ストーリとしての動画キャプションの自動評価法

藤田 綜一郎[†] 平尾 努[§] 上垣外 英剛[‡] 奥村 学[‡] 永田 昌明[§]
[†]東京工業大学 [‡]東京工業大学科学技術創成研究院 [§]日本電信電話株式会社
[†]{fujiso, kamigaito}@lr.pi.titech.ac.jp [‡]oku@pi.titech.ac.jp
[§]{tsutomu.hirao.kp,masaaki.nagata.et}@hco.ntt.co.jp

1 はじめに

Dense Video Captioning [1] は動画中の場面 (シーン) に応じてキャプション文を与えるタスクである．一般的には，2 分程度の動画を平均 3 から 4 文程度のキャプションで表現するタスクであり，Vision and Language 分野における主要な研究課題の一つである．動画中の場面に対しキャプションを与えインデキシングしておけば，自然言語をクエリとして高度な動画検索が可能となる．さらに，動画の場面に与えられたキャプションは動画全体のストーリを言語化したものともいえるので，人間が動画の概要を把握するのにも大いに役立つ．

一般的に Dense Video Captioning の評価には，システムが出力したキャプション文と対応関係にある正解のキャプション文の間の METEOR スコア [2] の平均が用いられるため，システムが生成したキャプション文のうち，正解のキャプション文と対応がとれるものしか評価の対象とならない．この評価法は，動画のインデキシングという観点からは理にかなった評価，つまり，検索の再現率を重視して評価するという点では問題ないが，百文を超えるような過剰に生成されたキャプションに対しても高いスコアを与えてしまうことがあるため，動画のストーリの評価という観点では問題がある．そこで本稿では，ストーリ生成という観点から Dense Video Captioning を自動的に評価する手法を提案する．

2 従来の Dense Video Captioning の自動評価

人間の動作に関する動画を大量に集めた ActivityNet データセット [3] に対して各動画中の場面に人間がキャプションを付与した Dense Video Captioning データセット [1] が公開されており，評価型ワークショップである ActivityNet Challenge では，このデータセットを用いて Dense Video Captioning システムの性能の比較評価が盛んに行われている．

ActivityNet Challenge では，文献 [1] で提案された自動評価法を公式指標として採用している．いま，ある動画に対する正解のキャプション文の集合を $\mathcal{G}$，システムが生成したキャプション文の集合を $\mathcal{P}$ とする．ここで，$g(\in \mathcal{G})$ と $p(\in \mathcal{P})$ をそれぞれ正解およびシステムキャプション文とする．なお，個々のキャプション文には，それが動画中のどの場面に対応するかをあらわす時間情報 (以降，プロポーザルと呼ぶ) が与えられている．プロポーザルの開始時間を関数 $s()$，終了時間を関数 $e()$ で得られるものとして，g と p の時間の重なり，IoU (Intersection of Union) を以下の式で定義する．

$$\mathrm{IoU}(g,p)=\frac{\min(e(g),e(p))-\max(s(g),s(p))}{\max(e(g),e(p))-\min(s(g),s(p))} \quad (1)$$

なお，分子が負になった場合には IoU はゼロとする．ここで，p に対しあるしきい値 τ 以上の IoU を持つ正解キャプション文の集合を以下で定義する．

$$G_{p,\tau} = \{g \in \mathcal{G} | \mathrm{IoU}(g,p) \geq \tau\} \quad (2)$$

次に，システムが生成したキャプション集合 $\mathcal{P}$ に対し，$\mathcal{G}$ の要素と対応関係を持つ部分集合 P を以下で定義する．

$$P = \{p \in \mathcal{P} | G_{p,\tau} \neq \phi\} \quad (3)$$

ここで，P と $G_{p,\tau}$ を用いてシステムキャプションの評価スコアを以下の式で定義する．

$$E(\mathcal{G},\mathcal{P},\tau) = \frac{\sum_{p\in P}\sum_{g\in G_{p,\tau}} f(g,p)}{\sum_{p\in P}|G_{p,\tau}|} \quad (4)$$

正解キャプション文とシステム生成キャプション文の間の自動評価スコアを返す関数 $f()$ としては，METEOR [2]，BLEU [4]，CIDEr [5] が利用可能であるが，通常は METEOR が用いられる．よって以下では $f()$ として METEOR を用いるものとする．$\tau = 0.9, 0.7, 0.5, 0.3$ のスコアを計算し，その平均が最終的なスコアとして用いられる．

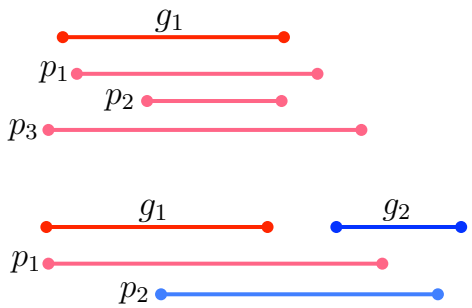

図 1: システムプロポーザルと正解プロポーザルの例．キャプションの内容は色に対応する．

3 従来法の問題点

3.1 冗長なシステムキャプションに寛容なペア生成

式 (4) より，METEOR スコアの計算対象となるのは τ 以上の IoU を持つすべての g と p のペアである．よって，1 つの g と複数の p がペアになること，1 つの p に対して複数の g がペアになることがある．前者の場合，同じキャプション文に対し長さの異なる複数のプロポーザルを用意しておくとしきい値 τ を変化させてもいずれかの p が g と対応するので METEOR スコアを効率的に稼ぐことができる．図 1 上段の例では，しきい値が低い場合には p_1 から p_3 のすべてが，高い場合には p_1 のみが g_1 と対応する．後者の場合，p に長いプロポーザルを与えると正しい g と対応する可能性が高くなり，これも METEOR スコアを効率的に稼ぐことができる．図 1 下段の例では，しきい値が低い場合には，p_2 は g_1 と g_2 の両方に対応し，高い METEOR スコアを得られるペア (p_2, g_2) を構成することができる．

3.2 キャプションの情報量と無関係なスコアの計算法

前章で説明した評価指標は，式 (4) の分母が $\sum_{p \in P} |G_{p,r}|$ であることから，システムが生成したキャプション文のうち，IoU に基づき何らかの g と対応づいた p のペアのみで METEOR スコアを平均する．つまり，システムが何文キャプションを出力するかは直接スコアに影響しない．ここで，前節で説明したように IoU がしきい値以上の g と p のペアのみを評価対象とするのであれば，(a) キャプション全体の情報量を重要視するため，複数のキャプション文を用意し，それぞれに複数のプロポーザルを与える，(b) 確度の高いキャプションのみを出力するため，確信度の高い 1 つの文に複数のプロポーザルを与えることが可能となり，従来の評価法では (a)，(b) とも高いスコアを獲得する可能性が高い．つまり，システムキャプションがどの程度正解キャプションの情報を網羅しているか (いわゆる再現率)，システムキャプションのうちどの程度が正解キャプションと合致しているか (いわゆる精度) という観点を無視した評価となっている．実際，正解キャプションが 1 つの動画に対し平均 3.5 文であることに対し，システムは数十から多いときには百を超えるキャプション文[1]を出力している．これは，主に (a) の理由であり，動画の検索結果を再現率を重視して評価したい，ある自然言語のクエリに対して合致する動画を漏れなく探したいという要求に対しては問題ない．しかし，動画キャプションのストーリ生成という観点で捉えると，そもそも大量のキャプションは人間が読むことのできるものではないし，冗長なキャプションが高いスコアを得ることは問題となる．

また，(b) の同じ文を異なるプロポーザルで生成し高いスコアを得たとしても，再現率を無視しているので基本的には何らかの応用に役立つことはない．

このように従来の自動評価法には IoU を用いたキャプションの対応決定に問題がある上に再現率，精度という観点を欠くため，大量の文を含むキャプションが高い評価を得る可能性がある．

4 IoU の最適割当に基づく Dense Video Captioning の自動評価

本研究ではシステムキャプションと正解キャプションの対応を時系列を考慮した上で IoU の和を最大化するように決定し，再現率，精度に基づき評価スコアを計算する手法を提案する．

4.1 システムキャプションと正解キャプションの最適な割当

冗長な対応を避けつつ IoU の重なりを重視するため，システムキャプション文に対応する正解キャプション文の数を高々1 つであること，時間順序に応じて対応すること[2]を制約として，対応関係にあるシステムキャ

[1] 公式評価スクリプトは 1000 文までのキャプション文を許容している

[2] g_i と p_j が対応する場合，$g_\ell (\ell > i)$ と対応することのできる p のインデックスは $m > j$ となる．

	p_1	p_2	p_3	p_4	p_5
g_1	0.1	0.3	0.2	0.8	0.1
g_2	0.1	0.3	0.1	0.8	0.5
g_3	0.9	1.0	0.3	0.9	0.8
g_4	0.3	0.5	0.6	1.0	0.1

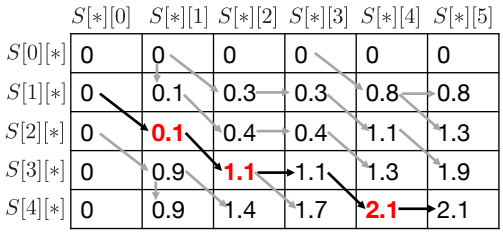

図 2: 動的計画法による対応関係の決定

プション文と正解キャプション文との間の IoU のスコアの和が最大となるよう対応関係を決定する．

まず，i 番目の正解キャプション g_i と j 番目のシステムキャプション p_j の時間の重なりをあらわすスコアを IoU としきい値 τ を用いて以下の式で定義する．

$$C_{i,j} = \begin{cases} \mathrm{IoU}(g_i, p_j) & \mathrm{IoU}(g_i, p_j) \geq \tau \\ 0 & (\text{otherwise}) \end{cases} \tag{5}$$

次に，i 番目までの正解キャプションと j 番目までのシステムキャプションの時間の重なりをあらわすスコア $C_{i,j}$ の和の最大値を格納するテーブルを $S[i][j]$ とする．$S[i][0] = 0 (0 \leq i \leq |\mathcal{G}|)$，$S[0][j] = 0 (0 \leq j \leq |\mathcal{P}|)$ を初期値として，以下の漸化式を用いることで，最適値，$S[|\mathcal{G}|][|\mathcal{P}|]$ を得る．

$$S[i][j] = \max \begin{cases} S[i-1][j] \\ S[i-1][j-1] + C_{i,j} \\ S[i][j-1] \end{cases} \tag{6}$$

g_i と p_j に対応がするのは $S[i][j]$ を $S[i-1][j-1]+C_{i,j}$ で与えたときである．

図 2 に例を示す．図 2 上段のように $C_{i,j}$ が与えられた場合，スコアテーブル S は図の下段となる．テーブルの任意のセル，つまり $S[i][j]$ の値は，真上の値，真横の値，左斜め上の値と $C_{i,j}$ の和のうち最も大きいものである．例では最大スコア $S[|\mathcal{G}|][|\mathcal{P}|]$ は 2.1 となり，それを与える g と p の対応は，(g_4, p_4)，(g_3, p_2)，(g_2, p_1) となる．

4.2 キャプション評価スコアの計算

前節で決定したキャプション間の対応に基づき，以下の式で精度，再現率を定義する．

$$\mathrm{Precision}(\mathcal{G}, \mathcal{P}) = \frac{\sum_{g \in \mathcal{G}} f(g, p_{a(g)})}{|\mathcal{P}|} \tag{7}$$

$$\mathrm{Recall}(\mathcal{G}, \mathcal{P}) = \frac{\sum_{g \in \mathcal{G}} f(g, p_{a(g)})}{|\mathcal{G}|} \tag{8}$$

そして，$\mathrm{F}_1 (= 2 * \mathrm{Precision} * \mathrm{Recall} / (\mathrm{Precision} + \mathrm{Recall}))$ を最終的なシステムキャプションの評価スコアとする．なお，$a()$ は g に対応する p のインデックスを返す関数である．

正解キャプション文とシステムキャプション文の対応を高々1 対 1 に絞ったこと，式 (7) の分母が $|\mathcal{P}|$ であることから，文数が多いキャプションは再現率は高いかもしれなが，精度が低くなるので，F_1 は低くなる．また，同じキャプション文を異なるプロポーザルで多数出力すると再現率，精度とも低くなる．

なお，提案手法のバリエーションとして，式 (5) のかわりに以下の式 (9) で $C_{i,j}$ を定義することで IoU と METEOR スコアの双方を同時に考慮した評価スコアを得ることができる．

$$C_{i,j} = \mathrm{IoU}(g_i, p_j) f(g_i, p_j) \tag{9}$$

5 実験

5.1 設定

Krishna ら [1] の ActivityNet に対するキャプションデータのうち，開発用データ (4883 件)[3]を用いて，End-to-end Transformer システム [6] のキャプションを利用して従来の自動評価指標と提案法とを比較した．

また，文献 [1] で配布されている評価用スクリプトでは，開発データとテストデータに対しては，1 つの動画に対して異なる 2 名のキャプションが与えられており，評価の際には 2 名のキャプションをまとめて 1 つの正解キャプションとして扱っているので，本稿でもこの設定に合わせる．

3 章で指摘したように従来の自動評価法はキャプションの内容が過剰であっても過少であっても正解のキャプション文と IoU が τ 以上であるものだけで評価スコアを計算する．よって，システムキャプションの文数によらず似たようなスコアをとることが予想される．一方，提案法は精度，再現率に基づくスコアなのでシ

[3]テストデータに対するキャプションは公開されていない．

		k			
		3	7	15	200
	従来法	5.00	5.58	5.69	6.06
Prop.(a)	Precision	3.31	2.93	2.37	0.215
	Recall	1.41	2.83	3.16	6.99
	F_1	1.98	2.89	2.71	0.410
Prop.(b)	Precision	9.24	7.18	5.95	0.326
	Recall	3.94	7.15	7.92	10.6
	F_1	5.53	7.16	6.80	0.631
Prop.(c)	Precision	5.18	4.02	3.47	0.290
	Recall	2.12	4.00	4.61	9.46
	F_1	3.10	4.01	3.96	0.570

表 1: k を変化させた場合の評価スコア

ステムキャプション数が多ければ，再現率が高く，精度が低い，少なければ再現率が低く，精度が高いことが予想される．これを確認するため，システムキャプションから重複なくランダムに k 文を抽出し各評価スコアがどう変化するかを調べた．文献 [6] のシステムは 1 つの動画あたり平均で 200 文のキャプションを生成すること，正解キャプションの平均文数が 7 文であることより，$k = 3, 7, 15, 200$ を試した．

5.2 結果

表 1 に実験結果を示す．なお，表中 $k = 200$ 以外の値はランダム試行を 5 回行った平均値であり，従来法，Prop.(a) は τ を 0.9,0.7,0.5,0.3 のスコアの平均，Prop.(b) は $\tau = 0$ の場合，Prop.(c) は式 (9) を用いた場合である．

表より，従来法は k を変化させても評価スコアの変動が小さい．200 文全体を評価したスコアとそこからランダムに選んだ 7 文を評価したスコアがほぼ変わらず，やや 200 文の方が良いスコアであることは動画に対するストーリ生成の評価という観点では明らかに問題である．

提案法はいずれの場合も k を大きくすると再現率が向上し，精度が低下する．k を小さくするとその逆の振る舞いをする．これは意図したとおりの結果であり，人間が読むことが負担になるような大量のキャプションに対しては低いスコアを与える．提案法のバリエーションを比較すると，Prop.(a) は高い τ，つまり $\tau = 0.9, 0.7$ のスコアが低いため全体で平均をとるとスコアレンジが低くなっている．Prop.(b)，Prop.(c) は，Prop.(a) よりもスコアレンジも高く，変動もやや大きい．k は人間がキャプションを読むうえで重要なパラメタであるがゆえ，それに対し敏感な Prop.(b) か Prop.(c) がストーリとしてのキャプションを評価するという観点では良いであろう．両者の違いについては，被験者実験などを利用して今後明らかにしたいと考える．

6 おわりに

本稿では，Dense Video Captioning を動画に対するストーリ生成としてとらえ，時系列を考慮した上で正解キャプションとシステムキャプションの対応を決定する，再現率，精度に基づく評価スコア計算法を提案した．End-to-end Transformer システムによるキャプションを利用した人工データによる実験結果から，従来の自動評価法はキャプションの分量に鈍感で動画のストーリの評価には向いていないこと，提案法はキャプションの分量に応じて敏感にスコアが変動することを確認した．

参考文献

[1] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of CVPR*, pp. 706–715, 2017.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65–72, 2005.

[3] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understan ding. In *Proceedings of CVPR*, pp. 961–970, 2015.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, pp. 311–318, 2002.

[5] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of CVPR*, pp. 4566–4575, 2015.

[6] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of CVPR*, pp. 8739–8748, 2018.