サムネイル画像への見出しレイアウト生成

成川 喬朗[1,a)]　出口 大輔[1]　川西 康友[2,1]　村瀬 洋[1]

概要

本稿では，サムネイル画像作成の簡単化を目的として，事前に用意した背景画像と見出し文字の 2 つを入力としてサムネイル画像中の見出しレイアウトを自動生成する手法を提案する．提案手法の見出しレイアウト生成には，Variational AutoEncoder（VAE）型のレイアウト生成モデルを使用し，背景画像から VAE の潜在空間への写像を求めることで背景画像から見出しレイアウトへの変換を実現する．学習はレイアウトからレイアウトを復元する第一段階と背景画像からレイアウトを復元する第二段階の二段階に分けて行なう．第一段階では潜在空間が適切なレイアウトを表現できるように学習を行なう．第二段階では背景画像から得た特徴量を第一段階で学習した潜在空間にマッピングするように学習する．このように潜在空間を経由することにより，背景画像からレイアウトへの適切な変換を実現する．実験では提案手法が背景画像中のどの領域に文字を配置しやすいかを調べることで，その有効性を評価した．

1. はじめに

インターネットコンテンツを作成する人々はクリエイターと呼ばれ，自身の作成物を多くの人々に見てもらえるように様々な工夫を行なっている．例えば，注目の得やすい単語をタイトルに含めたり，最近話題の内容を取り入れるなど様々なものがある．佐藤らの研究 [7] では，動画共有サービス YouTube で視聴者が見たい動画を決定する際は，タイトルの文章や単語よりもサムネイル画像を重視しているという結果が得られている．そのため YouTube のような動画投稿サイトでは人々の目を惹きつけるサムネイル画像の作成が特に重要であると考えられる．また，サムネイル画像は，動画共有サービスやブログなどの様々な場面でも使用されており，それらのコンテンツにおいても注目を集めるための重要な要素となっている．

図 1 に示すように，サムネイル画像はコンテンツと関連する画像にメッセージ性の高いフレーズを組み合わせて作成されるのが一般的である．しかし，サムネイル画像を

図 1　YouTube のサムネイルの例（youtube.com [1] より引用）

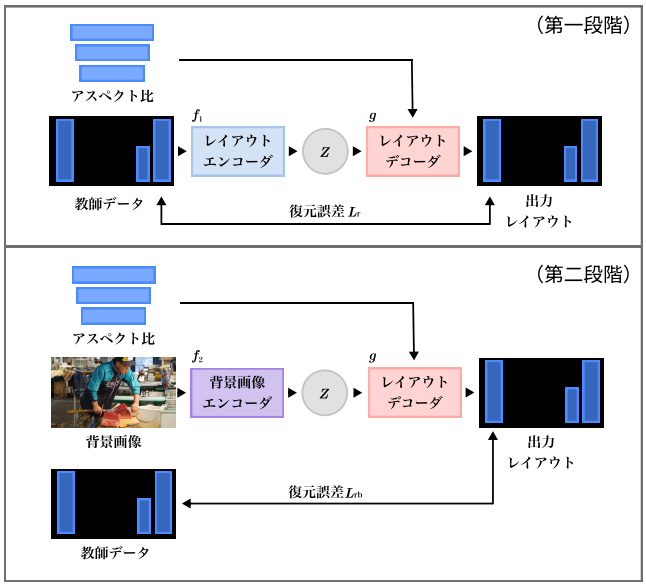

図 2　提案手法の学習概要図

作成するには，文字を配置する場所の全体的なバランスや背景に使用する画像（以下，背景画像と呼ぶ）の邪魔にならないように文字の大きさや配置を決める必要がある．Li らによる LayoutGAN [5] や Arroyo らによる Variational Transformer Networks [3] など，文書などのデータセットを用いたレイアウト自動生成についての研究は存在するが，このような背景画像を考慮して生成する手法は存在しない．そこで，本稿ではこの文字の大きさや配置を決定する部分の自動化を目的として，背景画像と配置する文字のアスペクト比を事前条件に与えてレイアウト生成を行なう手法を提案する．

2. 提案手法

本稿で扱うレイアウトのデータは，文字の配置領域の集合として表現される．以下，この文字の配置領域それぞれ

1　名古屋大学
2　理化学研究所 情報統合本部 GRP
a)　narikawat@vislab.is.i.nagoya-u.ac.jp

をエレメントと呼ぶ.

2.1 処理の概要

本稿では，背景画像の内容に合わせてサムネイル画像の文字レイアウトを自動生成する手法を提案する．提案手法では VAE と Transformer を組み合わせたモデルを使用しており，学習を図 2 に示すレイアウトからレイアウトを復元する第一段階と，背景画像からレイアウトを復元する第二段階の 2 つに分けて行なう．第一段階では適切なレイアウトを表現できるように潜在変数の学習を進める．第二段階では背景画像から得た特徴量を第一段階で学習した潜在空間にマッピングするよう学習する．このように，背景画像からレイアウトへ直接変換するのではなく，潜在空間を経由して処理を行なうことで，本来結びつけることが困難な背景画像からレイアウトへの変換を実現する．

また，提案手法の学習に関して，一般的にある背景画像に対して配置可能な見出し文字のレイアウトには曖昧性があり，必ずしも一意に定めることはできない．そのため，単に背景画像とレイアウトの組を学習するだけでは，背景画像とレイアウトの 1 対 1 の関係のみを学習してしまい，未知の背景画像に対応できないという問題が生じる．この問題の解決策として，背景画像とレイアウトそれぞれをランダムクロップして学習に用いることにより，データセットのバリエーションを増やすデータ拡張を行なう．これにより 1 対 1 の関係のみを学習させないようにする．

2.2 レイアウトの表現

レイアウトデータは複数のエレメントから構成される．各エレメント $\mathbf{e}_i$ にはクラス c_i, X 座標 x_i，Y 座標 y_i，縦幅 w_i，横幅 h_i，回転 r_i の 6 つのパラメータが含まれる．レイアウトデータには Transformer の特殊トークンも含まれるが，これはクラス c_i を用いて区別する．

$$\mathbf{e}_i = [c_i, x_i, y_i, w_i, h_i, r_i]$$

2.3 エレメントの大きさ決定

提案手法では事前にエレメントの文字数（アスペクト比）が与えられ，それに合わせた領域を出力する．そこで，文字配置領域のアスペクト比 a_i を計算し，モデルに入力する．ここで，a_i を単純に w_i/h_i としてしまうと，$w_i > h_i$ の場合は $1 \leq a_i < \infty$，$w_i < h_i$ の場合は $0 < a_i \leq 1$ の値をとることになり，a_i の表現力に偏りが生じる．そこで，a_i を式 (1) のように定義する．モデルにはエレメントの縦幅もしくは横幅を決定する基準となる大きさ s_i と，$w > h$ であるかを表す $c_{w>h} = \{0, 1\}$ を出力させ，式 (2), (3) に示すように a_i と組み合わせて w_i, h_i を決定する．

$$a_i = \frac{\min(w_i, h_i)}{\max(w_i, h_i)} \quad (0 < a_i \leq 1) \tag{1}$$

$$w_i = \begin{cases} s_i & (c_{w>h} = 1) \\ a_i s_i & (c_{w>h} = 0) \end{cases} \tag{2}$$

$$h_i = \begin{cases} a_i s_i & (c_{w>h} = 1) \\ s_i & (c_{w>h} = 0) \end{cases} \tag{3}$$

2.4 潜在空間を介したレイアウト生成の学習

第一段階では，学習データに含まれるレイアウトのルールや傾向などを捉えるために，アスペクト比のみを条件としたレイアウト生成モデルの学習を行なう．学習は図 2 の第一段階に示すように VAE の枠組みに基づいて行なう．学習においては，復元したレイアウトと教師データのレイアウトとの差分 L_{r} を最小化するように進める．ここで，L_{r} は 式 (4) ∼ (8) のように計算される．また，f_1 はレイアウトエンコーダ，g はレイアウトデコーダであり，教師データは $\hat{u}$ のように表記する．これにより，潜在変数の値を変化させることで様々なレイアウトの生成を実現する．

$$\{\mu, \sigma\} = f_1(\{\hat{\mathbf{e}}_1, ..., \hat{\mathbf{e}}_n\}, \{a_1, ..., a_n\}) \tag{4}$$

$$\{\mathbf{e}_1, ..., \mathbf{e}_n\} = g(\mathbf{z}), \quad \mathbf{z} \sim \mathcal{N}(\mu, \sigma^2) \tag{5}$$

$$L_{\mathrm{r}} = \sum_i L(\hat{\mathbf{e}}_i, \mathbf{e}_i) \tag{6}$$

ただし，$\mathcal{N}(\mu, \sigma^2)$ は平均 μ，分散 σ^2 の正規分布を表す．そして， L は以下の式で与えられる．

$$\begin{aligned} L(\hat{\mathbf{e}}_i, \mathbf{e}_i) = {} & L_{\mathrm{c}}(h(\hat{c}_i), h(c_i)) + L_{\mathrm{c}}(h(\hat{x}_i), h(x_i)) \\ & + L_{\mathrm{c}}(h(\hat{y}_i), h(y_i)) + L_{\mathrm{c}}(h(\hat{s}_i), h(s_i)) \\ & + L_{\mathrm{c}}(h(\hat{r}_i), h(r_i)) \end{aligned} \tag{7}$$

$$L_{\mathrm{c}}(p, q) = -\sum_x p(x) \log q(x) \tag{8}$$

第二段階では，第一段階で作成した生成モデルを拡張して，図 2 中の第二段階に示すように背景画像を条件としてレイアウト生成の学習を行なう．背景画像の画像特徴量を用いて潜在変数を求め，それを用いてレイアウトを出力する．そして，出力したレイアウトと背景画像のペアになっている教師データのレイアウトの差分 L_{rb} を最小化するように学習を進める．ここで，L_{rb} は 式 (10) ∼ (11) のように計算される．また，f_2 は背景画像エンコーダである．これにより，背景画像に対応するレイアウトの生成を行なう．

$$\{\mu, \sigma\} = f_2(I, \{a_1, ..., a_n\}) \tag{9}$$

$$\{\mathbf{e}_1, ..., \mathbf{e}_n\} = g(\mathbf{z}), \quad \mathbf{z} \sim \mathcal{N}(\mu, \sigma^2) \tag{10}$$

$$L_{\mathrm{rb}} = \sum_i L(\hat{\mathbf{e}}_i, \mathbf{e}_i) \tag{11}$$

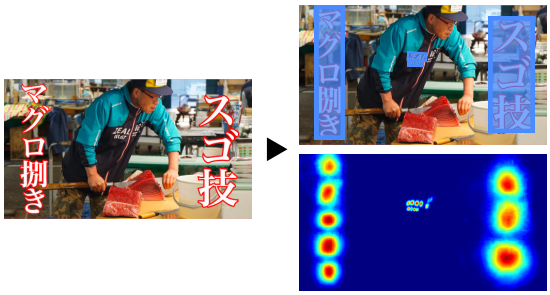

図 3 データセットの文字領域を検出した例

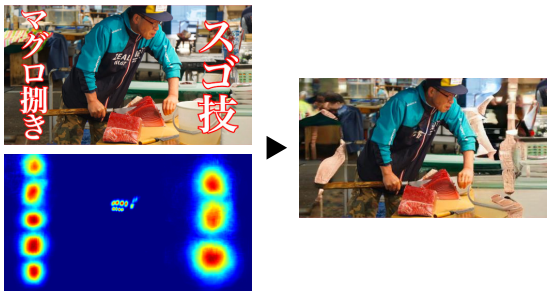

図 4 データセットに画像補完を適応した例

2.5 提案手法による文字レイアウト生成

提案手法によりレイアウトを生成する際には，まず事前に使用する背景画像と埋め込む文字列 $t_1, ..., t_n$ を用意する．そして，各文字列 t_i に対応する a_i を文字列の文字数 N_i から算出する．さらに，背景画像からエンコーダを通じて潜在変数 $\mathbf{z}$ を出力し，デコーダを用いて潜在変数からレイアウトを生成する．

3. データセットの構築

提案手法では，サムネイル画像の文字レイアウトと背景画像の組からなるデータセットを用いる．しかし，このような公開データセットは存在しないため，WEB 上で公開されているサムネイル画像から背景画像と文字レイアウトの組を抽出してデータセットの構築を行なった．

サムネイル画像を採用している動画サイトやブログは多く存在するが，それらの中でも YouTube [1] はバリエーション豊富で掲載数も多く，かつ API [2] によって収集が容易である．そこで本実験では，YouTube からサムネイル画像を収集して利用した．また，再生回数の多い動画のサムネイル画像は優れているという仮定のもと，再生回数によるフィルタリングを行ない，学習用として 12,350 枚，検証用として 1,259 枚を用意した．

まず，サムネイル画像から文字レイアウトとして文字領域の検出を行なう．文字領域の検出には Beak らが提案している文字領域検出モデル CRAFT [4] を使用した．本稿では， CRAFT を用いてサムネイル画像の文字領域を検出し，レイアウトと文字領域のヒートマップを取得した（図 3）．次に，サムネイル画像の文字領域をインペインティング手法を用いて塗りつぶし，背景画像を取得した．文字領域の塗りつぶしには Zhao らが提案している画像補完モデル Co-Mod-GAN [6] を使用した．本稿では，サムネイル画像に CRAFT を適用して得られた文字領域のヒートマップをマスク画像として入力することで，サムネイル画像に含まれる文字を塗りつぶした背景画像を得た（図 4）．

4. 評価実験

4.1 実験方法

本実験では提案手法を用いて背景画像からレイアウト生成を行ない，生成されたレイアウトと背景画像の関連性，二段階学習の有効性，ランダムクロップの効果，の 3 つについて評価する．

まず，レイアウトと背景画像の関連性について，同じ背景画像を用いて生成を繰り返した際に，画像上のどの領域にエレメントが配置されるかを集計（以下，背景ヒートマップと呼ぶ）して定性的に評価する．これは，いくつかの背景画像を選んで，エレメントが配置されやすい領域とされにくい領域を求め，それらをヒートマップを用いて視覚化することにより行なう．具体的には，使用する背景画像は変化させず，条件として与えるアスペクト比をランダムに選んでレイアウトを生成することを 100 回繰り返す．そして，生成された 100 個のレイアウトを集計し，背景画像上でエレメントが配置されやすい領域とされにくい領域を調べる．

また，二段階学習の有効性とランダムクロップ効果について，それぞれの手法を取り入れた場合と取り入れなかった場合の検証用データでの復元誤差を比較することで評価する．特にランダムクロップについてはランダム性を決定するパラメータ q が存在する．$0 \leq q \leq 1$ の範囲内で値が小さいほどランダム性が高くなり，この q を変化させながら評価する．

4.2 実験結果

いくつかの背景画像を用いて作成したヒートマップを図 5 に示す．図中の (1) ～ (4) ように，画像内の主なコンテンツである物体や人物が背景部分にはっきりと写っている場合は，その領域を避けてエレメントが配置されることを確認した．しかし (5), (6) ように，人に重なってしまう例も見られた．このような場合の背景画像には，画像端で見切れている，後ろ向きになっている，大きさが小さいといった傾向があり，人が背景の一部として認識されたことが原因であると考える．一方で (7), (8) は主なコンテンツが無い背景画像を用いてレイアウト生成した結果を示している．図から分かるように，ヒートマップの分布が画像全体に広がることを確認した．

次に，二段階学習についての評価実験の結果を図 6 に示す．第一段階の学習を行なった場合には学習がうまく進ん

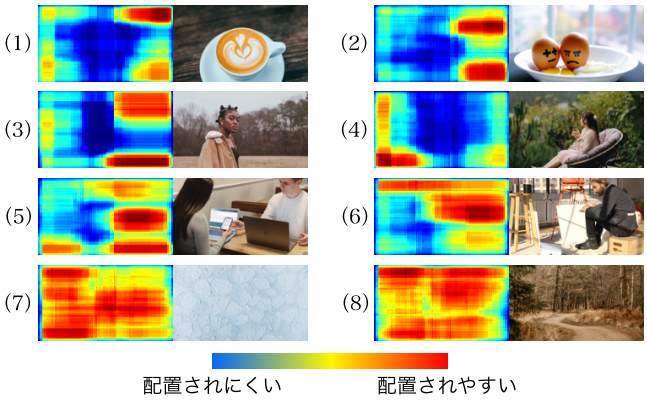

図 5 エレメント配置場所のヒートマップ

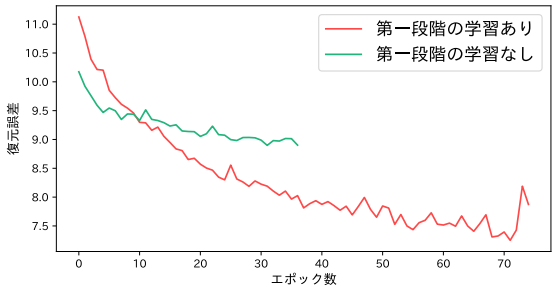

図 6 二段階学習の検証用データにおける復元誤差の推移

でいるが，第一段階を行なわない場合には途中で学習に失敗してしまった．これは背景画像からレイアウトへの直接的な変換が難しく，うまく潜在空間を構築できなかったことが原因であると考える．第一段階の学習を行なうことでレイアウトを表現する潜在空間がより適切に構築され，背景画像からレイアウトを出力するという課題をより簡単化できたと考える．

更に，ランダムクロップについての評価実験の結果を図 7 に示す．背景画像とレイアウトに対してランダムクロップを施すことで，学習時の検証用データでの復元誤差の値を小さくできることを確認した．特にパラメータ q を小さくするほど，つまりランダム性を高めるほど検証用データでの復元誤差の値を小さくできた．これはデータ拡張を行なうことで，モデルが 1 対 1 の関係のみではなく，背景画像に含まれる内容に注目して学習を行なうことができたからだと考えられる．

5. むすび

本稿では，背景画像の内容を考慮してサムネイル画像の文字レイアウトを自動生成する手法を提案した．背景画像と文字レイアウトは単純に結びつけることは難しく，直接的な変換は困難である．そこで，提案手法では二段階学習を行ない，潜在空間を経由することで背景画像から文字レイアウトへの変換を実現した．第一段階では潜在空間が適切な文字レイアウトを表現できるように学習し，任意の潜

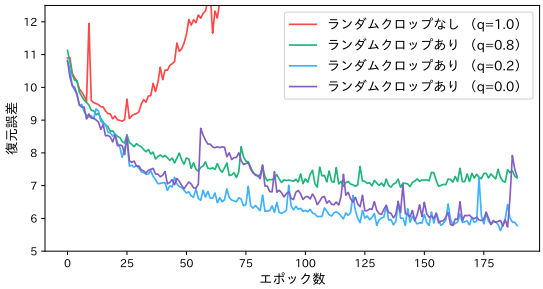

図 7 ランダムクロップの検証用データにおける復元誤差の推移

在変数に対して適切な文字レイアウトを生成できるようにした．第二段階では背景画像からレイアウトへの変換を学習し，背景画像から第一段階で学習した潜在空間へのマッピングを学習する．これにより，背景画像に対して最適なレイアウトを生成できるようにした．

提案手法を評価するために評価実験を行なった．1 つ目は背景画像を考慮できているかどうかを調査するための実験であり，ヒートマップを用いて定性的に評価した．この結果，一部で失敗する例も見られたものの，背景画像の内容を考慮してサムネイル画像の文字レイアウトを生成できることを確認した．2 つ目は提案手法で使用した二段階学習の有効性とランダムクロップの効果を評価するものであり，どちらも性能向上に繋がることを確認した．

提案手法では文字の配置場所や大きさのみを扱ったが，文字を見えやすくするために背景画像の色調に合わせて文字の色を変化させることが重要である．そのため，今後はレイアウトだけでなくスタイルも同時に出力する手法の検討が必要である．

参考文献

[1] YouTube, https://www.youtube.com.
[2] YouTube Data API — Google Developers, https://developers.google.com/youtube/v3.
[3] Arroyo, D. M., Postels, J. and Tombari, F.: Variational Transformer Networks for Layout Generation, *in* Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13642–13652 (2021).
[4] Baek, Y., Lee, B., Han, D., Yun, S. and Lee, H.: Character Region Awareness for Text Detection, *in* Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9365–9374 (2019).
[5] Li, J., Yang, J., Hertzmann, A., Zhang, J. and Xu, T.: LayoutGAN: Synthesizing Graphic Layouts With Vector-Wireframe Adversarial Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 7, pp. 2388–2399 (2021).
[6] Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E. I. and Xu, Y.Large Scale Image Completion via Co-Modulated Generative Adversarial Networks (2021). arXiv preprint arXiv:2103.10428.
[7] 佐藤亮介，田村良一：YouTuber の動画における視聴者に選択されるサムネイル画像とタイトルの研究，日本感性工学会論文誌，Vol. 18, No. 1, pp. 139–145 (2019).