Lukas Lasse Bickelmann
Matrikel-Nummer: 42673

Hochschule der Medien Stuttgart
AM3 - Aktuelle Fragen der Medientechnik
Wintersemester 2021/2022

Künstliche Intelligenz in der Videoübertragung

Abstract

Die Zahl der Anwendungen, welche künstliche Intelligenz im Bereich Videokompression nutzen, steigt. Es gibt eine große Vielfalt an verschiedenen Methoden, die an den unterschiedlichsten Punkten der Datenreduktion ansetzen. Dieser Artikel beleuchtet die aktuell am weitesten fortgeschrittenen Lösungsansätze, die Erzeugung von synthetischen Gesichtern in Videokonferenzen und das perzentuelle Precoding von Videomaterialien.

Einleitung

Instagram, Netflix oder Zoom - Der Videokonsum geht längst über klassische Fernsehabende hinaus und das Internet hat sich als Standardmedium der Übertragung etabliert. Für 2022 ist vorhergesagt, dass mehr als 80 Prozent des genutzten Internet Traffics von Endnutzern auf den Konsum von Videomaterial zurückzuführen ist [1]. Durch Trends wie UHD oder VR steigen die Bandbreiten gegenüber HD-Übertragungen weiter. Künstliche Intelligenz kann hierbei einen Ansatz bieten um den schnellen Anstieg der Datenmenge abzubremsen, indem die Effizienz der Übertragung gesteigert und mehr der menschlichen Wahrnehmung angepasst wird.

Eine Videokonferenz stellt andere Ansprüche als ein Filmeabend über eine Streamingdienst - Vielfältige Arten der Videoübertragung sind die Grundlage unterschiedlicher Ansätze um mit KI die Übertragung zu verbessern.

Das Ziel des Einsatzes von Künstlicher Intelligenz ist jedoch dasselbe: Höhere Qualität bei gleichbleibenden Datenmengen oder gleiche Qualität bei niedrigeren Datenmenge. In klassischen Videoproduktionsworkflows spricht man hier von Videokompression.

Grundlagen der Livevideoübertragung

An eine Videoübertragung, gerade in Echtzeitumgebungen, sind technisch höchste Anforderungen gestellt. Die Übertragung soll in nahezu Echtzeit erfolgen und dabei höchsten Qualitätsanforderungen entsprechen. Daher weisen Videoübertragungen in Echtzeit häufig hohe Bandbreiten auf. Die Faktoren Qualität, Latenz und Bandbreite stehen dabei in direkter Korrelation.

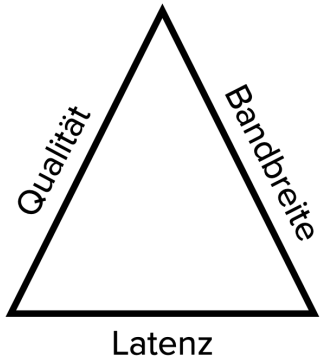

Abb. 01: Korrelation Qualität, Bandbreite, Latenz - Eigene Darstellung

In abgeschlossenen Produktionsumgebungen wie am Filmset oder im TV-Studio, kommt unkomprimiertes Video zum Einsatz. Bei dem Austausch und der Verbreitung von Videosignalen werden Videosignale komprimiert. In den meisten Fällen überschreite der Kompressionsfaktor hierbei den Faktor 1:100.Daher betrachtet man in diesem Fall die Signalkette, bestehend aus Encoder, Decoder und Übertragungsmedium, für die komprimierten Signalen.

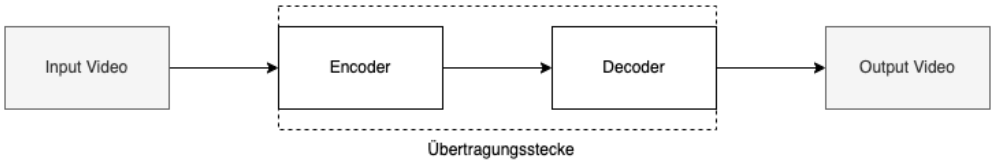

Abb. 02: Klassische Übertragung von komprimiertem Video - Eigene Darstellung

KI bei Videokonferenzen

Durch die Covid-Pandemie getrieben, stieg die Zahl der Videokonferenzen stark an. Ansätze die Bandbreite von Videosignalen in einer Videokonferenz zu minimieren zeigen erste vielversprechende Ansätze. Die benötigten Bandbreiten lassen sich hierbei um das zehn- bis hundertfache verringern.

Ausgangspunkt hierfür ist eine einzelnes übertragenes Standbild und dazu die Echtzeit-Information wie sich einzelne Keypoints innerhalb des Gesichts verändern. Diese Technik wird häufig als "Neural Head Synthesis" beschrieben. Hierbei werden die Mimik, die Kopfhaltung sowie die Blickrichtung eines Referenzbild mithilfe eines Eingabebildes in Echtzeit manipuliert [2]. Besondere Aufmerksamkeit liegt auf dieser Technologie unter dem Überbegriff "Deep Fake", wobei Gesichter von prominenten Menschen zu Aussagen manipuliert werden, welche synthetisch generiert sind.

Diesen Ansatz verfolgt ebenfalls NVIDIA bei ihrer Lösung für AI-basierte Videokonferenzen. Es wird ein Referenzbild des Teilnehmers übertragen und auf Basis dessen werden Informationen zu den Gesichtsmerkmalen der Person ermittelt. Hierbei handelt es sich um Schlüsselpunkte rund um Mund, Nase und Augen.

Die Datenübertragung während der Videoübertragung unterscheidet sich grundlegend zu klassischen Videokompressionsverfahren. Anstelle von aufeinanderfolgender Pixel, werden die veränderten Positionen der einzelnen Schlüsselstellen im Gesicht des Teilnehmers übertragen. Auf der Seite des Empfängers können die Daten mithilfe eines Generative Adversarial Network (GAN) auf das übertragenen Referenzbild synthetisiert werden. Dies benötigt hohe Rechenleistung auf der GPU. Hierbei wird ein Ansatz im dreidimensionalen Raum verfolgt, welcher die Ergebnisse signifikant verbessert. Darüber hinaus können die Schlüsseleigenschaften des Gesichts, beispielsweise Kopfneigung und Augenrichtung, manipuliert werden um ein immersiveres Erlebnis zu schaffen[3].

Die Grenzen von Neural Head Synthesis sind die menschlichen Gesichter, welche die Grundlage dieser Datenreduktion bilden. Das System ist auf die Schlüsselstellen eines Gesichtes ausgelegt und wenn es noch Besonderheiten wie Brillen und Hüte kompensieren kann, ist die Anwendung ohne Gesichter nicht realisierbar. Durch den speziellen Anwendungsfall einer Videokonferenz, lassen sich viele Daten reduzieren, jedoch auf Kosten der universalen Nutzen.

Precoding Algorithmen

Da häufig in Videomaterial deutlich mehr Inhalt vorhanden ist als menschliche Gesichter, müssen für universelleres Videomaterial andere Ansätze der Datenreduktion betrachtet werden. Precoding Algorithmen sind neuronale Netze, die Videomaterial perzeptuell anpassen, bevor dieses über die klassische Übertragungsstrecke übertragen wird.

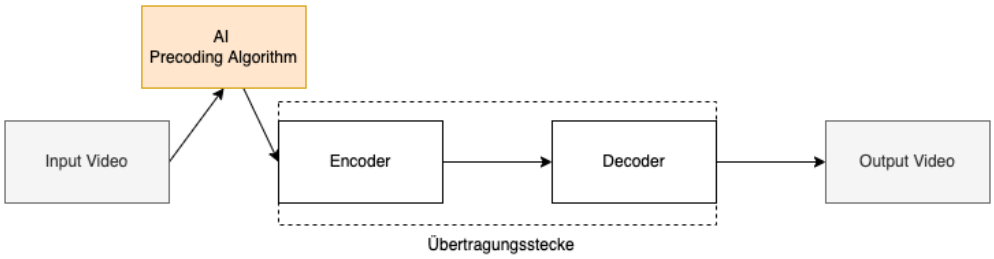

Abb. 03: Precoding AI Workflow - Eigene Darstellung

Encoder und Decoder bleiben hierbei in ihrer Funktion erhalten und Standardcodecs wie H264 oder H265 weiterverwendet werde. Hierbei lassen sich die Datenraten um 10 bis 25 Prozent gegenüber dem Standard En- und Decoder Workflow erreichen. In einer Echtzeitumgebung, mit ausreichend starker Hardware, beträgt die Laufzeit durch einen Precoding-Algorithmus 1 Frame, was 40ms entspricht[4].

Das Training des neuronalen Netzes erfolgt mit Videomaterial, welches das Precoding-Netz und anschließend eine Encoding-Simulation durchläuft. Dieser Prozess erfolgt auf der Luminanzkomponente eines Bildes, da diese für die hauptsächliche Wahrnehmung von spartialen Informationen relevant ist. Anhand der perzeptuellen Metrik SSIM und der Luminanzdifferenz aus encodiertem und ursprünglichen Bild wird der Fehler des Precodingnetzes minimiert[5].

Auf der Website www.bitsave.tech, bietet die Firma iSize eine kostenlose Demo von bis zu 20 Minuten KI-basiertem Precoding. Der Vergleich von Precoding-komprimierten-Signalen zu etablierten Encodern zeigt die Vorteile deutlich.

Hierzu wurde ein Video aus seinem Orginalformat ProRes 422 HQ in verschiedene H265 Bitraten encodiert. Für einen repräsentativen Vergleich wurden alle Encodings in CBR durchgeführt. Einerseits mit dem Adobe Media Encoder 2022 (AME) und andererseits über das Bitsave iSize Encoding. Das Ziel der Codierung war jeweils 2500 kbit/s und 5000 kbit/s.

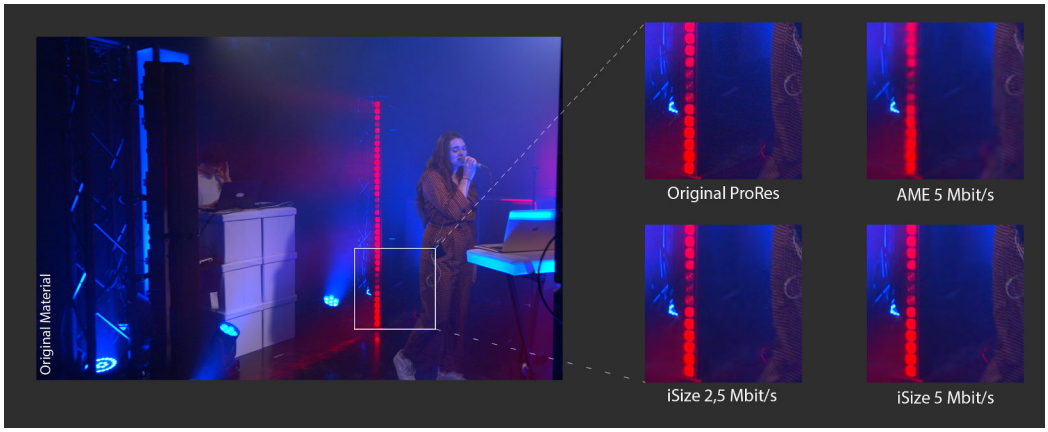

Abb. 04: Vergleich iSize und Adobe Media Encoder - Eigene Darstellung

Bei visueller Betrachtung werden die Vorteile der KI-basierten Encodierung sichtbar. Es erfolgt eine wesentlich bessere Abstufung in den Details bei halbierter Bitrate gegenüber der AME-Codierung. Feine Strukturen behalten ihre Zeichnung. Zudem Vergleicht man die 5 Mbit/s und 2,5 Mbit/s KI-basierte Kodierung, zeichnet sich die höhere Bitrate vor allem durch glattere Verläufe und weniger Halo-Artefakte aus.

Fazit/Ausblick

Zukünftig wird die Effizienz von KI-basierten Video-Codierverfahren weiter steigen. Es wird ein Effizenzsteigerung um 25% für die erste Generation von AI-Video-Codecs prognostiziert[6] . Dieser Fortschritt entspricht etwa der Evolution zwischen H264 und H265. Für einen Ende-zu-Ende-AI-Videocodec-Workflow, der viele Nutzer erreicht, muss zudem die Rechenleistung auf mobilen Endgeräten gegeben sein um dort die KI-Funktionalitäten zu gewährleisten um dann von den reduzierten Datenraten der Übertragung zu profitieren.

Einen Mehrwert könnte die Kombination verschiedener KI-Technologie bieten um die vielfältigen Anforderung an Videokompression universell abzudecken: Über Objekterkennung lassen sich einzelne Bereiche des Videomaterials unterschiedlich encodieren: Werden beispielsweise Gesichter erkannt, so können diese anders encodiert werden, als der Rest des Bildes.

Der Schwerpunkt für den Einsatz von KI zeigt sich am vielversprechendsten wenn Künstliche Intelligenz im Bereich zum Einsatz kommt um die menschliche Wahrnehmung zu modellieren. Um möglichst performante Systeme zu trainieren wird man um GAN-basierte Ansätze nicht herumkommen.

---

1. Cisco Annual Internet Report (2018–2023), https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html , *letzter Abruf: 26.03.2022*
2. M. R. Koujan, M. C. Doukas, A. Roussos, and S. Zafeiriou, “Head2Head: Video-based Neural Head Synthesis,” CoRR, vol. abs/2005.10954, 2020, [Online]. Available: https://arxiv.org/abs/2005.10954
3. T.-C. Wang, A. Mallya, and M.-Y. Liu, “One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing,” CoRR, vol. abs/2011.15126, 2020, [Online]. Available: https://arxiv.org/abs/2011.15126
4. iSize Datasheet v20211209: iSIZE BitSave: High-Quality Video Streaming at Lower Bitrates, https://www.isize.co/wp-content/uploads/2022/03/iSIZE-Datasheet_BitSave-v20211209_v4.pdf
5. A. Chadha, R. Anam, I. Fadeev, V. Giotsas, and Y. Andreopoulos, “Escaping the complexity-bitrate-quality barriers of video encoders via deep perceptual optimization,” in Applications of Digital Image Processing XLIII, 2020, vol. 11510, pp. 38–52. doi: 10.1117/12.2567549.
6. Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) - AI-Enhanced Video Coding (MPAI-EVC), https://mpai.community/standards/mpai-evc/evidence-project-description/ , *letzter Abruf: 26.03.2022*