video und ist somit auch in der Lage, Qualitätsvorhersagen auf Szenenbasis zu treffen und den Videoinhalt somit noch effizienter zu optimieren. Mittels Per-Scene Encoding sind Einsparungen von bis zu 55 Prozent der Speicherkosten und sogar bis zu 65 Prozent der Übertragungskosten möglich, ohne dabei an perzeptueller Qualität zu verlieren.

Während sich Per-Title oder Per-Scene Encoding hauptsächlich auf die inhaltliche Komplexität des zu enkodierenden Videoinhalts selbst konzentriert, werden aktuell Ansätze entwickelt, welche ein viel breiteres Spektrum an Informationen nutzen. **Context-Aware Encoding** bezieht zusätzlich auch Meta-Parameter über die bestehenden Netzwerkverhältnisse, Geräte-spezifische Informationen sowie Details der Anzeigeumgebung des Zuschauers in die Optimierung der Encoding-Ladder mit ein. Diese Daten können beispielsweise aus Quality-of-Experience (QoE) und Quality-of-Service (QoS) Metriken bestehen, welche Details wie die effektive Bandbreite des Nutzers, die Geräte zum Abspielen des Videos verwendet werden, und die Verteilung der tatsächlich abgespielten Bitraten über Ihre Encoding-Ladder beinhalten. Die gezielte Kombination dieser verschiedenen Kontextparameter ermöglicht es, eine noch effizientere Encoding-Ladder zu generieren, die nicht nur auf den Inhalt selber, sondern auch auf den jeweiligen Zuschauer und seine aktuelle Umgebung zugeschnitten ist.

www.fokus.fraunhofer.de

4) https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f

Quelle: Fraunhofer FOKUS

CHRISTOPH MÜLLER

ist Wissenschaftlicher Mitarbeiter bei Fraunhofer FOKUS, Geschäftsbereich Future Media and Applications
www.fokus.fraunhofer.de