複数ソースの転移学習を用いた TVCM 評価予測

熊谷　雄介†　　田原　将志†　　黄　　橙白†　　道本　　龍†

† 株式会社博報堂 研究開発局　〒 107–6332 東京都港区赤坂 5-3-1 赤坂 Biz タワー
E-mail: †{yusuke.kumagae,masayuki.tahara,chengbai.huang,ryo.domoto}@hakuhodo.co.jp

あらまし　広告は商品やサービス，ブランド，企業の認知や好意，購入意欲を高めるために消費者に向けて発信されるものである．特にテレビコマーシャル (TVCM) の放映は企業の広告費において大きな割合を占めているため，より効果的な TVCM を制作する必要がある．本論文では，TVCM を入力とし，それを視聴した人物が「その商品を購入したいと思ったか」「そのブランドへの好意が高まったか」といった，どのような印象を受けるかを推定するタスクに取り組む．本論文では異なる学習済みリソースを組み合わせ，転移学習を行うことで相関係数 $R^2 = 0.67$ での予測を実現した．

キーワード　広告, TVCM, 深層学習, 評価予測, 転移学習

1　イントロダクション

広告は商品やサービス，ブランド，企業の認知や好意，購入意欲を高めるために消費者に向けて発信されるものである．2017 年度の日本における総広告費は 6 兆 3,907 億円であり，GDP の 1.17%を占める巨大領域である．特に地上波・衛星放送を関わらず放送されるテレビコマーシャル (TVCM) にまつわる広告費は 2017 年度では 1 兆 9,657 億円であり，総広告費の 30.4 % を占めている．この高い広告費割合を支えているのは TV の 87.7 % という高いリーチ率 [1] であり，これは他のマスメディア（新聞 27 %，雑誌 8.2 %，ラジオ 18.5 %）と比べると圧倒的である [1]．そのため，一年で 10 万本以上の TVCM が日々放映されている．よって，広告主や広告代理店としては，数多くの TVCM の中から少しでも多くの人々の記憶や印象に残るものを制作・放映することが重要である．

本論文では，より効果的な TVCM の制作を実現すること，および TVCM の印象評価調査の効率化を目的とし，音声と映像から構成される TVCM と付随する情報とを入力とし，それを視聴した人物が「その商品を購入したいと思ったか」「そのブランドへの好意が高まったか」といった，どのような印象を受けるかを推定するタスクに取り組む．

大規模な調査パネルに対して行った TVCM に対する印象調査データと実際の TVCM 動画を用いた予測実験の結果，本論文で提案した推定手法によって相関係数 $R^2 = 0.67$ での予測を実現した．

本論文で取り組むのは TVCM に対する視聴者の評価予測であるが，本研究の最終目標は高精度な評価予測アルゴリズムを構築することにより，以下の作業を実現することである．

- 出演者や台詞，テロップ，BGM，場面構成をどのように変更すればより印象に残る TVCM となるかを示唆すること
- インターネットにおける動画広告のように，本来の TVCM より短い動画長の広告素材作りにおいて，どのように編集すれば最も効果が高い動画広告になるかの示唆を与えること
- 印象に残る TVCM 製作のための絵コンテ自動生成

論文の構成を以下に述べる．2 章では関連研究について二つの側面から説明を行う．3 章では本論文にて用いる分析手法について，主にいかに特徴量を抽出するかについて説明する．4 章では評価予測実験の設定および結果をまとめ，5 章では実験において試みたものの有効に機能しなかった取り組みについて説明し，6 章で本論文をまとめる．

2　関 連 研 究

ここでは大きく分けて動画像に対する評価予測と広告にて用いられる動画像素材に対する分析の二つについて振り返る．

まずは動画像に対する評価予測について関連研究を振り返る．DWANGO [2] はユーザ投稿型のサービスにおいて投稿されたイラストについてその閲覧数および「お気に入り」の数を GoogLeNet を用いて予測している．Shu ら [3] および Esfandarani ら [4] は与えられた画像が「本来の画像からどの程度劣化したか」といった (劣化前の元画像が存在すれば) 定量化可能な指標，および「美しいか」や「独創的か」といったような定量化不可能な指標について人手での評価を学習・予測するタスクに取り組んでいる．Karpathy [5] は画像投稿サービスにおいて自身の顔写真を撮影したものを収集・蓄積し，どのような写真であれば高い評価がつくのかを学習・推定している．また，その中で入力された写真をどのようにトリミングすれば (学習されたモデルにおいて) 評価が改善するかの検証を行っている [2]．

より本論文に関係するテーマとして，広告において用いられる動画像に対する分析がある．Hussain ら [6] はアルゴリズム

1：全人口における当該メディアによって到達可能な人口の割合

2：たとえば，画像中において顔領域が占める割合が高いほど評価が高くなることなどが示されている．

動画 ID	テキスト情報
1000001	売上ナンバーワン，ビールなら XXX
1000002	新発売

表 1　テキスト情報の例.

動画 ID	出演者情報
1000001	織田信長, 明智光秀
1000002	徳川家康

表 2　出演者情報の例.

による分析に向けた広告の動画像データの収集とその基礎集計を行っている．Junxuan ら [7] はバナー画像を用いてインターネット広告における Click Through Rate の予測を行っている．Samaneh ら [8] は音楽配信サービスにて採用されている音声のみの広告を対象に，Long Click Rate[3] を代替指標として音声広告の品質推定に取り組んでいる．河原 [9] は TVCM について人手で付与した表現要素とその TVCM を視聴した者がどのような印象を受けるかを Random Forest を用いて分析を行っている．中村ら [10] は TVCM の動画を対象として視聴した者が受ける印象を深層学習を用いて学習・推定している．この取り組みは本論文に最も近いものであるが，本論文では動画だけではなくその音声や出演者情報，テキスト文などを用いている点が異なっている．また，Anjan ら [11] は商品検索における Click Through Rate と画像特徴量との関係について分析している．

3 提 案 手 法

ここからは，本論文で取り組むタスク，そのタスクで用いる特徴量の抽出方法，および予測手法について述べる．

3.1 タ ス ク

本タスクは以下に定義する **CM 動画情報**を受け取り，**評価値**を出力し，未知の CM 動画情報への評価値を予測する関数を学習することである．

3.1.1 入力：CM 動画情報

本タスクで用いる CM 動画情報は次の情報で構成されている．

動画　動画はサイズ 1440x1080 の画像 (フレーム) の集合で構成されている．動画のフレームレートは 29.0 fps であり，これは動画 1 秒が 29 枚のフレームで構成されていることを意味している．各動画の長さは 15 秒，30 秒，60 秒と異なるものが混在している．今回は処理の簡単のため，1 秒あたり 8 枚のフレームをランダムにサンプリングしたものを用いる．また，動画には同じ長さの音声も含まれているが，今回は音声と画像の集合は別の特徴量として取り扱う．

音声　音声はビットレート 64 kbps，サンプリングレート 48 kHz，長さ 15 秒， 30 秒， 60 秒の mp3 ファイルである．今回は BGM とナレーションとの分離などは行わず， CM 動画に含まれる音声全てを用いた．

テキスト　テキストはアノテーターによって付与された，動画のナレーションおよびテロップを自然言語で記述したものである．テキストの平均文字数は 71.96 字である．表 1 にテキストの例を記す．このようにテキストは網羅性が低く，すべての事象を記録できていないことがほとんどである．

出演者情報　出演者情報は自然言語で記述された，どの俳優 (名寄せ済み) が出演しているかの集合を表現した情報である．出演者情報の例を表 2 に記す．

カテゴリ情報　カテゴリ情報はその CM が取り扱っている商品のカテゴリを示したものであり，31 種の大カテゴリ，104 種の中カテゴリ，224 種の小カテゴリで構成されている．

3.1.2 出力：評価値

TVCM の評価値は，博報堂の広告好感度調査 BestHIT を用いる．BestHIT は東京キー局で放送された全ての TVCM について，商品・サービス・ブランドなどの様々なテーマの表現効果を測定するための調査である．調査は 2007 年から行われており，のべ 10 万人以上の規模の調査パネルに対して TVCM に関する様々な質問を行っている．

今回分析対象とするのは「表現に好感を持ったか」という質問に対して,「とてもそう思う」から「全くそう思わない」の 5 段階の質問を行い，最大を 2 点，最小を-2 点とした時の加重平均の値である． すなわち，目的変数 $y \in [-2, +2]$ である．また，分析においては一定数以上の調査パネルが答えた TVCM のみを予測の対象として採用した．

3.2 特徴量抽出

本論文では， CM 動画情報に含まれるそれぞれの情報から特徴量を抽出し，Early Fusion [12] を行うことで学習器に与える特徴量とする．ここからは，それぞれの情報からどのように特徴量を抽出するかについて説明する．

詳しくは後述するが，今回分析対象とするデータは次元数の大きさに対して数が少なく，データのみで学習を行うことが困難である．よって，特徴量抽出においてはさまざまな外部リソースや学習済みモデルを活用し，転移学習を行うことで精度改善を図る．抽出の大まかな方針は Lee らの取り組み [13] に基づいている．図 1 は特徴量抽出の概要を描いたものである．

3.2.1 動画からの特徴量抽出

動画からの特徴量抽出は

- a) キーフレームの抽出
- b) キーフレームからの特徴量抽出
- c) 複数のキーフレームから得られた特徴量の統合

の 3 ステップを行う．

3：一定の時間より長くランディングページを閲覧していたもののみを対象にした Click Through Rate

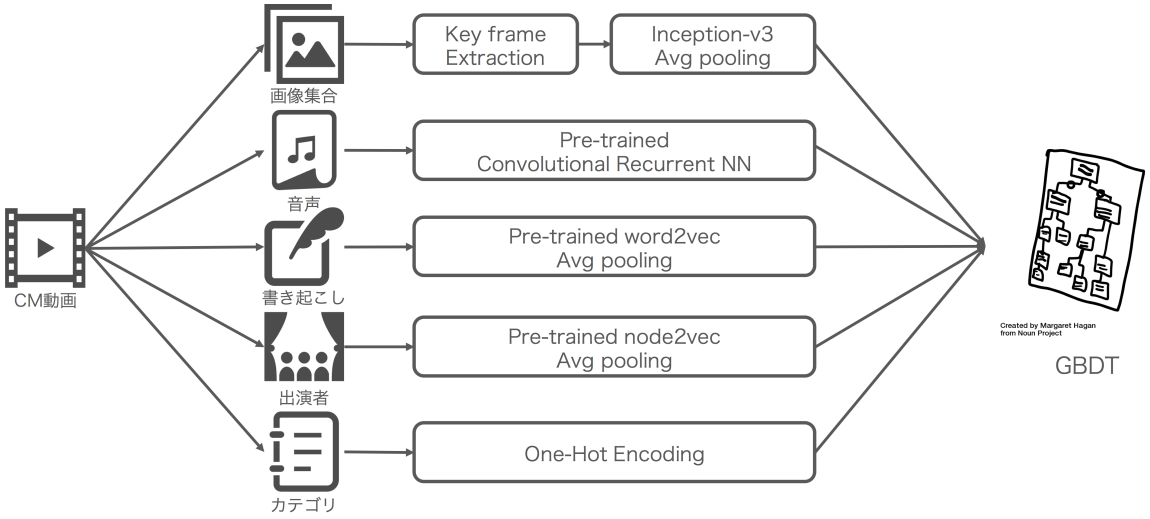

図 1　特徴量抽出の概要図．CM 動画を構成するそれぞれの情報から個別に特徴量抽出を行い，最終的に連結して入力ベクトルを構築し，学習器に与える．

a）　キーフレームの抽出

動画は前述したように複数枚のフレームの集合であるため，動画を代表する特徴的なフレーム (キーフレーム) を抽出する．今回は Zhuang ら [14] の手法にもとづきクラスタリングを用いてキーフレームを抽出する．まずはそれぞれのフレームを HSV 色空間で表現し，カラーヒストグラムとする．その後，クラスタ数 15 の k-means クラスタリングを行い，それぞれのクラスタの中心点に最も近い 15 枚のフレームをキーフレームとして得る．

b）　キーフレームからの特徴量抽出

続いてキーフレームから特徴量を抽出する．抽出には ImageNet [15] で学習済みの Inception-v3 [16] を採用する．Inception-v3 にキーフレームを入力として与え，出力層の前の層である最終層の出力である 2048 次元の表現をキーフレームの特徴量として用いる．

c）　複数のキーフレームから得られた特徴量の統合

ここまでの結果として 2048 次元のベクトルが 15 本存在していることになる．最後にこれらのベクトルをまとめて 2048 次元の 1 本のベクトルに変換する．予備実験の結果，変換には average pooling を用いる．

3.2.2　音声からの特徴量抽出

音声からの特徴量抽出には学習済みの CRNN [17] を用いる．CRNN は Convolutional Neural Network と Recurrent Neural Network を組み合わせたモデル [18] であり，畳み込みによる周辺情報と系列性との両者を考慮することにより精度を改善したモデルである．

今回用いた学習済みモデル [4] は，入力として音声の時間とメル周波数スペクトログラムを取り，その音声のジャンルを推定するように学習されたモデルである．今回はモデルに CM 音声を入力し，最終層にて得られる 32 次元のベクトルを音声の特徴量として用いる．

4：https://github.com/keunwoochoi/music-auto_tagging-keras

3.2.3　テキストからの特徴量抽出

テキストからの特徴量抽出は

- a) 形態素解析による品詞集合への変換
- b) 各品詞の埋め込みベクトルへの変換
- c) 埋め込みベクトルの集合の統合

という 3 ステップの操作を行う．まず追加の辞書として neologd [19] を使用した MeCab [20] を用いて形態素解析を行う．次にその結果得られた品詞集合それぞれに対し，学習済み日本語 Wikipedia エンティティベクトル [21] を参照し，200 次元の単語埋め込みベクトルを得る．最後に，単語埋め込みベクトル集合に対して average pooling を行い， 200 次元の 1 つのベクトルを得，これを特徴量とする．

3.2.4　出演者情報からの特徴量抽出

出演者情報の特徴量抽出においては，学習済み日本語 Wikipedia エンティティベクトルを用いるのではなく，「誰が誰と共演したか」についてより着目して処理を行う [5]．

より共演関係に敏感な特徴量抽出のために，日本語版 Wikipedia におけるページ間のリンク情報に着目する．Wikipedia の映画やドラマなどの作品ページには，多くの場合出演者の情報が箇条書き，相互にリンクされている．よって，リンク情報に着目することにより，本文情報よりも共演関係をより反映した情報が抽出できると考えられる．今回は日本語版 Wikipedia におけるそれぞれのページをノード，ページ間のリンクをエッジとして扱うことで巨大な重み無し有向グラフを構築し，node2vec [22] を用いたグラフ埋め込みを学習することで，各ページごとの潜在表現を得る [6]．その後，各出演者

5：学習済み日本語 Wikipedia エンティティベクトルは Wikipedia 本文に対して word2vec を適用している．共演関係を直接学習するには共演者の名前が同一文中に登場している必要があるが， Wikipedia において出演者名は箇条書きされていることが多いため，十分に学習できていないと考えられる．

6：この時，潜在表現の次元は 128 次元として学習を行った．

の名前で参照することで潜在表現の集合を得，最後に average pooling を行い 128 次元の 1 つのベクトルを構築し，これを特徴量とする．

3.2.5 カテゴリ情報からの特徴抽出

前述したようにカテゴリ 31 種の大カテゴリ，104 種の中カテゴリ，224 種の少カテゴリで構成されているため，それぞれを one-hot encoding ベクトルとして扱う．すなわち，カテゴリ情報にもとづく特徴量は $31 + 104 + 224 = 359$ 次元の，最大でも 3 つの要素が 1 になったベクトルである．

3.3 予測手法

それぞれの特徴量を入力とし，評価を学習・推定する学習器には Gradient Boosting Decision Tree(GBDT) [23] を用いる．複数ある GBDT の実装の中でも今回は LightGBM [24] を用いる．

4 実　　験

4.1 実験設定

データセット

今回予測対象とするのは 2017 年度に日本国内において放映された TVCM 3343 本である．これを 5 分割し， 5-fold cross validation の平均値によって予測精度を検証する．また，今回取り扱うデータセットでは商品ごとの TVCM の本数に偏りがある．よって，企業ごとに最大 20 本までの TVCM を分析対象とし，なおかつ，学習時に leak を引き起こさないよう [7]，厳密なランダムではなく，企業およびカテゴリが合致する TVCM がある分割時において学習データと予測データとに同時に存在しないように分割を行った．

評価指標

TVCM i に対する評価の予測値を $\hat{y}_i$ ，真の値を y_i とする時，N 本の TVCM に対する予測の評価は Root Mean Squared Error (RMSE)

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1}$$

と相関係数 R^2

$$1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \mu_y)^2} \tag{2}$$

を用いる [8]．RMSE は実際の値と予測値との差の二乗値の平均値の平方根であり，小さければ小さいほど予測が正確であることを意味する．また， R^2 は -1 から 1 を取る値であり，大きければ大きいほど予測が正確であることを意味する．また，$R^2 = 0.0$ の場合は予測が全くのランダムであることを意味する．

特徴量	RMSE	R^2
カテゴリのみ	0.27	0.16
テキストのみ	0.19	0.57
出演者のみ	0.24	0.36
音声のみ	0.22	0.42
動画のみ	0.19	0.57
全特徴量	**0.17**	**0.67**
全特徴量 (NN)	0.22	0.44

表 3　実験 1 の結果である異なる特徴量を用いて学習させた場合の評価指標の違いを示した表．RMSE は小さければ小さいほど，R^2 は大きければ大きいほどモデルが優れていることを意味する．もっとも優れた RMSE および R^2 は太字で表記した．

ここでふたつの評価指標を用いる理由は， RMSE はそれぞれの評価をどれだけ正確に予測できたかを確認するため，R^2 は評価が高いものは高い値を，低いものには低い値が予測できているかを確認するためである．実用を考えた場合には，個々の CM 動画に対する評価の正確さではなく，何が優れ，何が劣っているのかの傾向が予測できれば十分であるため， R^2 に着目して結果を検証することが適切であると考えられる．

実験設定

本論文では以下に示す二種類の実験を行う．

実験 1：全調査パネルに対する予測　全てのパネルを対象に計算した評価値の予測がどの程度の精度で実現可能か，また，テキストや画像，音声といったそれぞれの特徴量がどのように，どの程度予測に貢献するかを検証する．

実験 2：集計区分ごとの予測　TVCM や TV の視聴率は多くの場合，F (女性) および M (男性) と呼ばれる性別の区分と 1 (20 歳から 34 歳)，2 (35 歳から 49 歳) および 3(50 歳以上) と呼ばれる年齢の区分の組み合わせの計 6 パターン (M1 や F3 など) の集計区分にもとづき計算やターゲティングが行われる．これらの区分はビジネス上重要であるため，調査パネルをこの区分に分割し，区分ごとに求めた評価値の予測が実現可能かどうかを検証する．

4.2 実験 1：全調査パネルに対する予測

実験 1 の結果を表 3 に示す．もっとも優れた RMSE および R^2 は太字で表記した．実験 1 ではそれぞれの特徴量ごとにモデルを構築し，どの特徴量が予測に対して効果的に働くかを確認した．表 3 に示したように，単一種の特徴量で RMSE および R^2 の精度が高かったもの，すなわち支配的だったのは動画情報である．つぎにテキスト，音声が効果的だった．また，全ての特徴量を用いた予測精度は $R^2 = 0.67$ であった [9]．

また，参考までに学習器を GBDT ではなく，活性化関数に ReLU を用い Adam によって学習したユニット数 100 の 3 層ニューラルネットワークを用いて予測したもの (NN) を比較として追記した．同じ特徴量を使っているにも関わらず， GBDT

7：なぜなら同一商品に対する評価は類似するため

8：μ_y は y の平均である．

9：この値は (データセットが異なるため直接の比較は困難であるが) 中村ら [10] によって報告されている値よりも高かったことを付記しておく．

区分	M1	M2	M3	F1	F2	F3
RMSE	0.44	0.42	0.43	0.40	**0.38**	0.41
R^2	0.24	0.29	0.31	0.38	**0.42**	0.35

表 4　実験 2 の結果である異なる区分に対する予測精度の差を示したもの．RMSE は小さければ小さいほど， R^2 は大きければ大きいほどモデルが優れていることを意味する．もっとも優れた RMSE および R^2 は太字で表記した．

を用いた場合より RMSE および R^2 ともに予測精度が著しく悪化していることがわかる．

4.3　実験 2：集計区分ごとの予測

実験 2 では 6 つの区分ごとに評価値を計算した上で学習・予測を行った．この時，本来の 2048+32+200+128+359 = 2767 次元の特徴量に加え，区分を示す 6 次元の one-hot encoding 特徴量を加え，同時に cross validation における各 fold の学習データを 6 倍に増やすことで学習および予測を行った．

表 4 はその結果である．もっとも優れた RMSE および R^2 は太字で表記した．どの区分においても RMSE および R^2 も実験 1 より低下していることがわかる．これは区分ごとに抽出された調査パネルの数が少なくなってしまい [10]，評価値そのもののばらつきが相対的に大きくなってしまい，予測が全体のそれより困難になったためと考えられる．

区分ごとに評価指標の比較を行うと，最も予測がしやすいのは F3 (RMSE=0.38，R^2=0.42) であり，反対に最も困難なのは M1 (RMSE=0.44 ，R^2=0.24) であることがわかる．

5　考　　察

ここでは，本論文において試したものの精度の改善が見られなかった取り組みとその理由について考察する．

5.1　ニューラルネットワークによる学習

一般的には特徴量を個別に抽出するのではなく，ひとつの巨大なニューラルネットワークを構築し，特徴量抽出から評価予測を行う方が改善すると言われている [25]．これは，段階的に推定や抽出を行うことで，誤差が累乗的に蓄積するためであると言われている．

しかし，今回の推定においては End-to-End なネットワーク以前に特徴量抽出後の学習器をニューラルネットワークに置き換えたものですら精度が悪化してしまった (表 3)．これはおそらく学習データが 3,000 件程度と少数だったためと考えられる．そのため，対策としてより多くの TVCM を用いた学習が必要不可欠であると考えられる．

また，ニューラルネットワークの学習においてはデータの水増しが精度に改善することが知られている．今回はキーフレームに対して Random Erasing [26] による Data augmentation を行ったが予測精度が若干悪化するのみに留まった．

5.2　三次元方向の畳み込み

3D ResNet [27] に代表されるような三次元方向の畳み込み構造を持つモデルを用いて学習および推定を行ったが，三次元方向の畳み込みを用いないモデルよりも予測精度が悪化した．この原因は二つ考えられる．

ひとつは三次元方向の畳み込みを持つモデルの事前学習が十分ではなかったという可能性である．この課題は Youtube-8M [28] などの大規模な動画像データセットを用いた学習 [11] を行うことで改善されるのではないかと考えられる．

もうひとつは「そもそも TVCM に三次元方向の構造があるのか」という，より根源的な問題である．動画を対象にした識別問題では，UCF-101 [29] や HMBD-51 [30]，ActivityNet [31]，Kinetics [32] に代表されるように人間の動作を推定することが一般的である．そのような動画においては映像は激しく切り替わることはなく，連続した場面がフレームに記録されていることが多い．一方， TVCM は視聴者の記憶に残る必要があるため，必然的にある時は出演者，ある時は商品，ある時は企業名やテロップ，といったようにめまぐるしく場面が切り替わるものが多い．そのため，三次元方向の構造が存在していない (正確には局所的にのみ存在している) とも考えられる．

本論文では動画を時間軸上で連続したフレーム集合としては扱わず，キーフレームを抽出することでその動画としての構造を無視して分析を行った．動画をより動画らしく扱うために，どのように三次元方向の関係を扱うかは今後の課題である．

6　結　　論

本論文では， TVCM 動画を入力とし，視聴者がその広告に対して抱く印象の学習・推定について取り組んだ．複数の特徴量を用いたモデルによる実験の結果，相関係数 R^2 において 0.67 での予測を実現した．

今後は，よりデータを増やしての推定を検討している．その際，時間変化に伴う出演者や表現の流行の変化などを把握することを目標としている．また，制作側に示唆を与えるためにも，どの場面や音声，テキストが印象に働きかけているのかを知るために注意機構 [33] の導入などを検討している．同時に，動画特徴量と音声特徴量を分離すること無く効果的に扱う方法についても検討を進めたい．

文　　献

[1] 博報堂 DY メディアパートナーズ. **広告ビジネスに関わる人のメディアガイド** *2018*. 宣伝会議, 2018.

[2] DWANGO MEDIA VILLAGE. イラストの閲覧数・お気に入り数予測. Technical report, 2015.

[3] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.

[4] Hossein Talebi Esfandarani and Peyman Milanfar. NIMA: neural image assessment. *CoRR*, abs/1709.05424, 2017.

[5] Andrej Karpathy. What a deep neural network thinks about

10：何名が各区分に属していたのかは公開できない．

11：Youtube8M の場合には動画を入力とした総閲覧数の予測問題など

your # selfie. Technical report, 2015.

[6] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1100–1110. IEEE Computer Society, 2017.

[7] Junxuan Chen, Baigui Sun, Hao Li, Hongtao Lu, and Xian-Sheng Hua. Deep ctr prediction in display advertising. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 811–820, New York, NY, USA, 2016. ACM.

[8] Samaneh Ebrahimi, Hossein Vahabi, Matthew Prockup, and Oriol Nieto. Predicting audio advertisement quality. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 153–161, New York, NY, USA, 2018. ACM.

[9] 河原 達也. Tvcm 表現要素の消費者反応に対する効果. **行動計量学**, 43(1):85–105, 2016.

[10] 中村 遵介, 河原達也, and 山崎俊彦. 畳み込みニューラルネットによるテレビ広告動画における魅力予測. In **信学技報**, volume 117, pages 21–24, 2017.

[11] Anjan Goswami, Naren Chittar, and Chung H. Sung. A study on the impact of product images on user clicks for online shopping. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 45–46, New York, NY, USA, 2011. ACM.

[12] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 399–402, New York, NY, USA, 2005. ACM.

[13] Joonseok Lee, Sami Abu-El-Haija, Balakrishnan Varadarajan, and Apostol (Paul) Natsev. Collaborative deep metric learning for video understanding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 481–490, New York, NY, USA, 2018. ACM.

[14] Yueting Zhuang, Yong Rui, T.S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *Proceedings 1998 International Conference on Image Processing (ICIP)*, 1998.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[17] Keunwoo Choi, George Fazekas, Mark B. Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396, 2017.

[18] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 580–587, Washington, DC, USA, 2014. IEEE Computer Society.

[19] Taiichi Hashimoto Toshinori Sato and Manabu Okumura. Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese). In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pages NLP2017–B6–1. The Association for Natural Language Processing, 2017.

[20] Taku Kudo. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*, 2005.

[21] 鈴木 正敏, 松田 耕史, 関根 聡, 岡崎 直観, and 乾 健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. In *The Association for Natural Language Processing*, 2016.

[22] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA, 2016. ACM.

[23] Jerome H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, February 2002.

[24] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017.

[25] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc., 2015.

[26] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017.

[27] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.

[28] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016.

[29] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[30] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[31] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 961–970, June 2015.

[32] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[33] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.