H-002

深層強化学習を利用した動画要約
Video summarization using deep reinforcement learning

神 琛麒† 張 善俊†
Chenqi Shen Shanjun zhang

1. はじめに

動画要約は、膨大な動画コンテンツを効率的に管理するための重要な技術です。本研究では、深層強化学習（Deep Reinforcement Learning, DRL）を利用して動画要約の品質を向上させる方法を探求します。この研究の目的は、モデルのフレーム評価と人間評価の相関係数を高めることです。DSN（Deep Summarization Network）をベースラインモデルとして使用し、他の 2 つのモデルと比較実験を行いました。評価指標として F1 スコアと相関スコア（Spearman および Kendall）を用い、要約の品質を評価しました。また、DSN ネットワーク内の報酬関数の調整を通じて、要約の品質を向上させる方法について検討しました。

2. モデル

DSN-DR モデルは、深層強化学習を利用して動画の重要部分を選択し、簡潔な要約動画を生成するために設計されています。DR（Diversity Re-weighting）は、要約ビデオを評価する多様性と代表性の要素であり、学習に使用されます。DSN は教師なし学習でも訓練可能です。

2.1 特徴抽出

特徴抽出の前処理として、以下の手順を行います：

- 色空間変換：BGR から RGB に変換
- リサイズ：256x256 ピクセルに調整
- センタクロップ：224x224 ピクセルにクロップ
- 正規化：画面データを正規化にする

2.1.1 フレーム抽出と特徴抽出

動画から 15 フレーム間隔でフレームを抽出し、GoogleNet モデルに入力して特徴データを取得します。それを Numpy 配列に変換し、flatten()メソッドを使用して 1 次元ベクトルに平坦化します。

2.2 ネットワーク

2.2.1 要約動画の生成

DSN の核心は双方向 LSTM（BiRNN）であり、入力された動画フレームシーケンスを処理します。LSTM ユニットは各フレームv_tを受け取り、選択確率p_tと隠れ状態h_tを出力します。

- **アクション A**: 各フレームが要約動画に選ばれるかどうかを示す二元変数$\{a_i\}$を含みます。
- **出力要約 S**: 選択されたフレーム $\{v_(y_i)\,|\,a_(y_i)=1, i=1,2,\ldots\}$を含みます。ここで、$y\_i$ は選択されたフレームのインデックスを示します。

2.2.2 報酬関数

報酬関数は、生成された要約の多様性と代表性を評価るために設計されています。具体的には、以下の 2 つの要素が含まれます。

- **多様性（Diversity）**：生成された要約の多様性を評価し、選択されたフレーム間の特徴空間での相違を測定します。計算方法は以下の通りです：

$$R_{div} = \frac{1}{|Y|(|Y|-1)} \sum_{t \in Y} \sum_{t' \in Y\ t \neq t'} d(x_t, x_{t'})$$

ここで、$d(x_t, x_{t'})$ は相違関数で、計算方法は次の通りです：

$$d(x_t, x_{t'}) = 1 - \frac{x_t^T x_{t'}}{||x_t||_2 ||x_{t'}||_2}$$

調整： 時間構造を無視しないために、時間的に離れたフレーム（距離がλ以上）に対して$d(x_t, x_{t'}) = 1$

- **代表性（Representativeness）**：生成された要約が元の動画をどれだけ代表できるかを評価します。代表性は k-メドイド問題として定義され、動画フレームとその最も近いメドイドとの間の平方誤差の平均を最小化することによって実現されます:

$$R_{rep} = \exp\left(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in Y} ||x_t - x_{t'}||_2^2\right)$$

ここで、T は元の動画のフレーム、Y は要約動画のフレーム。

計算された多様性と代表性を利用して R(S)でモデルを更新します。

$$R(S) = R_{div} + R_{rep}$$

2.3 実験

2.3.1 データセット

本実験では、SUMME データセットを利用します。SUMME データセットには、25 本の異なるタイプの動画が含まれており、これらの動画は休日、イベント、スポーツなどをカバーしています。これらの動画は、ほとんど編集されていない生のユーザビデオであり、編集された動画に比べて圧縮率が高くなっています。動画の長さは約 1 分から 6 分です

† 著者所属 神奈川大学

2.3.2 評価手法

ビデオ要約モデルの評価には、F1 スコアに加えて、Kendall および Spearman 相関係数を使用しました。これらの指標を用いて、モデルが生成した要約と人間が作成した要約の類似性、および動画のフレームについて評価したスコアと人間が評価したスコアの関連性を評価します。

2.3.3 実験設定

本実験は 2 つの部分に分かれています

比較モデル:

Random モデルは、動画フレームにランダムなスコアを付与します。

VASNet モデルは、注意メカニズムに基づいたニューラルネットワークモデルで、動画フレームのスコアリングに使用されます。

実験 1： 実験 1 では、SUMME データセットから GoogleNet を使用して特徴を抽出しました。次に、DSN モデル、Random モデル、VASNet モデルを用いて比較実験を行いました。具体的な流れは、モデルが動画フレームを評価し、KTS 法で分割されたビデオセグメントの各セグメントスコアを計算し、そこから要約としてのセグメントを抽出します。要約ビデオの長さは原動画の 15%に統一されます。各モデルの平均 F1 スコア、Kendall および Spearman の相関スコアを計算しました。

実験 2： 実験 2 では、DSN モデルを用いて、単一ビデオセグメントの長さを基に離散度を計算し、最長セグメントが総フレーム数の割合を占める割合を制御することで、多様性と代表性の比重を調整しました。評価には Kendall 相関係数を使用しました。

2.5 結果と考察

まず、実験 1 の結果は図 1 に示されている通りです(DSNsup は教師ありの場合)。同一の分割手法を使用した場合、ランダムにフレームにスコアを付与したランダムモデルであっても、F1 スコアの評価で比較的良好な結果を得ることができました。この点は、[2]Mayu Otani ら（2019）の論文に示されている結果と一致しています。F1 スコアの評価は分割手法に関係が深いです。

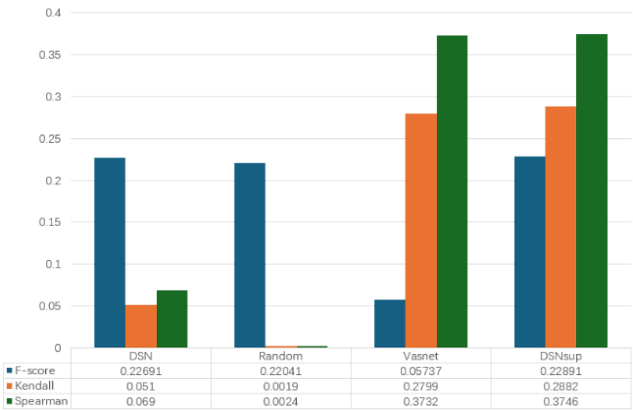

図 1 実験 1 の結果

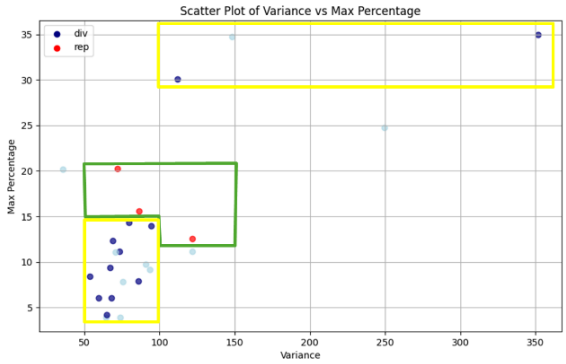

図 2　summe ビデオ離散と最長セグメント割合図

次に、実験 2 では Kendall 相関係数の向上を目指し、最長セグメントが総フレーム数に占める割合に基づいて多様性と代表性の比重を調整しました。図 2 に示すように、SUMME データセットの動画において、青色は多様性スコアのみを使用した DSN が Kendall 係数で高得点を得た動画を示し、赤色は代表性スコアのみを使用した DSN が高得点を得た動画を示しています。これに基づき、多様性が重要な場合は報酬の比重を 0.8(div)対 0.2(rep)に、代表性が重要な場合は 0.4(div)対 0.6(rep)に、それ以外の場合は 0.5(div)対 0.5(rep)に設定しました。(元の DSN モデルは R(S)=0.5(div)0.5(rep))

図 3 に示すように、比重を調整した後のモデル(実験 DSN)は、20 回の平均相関係数スコアにおいて、元の DSN モデルに比べてわずかに向上しました。

	DSN	実験 DSN	DSNsup	実験 DSNsup
平均 Kendall	0.027	0.029	0.258	0.26
平均 Spearman	0.035	0.039	0.335	0.337

図 3 実験 2 の結果

3. おわりに

最終的な結果として、離散度と最長セグメントの割合を調整することで、無教師学習の DSN および有教師学習の DSNsup のいずれも Kendall 相関係数のスコアが向上しました。今後の研究方向としては、報酬関数に注目度スコアを追加し、要約画面の色の豊かさやコントラストなどの面で注目度を評価することが求められます。

参考文献

[1] Kaiyang Zhou, Yu Qiao, Tao Xiang, "Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward", The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018.

[2] 2, Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, "Rethinking the Evaluation of Video Summaries", arXiv:1903.11328 [cs.CV], 2019.