深層学習を用いた映像伝送トラフィック削減技術の実験と考察

安藤 祐斗

富山県立大学　情報基盤工学講座
t815008@st.pu-toyama.ac.jp

June 17, 2021

本研究の背景

背景

スマートフォンの普及により,SNS サービスが急速に発展している. その中でもビデオ通話は表情やしぐさがわかりやすいためコミュニケーションが円滑に進めることができる.
しかし, 一般的に映像情報はデータ量が大きいため, ビデオ通話の使用率の上昇は通信帯域のひっ迫を招いてしまう. 現在, 映像情報は映像符号化技術を用いて映像情報を圧縮し伝送しているが, ビデオ通話の使用率が上昇することを鑑みると, 符号化技術だけでなく, ビデオ通話に特化してさらなるデータ量の圧縮を達成する仕組みが求められる.

研究の目的

目的

本研究では, 深層学習による超解像技術を用いてビデオ通話に生じるデータを削減する手法を提案する. 具体的には, 送信側ではカメラから得た映像情報の各ビデオフレームをダウンサンプリング（低解像度化）し解像度を落とすことで画像サイズを縮小, 映像符号化技術で圧縮し送信する. 受信側では受け取った各ビデオフレームを深層学習を用いて高解像度化する.

関連研究

超解像技術

本来ディスプレイに表示される画像よりもさらに詳細に画像を映し出すために, もとの画像から細部を予測する技術.

映像符号化技術

映像情報を圧縮する技術のこと.2018 年の時点で最先端の映像符号化として H265/HEVC が挙げられる.

GAN を用いない超解像技術

深層学習の画像生成モデルには変分推論モデル, 自己回帰モデル, 敵対的生成モデルがある. それぞれ VAE や PixelCNN 等が例に挙がる.

関連研究

GAN

敵対的生成ネットワーク (GAN:Generative Adverrsarial Network) は, 機械学習のひとつで, 生成モデルと判別モデルの 2 種類のニューラルネットワークが交互に精度を高め合うことによって, 入力データと非常によく似たデータを生成することができる.

GAN を用いた超解像技術

畳み込みニューラルネットワーク (CNN) を GAN に適応した DCGAN(Deep Convolution Generative Adversarial Network) や LPGAN(Lapia-cian Pyramid of Generative Adversarial Network),SRGAN(Super-Resolution Using a Generative Adversarial Network) などがある

提案手法

・送信側でダウンサンプリングした映像情報を圧縮・送信したあと, 受信側で顔の輪郭情報を損失関数に用いた DCGAN を利用して元の映像を高品質に復元する.
・特徴として, 低解像度化と高解像度化のプロセスにおいて各ビデオフレームにおける顔の輪郭情報を利用する点,DCGAN を用いた生成モデルを送受信端末間で共有する点が挙げられる.

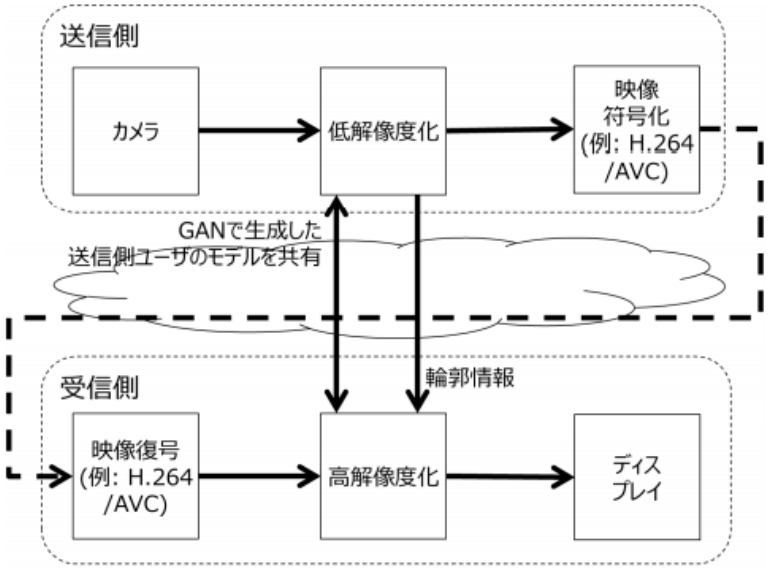

図 1: 提案システム

システム構成

DCGAN を用いた生成モデルの共有方法は, 次の 2 種類を想定している.

逐次生成方式

学習通信フェーズと高解像度通信フェーズの 2 つのフェーズから構成される. 学習通信フェーズにおける受信側の学習が終了すると, 学習終了の信号を送信側に送り高解像度化通信フェーズに移行する

事前共有方式

送信側ユーザに関する生成モデルをビデオ通話開始前に受信側に送っておくモデルである. 高解像度化通信フェーズと同じ動作を行う.

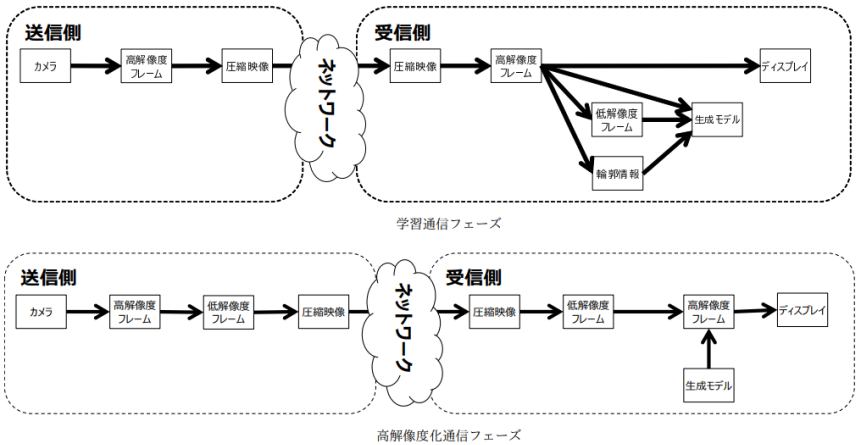

図 2: 学習通信フェーズと高解像度通信フェーズ

DCGANで用いる生成モデルと判別モデル

サイズ $W \times H \times C$ の元画像を I_R, これをダウンサンプリングした画像をサイズ $rW \times rH \times C$ の I_M とする. サイズはそれぞれ画像の幅, 高さ, チャネル数を表し,r はダウンスケーリング関数である.
また, 生成モデル $G(= G_{\theta_G}(I_M))$ は I_M とパラメータ θ_G を入力として元画像 I_R と類似する画像を生成するモデルである.
このパラメータ θ_G の最適化は, 損失関数を l,N 枚の元画像 $I_R^{(i)}, i = 1, ..., N$ と,N 枚のダウンサンプリング画像 $I_M^{(i)}, i = 1, ..., N$ を用いて以下の式で表される.

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l(G_{\theta_G}(I_M), I_R)$$

DCGAN で用いる生成モデルと判別モデル

生成モデル G の最適化のために判別モデル D を用いる.
判別モデルは与えられた画像が元画像 I_R による正解データ群と生成モデルが作成した画像 $G_{\theta_G}(i_M)$ による偽データ群のどちらかに属するかを判別する 2 値分類器で, 二つのモデルがお互いに制度を高めあうことで学習が行われる.
学習はゲーム理論を応用した次のミニマックス法によって最適化する.

$$\min_{\theta_G}\max_{\theta_D}\mathbb{E}_{I_R\sim P_{\mathrm{data}}(I)}[\log(D_{\theta_D}(I_R))]$$
$$+\mathbb{E}_{I_M\sim P_{mosaic}(I)}[\log(1-D_{\theta_D}(G_{\theta_G}(I_M)))]$$

$D_{\theta_D}(i_R)$ は入力画像 I_R が正解データ群に属する確率,$1-D_{\theta_D}(G_{\theta_G}(I_M))$ は低解像度画像 I_M から生成モデル G によって生成されたデータが偽データ群に属する確率,$\mathbb{E}[\cdot]$ は期待値を表す.

学習モデル

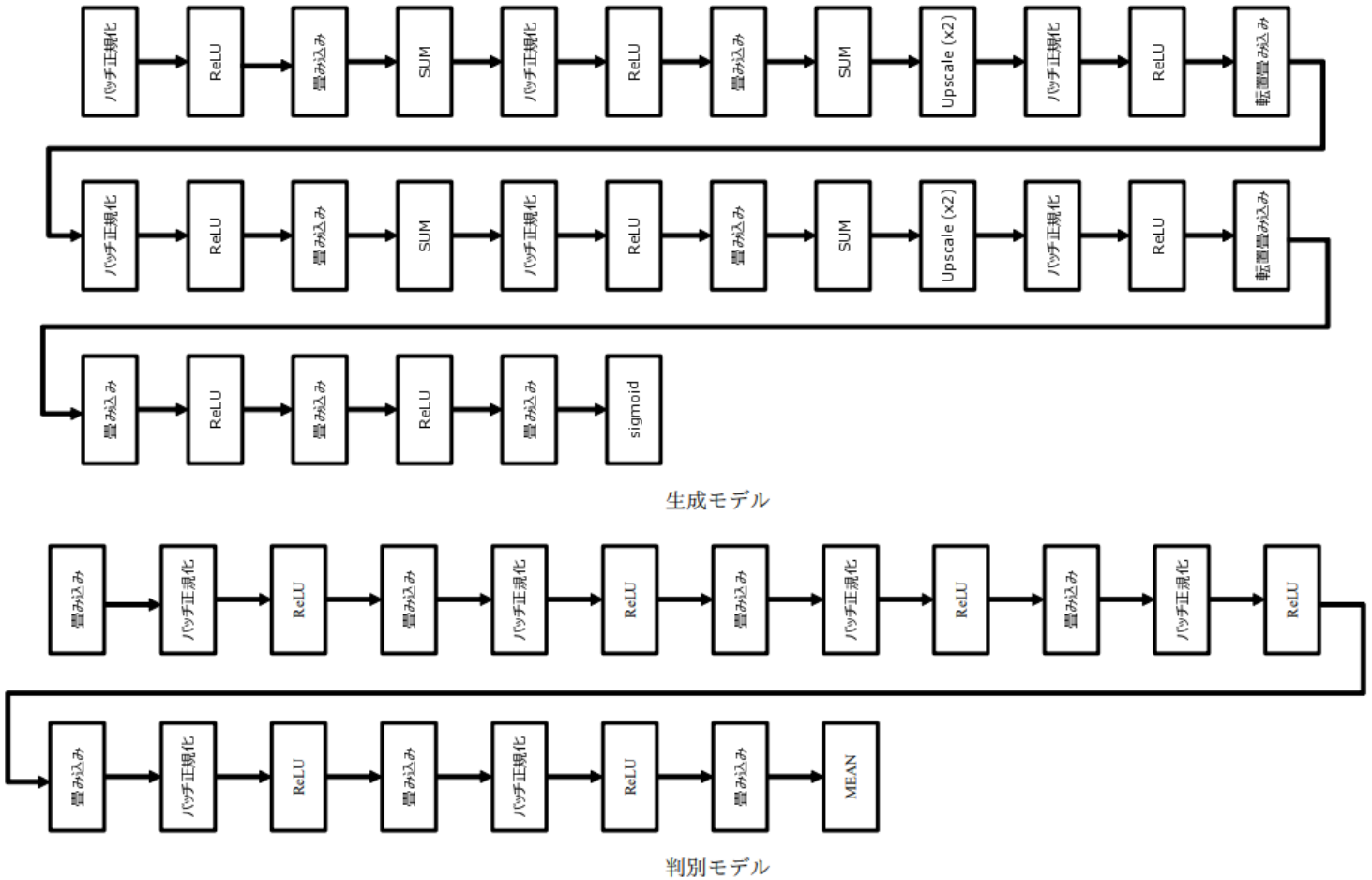

図 3: 生成モデルと判別モデル

各層の説明

バッチ正規化

入力データの各要素のバランスが取れていない場合に, 値の大きい要素が大きく影響してしまうといった問題を解消するために調整する仕組み. 要素ごとに平均 0, 分散 1 に正規化している.

ReLU

CNN によく利用される活性化関数. 入力が 0 より大きいときは入力の値を出力し, 入力が 0 以下のときは常に 0 を出力する.
ReLU$(x) = max(0, x)$

シグモイド関数

生物の神経細胞が持つ性質をモデル化した活性化関数.
sigmoid$(x) = \frac{1}{1+e^{-x}}$

畳み込み層

フィルタの濃淡パターンと類似した濃淡パターンが画像のどこにあるかを抽出する働きがある.
深層学習ではフィルタが重みで構成されているため, 抽出したい濃淡構造を学習することになる.

生成モデルの損失関数

生成モデルの損失関数 l は次の式で表される.

$$l = \alpha_1 l_{\text{adversarial}} + \alpha_2 l_{\text{pixel}} + \alpha_3 l_{\text{face}} \qquad (1)$$

α_1,α_2,α_3 は 0 から 1 の実数値をとるパラメータであり,$l_{\text{adversarial}}$ は敵対的ニューラルネットワークにおける生成器の損失であり, 以下の式で表される.

$$l_{\text{adversarial}} = \sum_{n=1}^{N} \log 1 - D_{\theta_D}(G_{\theta_G}(I_M))$$

l_{pixel} はピクセル単位での損失であり, 元画像をダウンサンプリングした画像 $I_{M_{(x,y)}}$ と生成器が生成した画像をダウンサンプリングした画像 $G_{\theta_G}(I_M)_{(\frac{x}{r},\frac{y}{r})}$ との画素値の差分の平均をとったものとして以下の式で表される.

$$l_{\text{pixel}} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} |I_{M_{(x,y)}} - G_{\theta_G}(I_M)_{(\frac{x}{r},\frac{y}{r})}|$$

l_{face} は顔の特徴点の損失である.HOG(histgram of oriented gradient) 検出器を顔の特徴点検出に使い,dlib と iBUG 300-W が公開しているデータセットを用いた.dlib では顔におけるあご, 右眉, 鼻, 右目, 左目, 口の合計 68 個の輪郭情報が検出できる.68 個の輪郭情報の点の座標を検出する関数を Ψ とすると,l_{pixel} は次の式で表される.

$$\frac{1}{68} \sum_{k=1}^{68} ||\Psi(I_R) - \Psi(G_{\theta_G})||$$

生成モデルの損失関数

l_{face} は顔の特徴点の損失である.HOG(histgram of oriented gradient) 検出器を顔の特徴点検出に使い,dlib と iBUG 300-W が公開しているデータセットを用いた.dlib では顔におけるあご, 右眉, 鼻, 右目, 左目, 口の合計 68 個の輪郭情報が検出できる.68 個の輪郭情報の点の座標を検出する関数を Ψ とすると,l_{pixel} は次の式で表される.

$$\frac{1}{68}\sum_{k=1}^{68}||\Psi(I_R)-\Psi(G_{\theta_G})||$$

また,dlib による顔の特徴点検出によって顔が検出できたときだけ l_{face} は計算できるため, 顔がうまく生成されない初期段階では l_{face} が計算できない. そこで, 顔が検出されない場合は次の損失関数式を用いる.

$$l=(1-\alpha)l_{\mathrm{adversarial}}+\alpha l_{\mathrm{pixel}} \qquad (2)$$

α は 0 から 1 の実数地をとるパラメータである.

評価環境

実際にビデオ通話映像を用いて性能評価を行う.

評価環境

評価に用いたテスト画像群は,PENTAX KS-2 を用いて撮影した 10 分間の動画から取得した. 撮影した動画の各ビデオフレームは JPEG を用いて圧縮した後, 解像度 80 × 80 画素と 160 × 160 画素にリサイズすることで 2 種類の入力画像群を作成.

学習

200000 枚作成した入力画像群から無作為に 16 枚を選択して学習に, 無作為に 8 枚を選択して評価に用いた.

パラメータ

損失関数のパラメータ α : 0.90
学習係数の初期値 : 0.0002
重みの初期値 : 正規分布から取得したランダム値
バイアスの初期値 : 0
学習パラメータの更新 : Adam を用いた

定性評価

図 4: 学習回数に対する生成画像の変化

定性評価

図 7 の補足

左からモザイク画像,bicubic 補完画像, 学習回数が 100 回,200 回,300 回,400 回,500 回,800 回,1000 回,2000 回,3000 回,4000 回,5000 回,8000 回,10000 回の生成画像, 元画像となっている.

モザイク画像

対象画像の解像度を縦横それぞれ 1/4 倍にした低解像度画像に Nearest Neighbor 補間を施したもの.

データサイズ削減効果

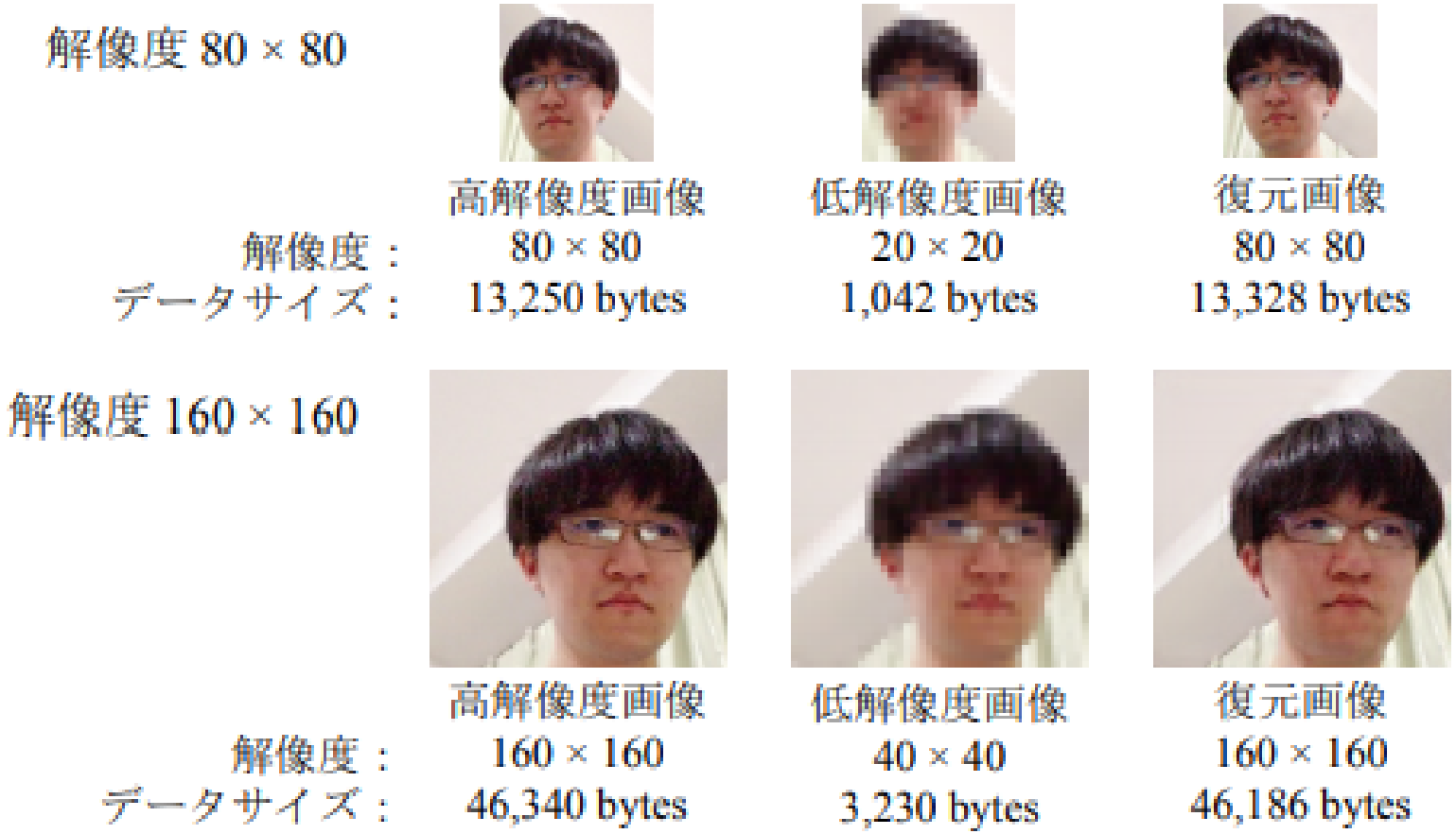

図 5: 評価に利用した画像とデータサイズ

画像品質

生成画像の品質について議論するため, 低解像度画像, 生成画像および bicubic 補間画像の WPSNR(Weighted Peak Signal-to-NoiseRatio) を比較する.
WPSNR とは,YCbCr 色空間のそれぞれの PSNR の真値に Y:Cb:Cr=8:1:1 の重み付き平均を行ってデシベル値に直したもの.

$$\mathrm{PSNR} = -10\log_{10}\frac{\mathrm{MSE}}{255^2}$$

$$\mathrm{MSE} = \frac{1}{nm}\sum_{i=0}^{n}\sum_{j=0}^{m}\{\mathrm{original}(i,j) - \mathrm{encoded}(i,j)\}^2$$

(n, m) は画像の縦横のピクセル幅,original(i, j) は高解像度画像におけるピクセル (i, j) の階調値,encoded(i, j) は復元後画像におけるピクセル (i, j) の階調値を表す.
MSE(Mean Squared Error) は元画像と復元画像の平均二乗誤差を表す.
ここで WPSNR は復元画像 8 枚それぞれについて PSNR を計算し, 真値で重み付けをしてから画像 8 枚の平均を計算したデシベル値とする.

画像品質

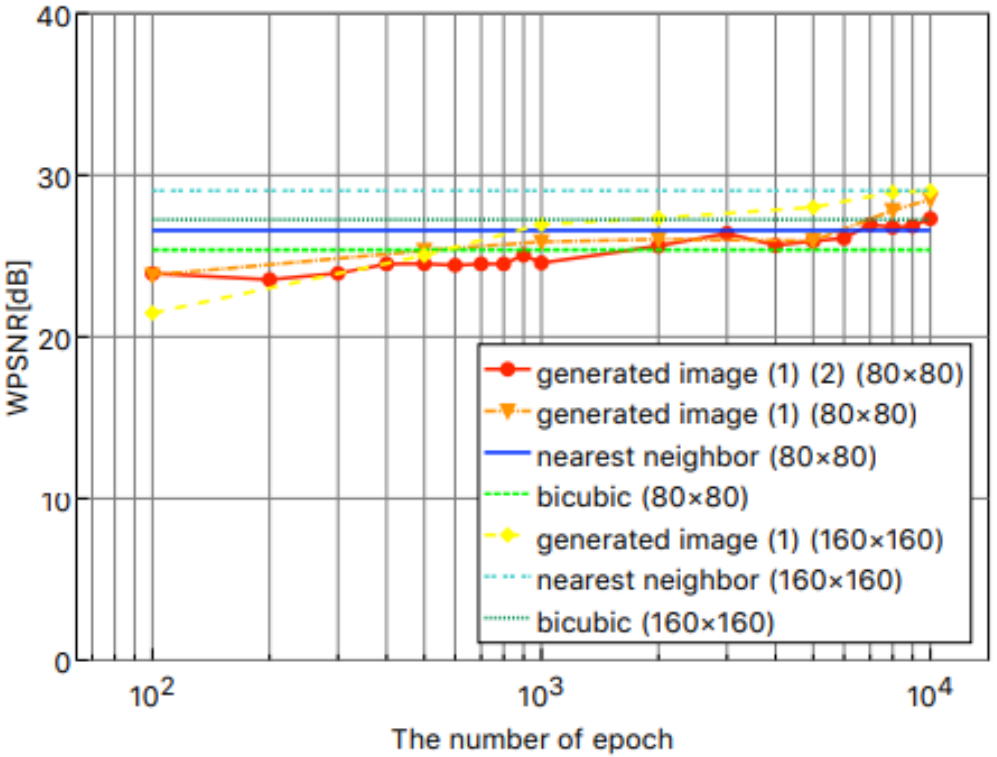

図 6: 学習回数と WPSNR との関係

画像品質

次に SSIM を用いて評価を行った.PSNR と比較して輝度, コントラスト, 構造を軸として各画素およびその周囲との相関を考慮している.

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

x, y はそれぞれ元画像と符号化後の画像における各画素を要素とするベクトル,μ_x, μ_y はそれぞれ画像 x, y の平均画素値,σ_x, σ_y はそれぞれ画像 x, y の標準偏差,σ_{xy} は画像 x, y の共分散,$C_1 = (255K_1)^2$,$C_2 = (255K_2)^2$,$K_1 = 0.01, K_2 = 0.03$ と表される.

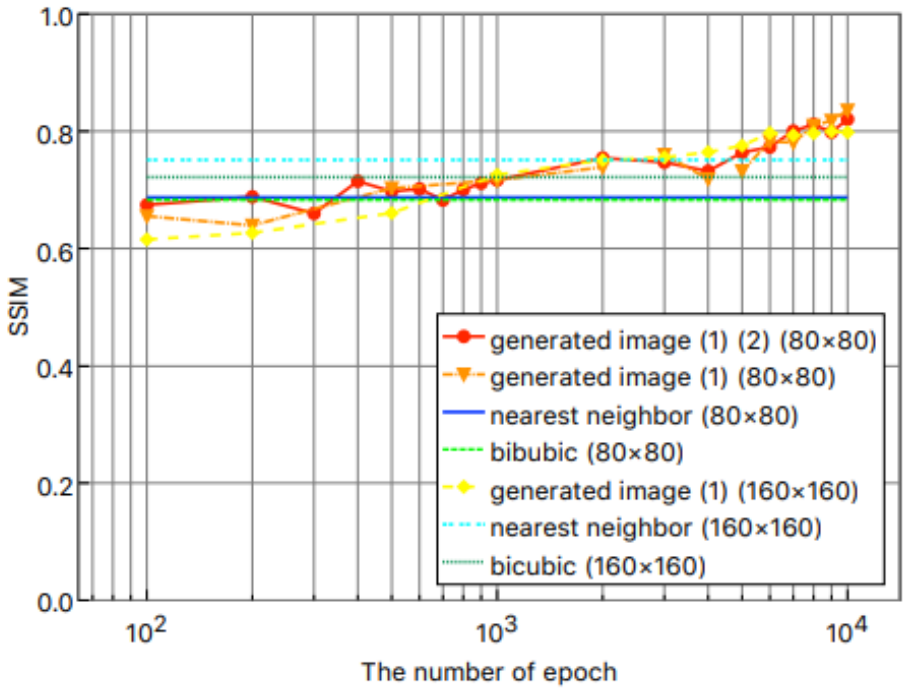

図 7: 学習回数と SSIM との関係

計算量の評価

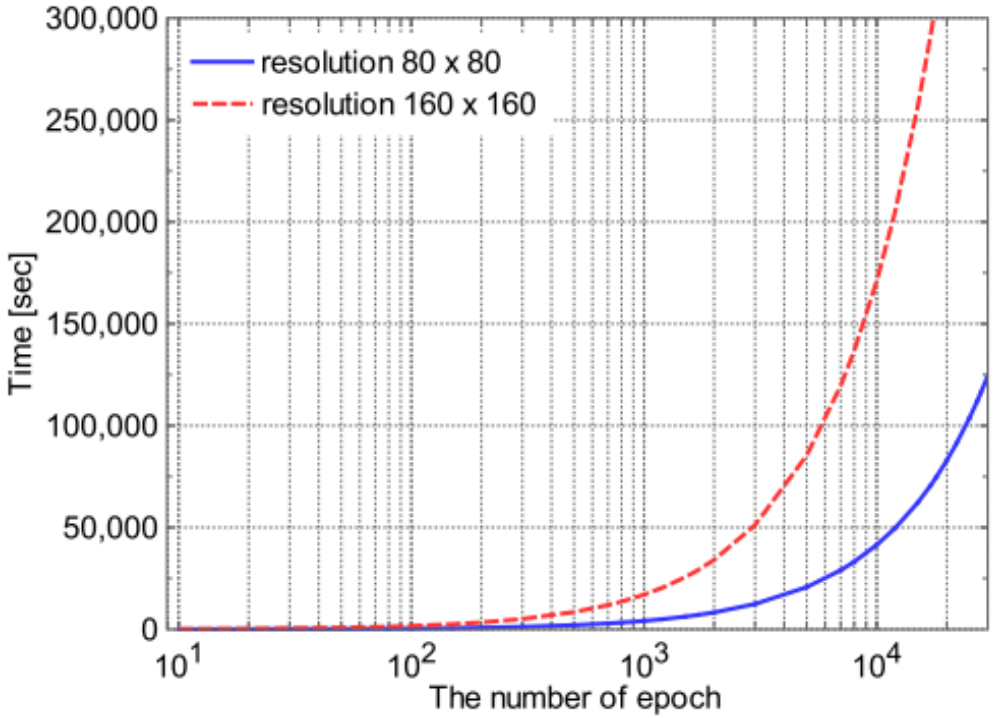

図 8: 学習回数と計算の関係

グレースケール画像化によるデータ削減効果

映像情報のデータ量をさらに削減するために, 画像をグレースケール化して提案システムを用いた場合の評価を行った.
入力画像群は 10000 枚, 無作為に 16 枚選択して学習に, 無作為に 8 枚選択したものをテストに用い. 学習回数は 10000 回, 損失関数は（2）を使用した.

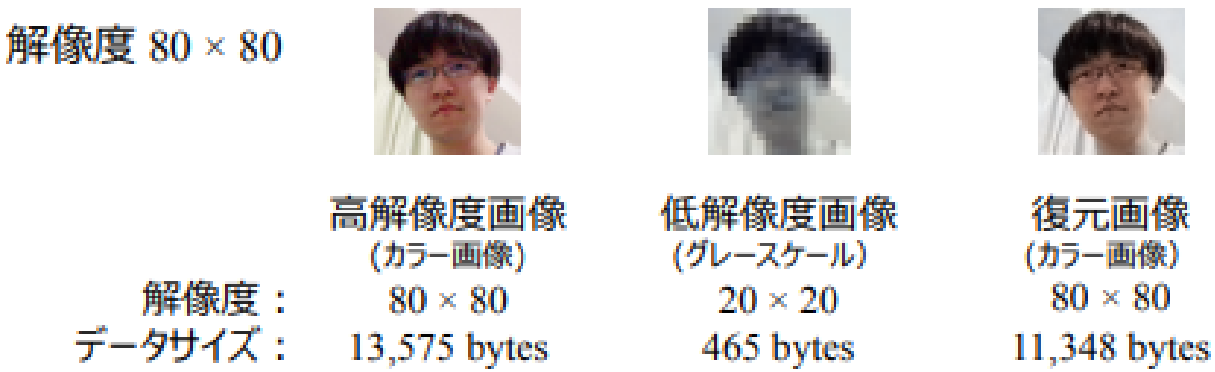

図 9: 評価に用いた画像とそのデータサイズ

まとめ

・ビデオ通話における映像トラフィックを削減する方式として, 深層学習による超解像技術を用いた手法を提案した.
・画像品質をある程度保ったままデータ量を大きく削減することが確認できた.