

料理タスクにおける将来イベントのクロスモーダル説明文生成

神原元就† 杉浦孔明†

†慶應義塾大学

E-mail: motonari.k714@keio.jp

1 はじめに

日常生活において、エージェントが近未来に起こりうるイベントを予測し説明できることは有用性が高い。例えば日常生活の介助では、エージェントへの指示に対し、動作中の危険を予測し自然言語を用いてユーザへフィードバックすることで、衝突等の危険なイベントの回避を促せる。一方で、このタスクを行う能力は十分ではない。なぜなら、このタスクは、イベントの自然言語による説明だけでなく現在までのイベントから将来のイベントを推論することも求められる点で、困難なためである。

本論文では、時刻 t までのクリップを基に時刻 $t+1$ のイベントについての説明文を生成するタスクである、Future captioning タスク [Hosseinzadeh 21] を扱う。特に、料理タスクにおける Future captioning タスクを扱う。

このタスクは、モデルは将来のクリップを利用することができないという点で難しい。そのため、過去イベントを用いた将来イベントの予測、及びキャプションの生成という2つの要素が求められる。実際、4章で示すように、動画キャプション生成タスクのためのモデルを適用した場合、正解文と生成文の品質には大きな差が見られる。

これは、既存手法では、イベントの視覚特徴系列と文の関係のモデル化が不十分であるためである。それらの手法における transformer の自己注意では、イベントの即時的な視覚特徴と文間の関係がモデル化されている。また、多くの動画キャプション生成モデルは、次時刻以降のキャプションを用いて次時刻のキャプションを行うモデルもあるため Future captioning タスクには不適切である。

Future captioning タスクが含まれるキャプション生成分野には多くの研究がある [Wang 18, Krishna 17, Lei 20, Kambara 21, Magassouba 21]。このタスクにおける代表的な手法として [Hossain 19, Mori 21, Mahmud 21] が挙げられる。

提案手法は3つのモジュールから構成され、それぞれ Relational Self-Attention (RSA [Kim 21]) エンコーダ、transformer エンコーダ、および transformer デコー

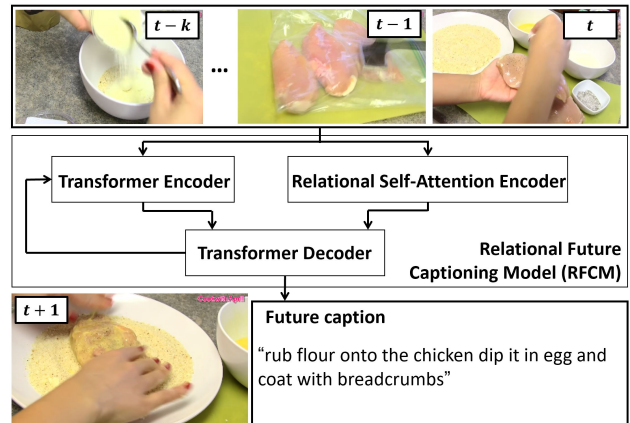


図1 RFCMの概要図。

ダである。提案手法はRSA エンコーダによって、過去のイベントとの関係性を適切に考慮した将来のイベント表現を獲得できる。なぜならRSAは、既存の注意機構よりも効果的に、過去のイベントのうちどのイベントに注目すればよいかを学習できるためである。また、transformer デコーダは過去のイベントの関係性から、適切に将来イベントのキャプションを生成できる。transformer デコーダは、RSA エンコーダの出力をクエリとし、transformer エンコーダの出力をキー、及びバリューとする source-target 注意機構を持つ。これによって、過去のイベント間の関係性を自然言語に適切に接地できる。

本研究の独自性は以下である。

- Future captioning タスクのためのクロスモーダル言語生成モデル、RFCMを提案する。
- RSA エンコーダの導入により、既存の自己注意機構に比べ、より効果的にイベント間の関係性を抽出できる。

2 問題設定

本論文では Future captioning タスクを扱う。Future captioning タスクでは時刻 t までのクリップから時刻 $t+1$ におけるクリップの説明文を生成することを目的とする。図1に本タスクの典型例を示す。本タスクでは、図1に示したように、時刻 t までのクリップから時刻 $t+1$ での状況を推論し説明文を出力することが望ましい。本タスクにおいて、モデルへの入力時刻 t ま

でのクリップであり、出力は時刻 $t+1$ に起きる出来事についての説明文である。ここで、本稿ではクリップをイベントを描写したフレーム系列と定義する。本稿では、他タスクでの事前学習は許可されていないものとする。これは、本研究は知識転移を目的としたものではないためである。

3 提案手法

ネットワークの構造を図2に示す。図において、Norm はレイヤ正規化 (layer normalization, LN) 層を表す。ネットワーク入力以下で表される、時刻 $t-k$ から時刻 t までの動画クリップとする。

ネットワーク入力 \mathbf{x} は $\mathbf{x} = \{\mathbf{x}_{t-k}, \dots, \mathbf{x}_t\}$, で定義される。ここで、 $\mathbf{x}_t \in \mathbb{R}^{d_{in}}$ 及び $k \in \mathbb{N}$ はそれぞれ時刻 t におけるクリップ及び時刻 t 以前のクリップ数を表す。

\mathbf{x} を基に、 $\mathbf{Z}_{t-k:t+1} \in \mathbb{R}^{(k+2) \times d_{in}}$ を以下の式により獲得する。

$$\mathbf{z}_\tau = \begin{cases} f_z(\mathbf{x}_\tau) & (\tau > t) \\ f_z(\mathbf{x}_\tau, \mathbf{x}_t) & (\tau \leq t), \end{cases}$$

$$\mathbf{Z}_{t-k:t+1} = \{\mathbf{z}_{t-k}; \dots; \mathbf{z}_{t+1}\},$$

ここで、 $\tau = t-k, \dots, t+1$, $f_z(\cdot)$ は線形変換を表す。 $\mathbf{Z}_{t-k:t+1}$ はイベント系列に関する情報を含む。

3.1 RSA エンコーダ

RSA エンコーダは Relational Self-Attention [Kim 21] を利用した構造を持ち、イベント間の関係性を抽出する。RSA エンコーダは N_r 層の RSA 層から構成される。第一層目では、 $\mathbf{Z}_{t-k:t+1}$ に対して [Vaswani 17] と同様の手順で、三角関数を用いた位置埋め込みを行う。

その後、クエリ $\mathbf{Q}_r \in \mathbb{R}^{d_{rsa}}$, キー $\mathbf{K}_r \in \mathbb{R}^{(k+2) \times d_{rsa}}$, 及びバリュー $\mathbf{V}_r \in \mathbb{R}^{(k+2) \times d_{rsa}}$ を以下のようにして獲得する。

$$\mathbf{Q}_r = \mathbf{z}_t,$$

$$\mathbf{K}_r = \mathbf{V}_r = \mathbf{Z}_{t-k:t+1},$$

ここで、 d_{rsa} は各層の次元数を表す。続いて、basic kernel $\varphi_p \in \mathbb{R}^{k+2}$ 及び relational kernel $\varphi_h \in \mathbb{R}^{k+2}$ を以下のようにして得る。

$$\varphi_p = \mathbf{W}_p \mathbf{Q}_r,$$

$$\mathbf{Q}_{r'} = \{\mathbf{Q}_r; \dots; \mathbf{Q}_r\} \in \mathbb{R}^{(k+2) \times d_{rsa}},$$

$$\varphi_h = \mathbf{W}_h f_{\text{flatten}}(\mathbf{Q}_{r'} \odot \mathbf{K}_r),$$

ここで、 $f_{\text{flatten}}(\cdot)$ はベクトル化を表す。続いて、relational context $\Phi_g \in \mathbb{R}^{(k+2) \times d_{rsa}}$ を以下のようにして得る。

$$\Phi_g = \mathbf{V}_r + \mathbf{W}_g \mathbf{V}_r^\top \mathbf{V}_r.$$

最終的に RSA $\varphi \in \mathbb{R}^{d_{rsa}}$ を以下のようにして得る。

$$\varphi = (\varphi_p + \varphi_h)^\top \Phi_g.$$

RSA φ は入力イベント間の関係性についての情報を含む。 φ を時刻 t における潜在特徴量として用いるため、 \mathbf{z}_t を φ によって置き換える。これにより、中間特徴量 $\mathbf{h}_r = \{\mathbf{z}_{t-k}; \dots; \mathbf{z}_{t-1}; \varphi; \mathbf{z}_{t+1}\} \in \mathbb{R}^{(k+2) \times d_{rsa}}$ を獲得する。各層の出力 \mathbf{h}_{n_r} ($n_r = 1, \dots, N_r$) は \mathbf{h}_r をフィードフォワードネットワーク (feedforward network, FFN) 及び LN 層を用いて変換することで得る。エンコーダの出力 $\mathbf{h}_{rsa} \in \mathbb{R}^{(k+2) \times d_{rsa}}$ は $\mathbf{h}_{rsa} = f_{LN}(f_{FFN}(\mathbf{h}_{N_r}))$ により得られる。ここで、 $f_{FFN}(\cdot)$ 及び $f_{LN}(\cdot)$ はそれぞれ FFN 及び LN 層を表す。

損失関数 \mathcal{L} は以下で定義される。

$$\mathcal{L} = \lambda_{ce} \mathcal{L}_{CE}(y_{t+1}, p(\hat{\mathbf{y}}_{t+1})) + \lambda_{iwp} \mathcal{L}_{iwp}(\mathbf{y}_{t+1,1}, p(\hat{\mathbf{y}}_{t+1,1}))$$

$$+ \lambda_{corr} \mathcal{L}_{corr} + \lambda_{mse} \mathcal{L}_{MSE}(\mathbf{x}_{t+1}, \mathbf{z}_{t+1}),$$

ここで、 $\mathcal{L}_{MSE}(\cdot, \cdot)$ 及び $\mathcal{L}_{CE}(\cdot, \cdot)$ はそれぞれ平均二乗誤差関数及び交差エントロピー誤差関数を表す。また、 λ_{ce} , λ_{corr} , λ_{mse} 及び λ_{iwp} はハイパーパラメータである。 \mathcal{L}_{corr} は $\hat{\mathbf{y}}_{t+1}$ が時刻 t よりも早い、または遅い時刻のイベントについて述べている場合にペナルティを科す。 \mathcal{L}_{iwp} は生成文の1単語目、 $y_{t+1,1}^*$ が適切でない場合にペナルティを科す。これは $\mathcal{L}_{iwp}(y_{t+1,1}, p(\hat{\mathbf{y}}_{t+1,1})) = \gamma_{iwp} \mathcal{L}_{CE}(y_{t+1,1}, p(\hat{\mathbf{y}}_{t+1,1}))$ のように定義される。ここで、 $\gamma_{iwp} = 1/W$ である。

4 実験

4.1 実験設定

データセットとして、YouCook2-FC データセットを利用した。YouCook2-FC データセットは Future captioning タスクのためのデータセットの一つである。一般的に、料理において、次の調理手順はそれまでの手順により決定される。本研究では YouCook2 データセット [Zhou 18] に基づき YouCook2-FC データセットを構築した。訓練、検証、及びテスト集合のサイズはそれぞれ 7435, 1569, 3035 とした。

以下にデータセットの事前処理における手順を示した。YouCook2-FC データセットにおいて、与えられた開始、及び終了時刻に基づきクリップを切り出した。各クリップは、それぞれ 0.6 に変換された。その後、[Ging 20] に示された手順により処理を行った。

実験設定は以下とした。最適化関数、学習率、バッチサイズ、及びエポック数はそれぞれ Adam ($\beta_1: 0.9$, $\beta_2: 0.999$), $1.0e-4$, 16, 及び 25 とした。各モジュールにおいて、 N_e , d_{enc} , N_h , N_d , d_{dec} , N_r , 及び d_{rsa} はそれぞれ、3, 384, 12, 3, 384, 2, 及び 384 とした。損失関数において、 λ_{ce} , λ_{iwp} , λ_{corr} , 及び λ_{mse} はそれぞれ

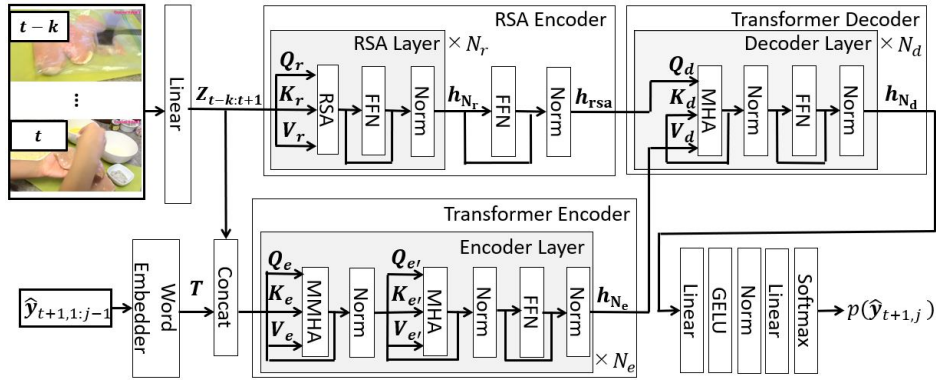


図2 RFCMのネットワーク図.

表1 各手法による生成文の評価結果

Methods	BLEU4↑	METEOR↑	ROUGE-L↑	CIDEr-D↑
MART [Lei 20]	6.85±0.18	14.24±0.07	30.80 ±0.21	20.86±1.07
Ours (w/o RSA)	6.70±0.36	14.14±0.50	30.18±0.18	21.26±2.83
Ours (w/o Transformer Decoder)	6.68±0.13	14.09±0.15	30.16±0.31	19.75±1.41
Ours (RFCM)	7.03 ±0.15	14.53 ±0.09	30.49±0.21	21.32 ±1.09

30, 1.0, 0.1, 0.005, 及び 10 とした. \mathcal{L}_{iwp} において, n_{th} 回以上出現した単語を扱った. ここで, $n_{th} = 30$ とした. また, W は 3000 とした. RFCM の訓練可能パラメータ数及び積和演算数はそれぞれ 3.1M 及び 540M であった.

訓練は 16GB メモリ搭載 NVIDIA Tesla V100 SXM2, 240 GB RAM, 及び Intel Xeon Gold 6148 プロセッサを用いて行われた. YouCook2-FC における訓練時間はそれぞれ 6.2 時間であった. 同様に, 推論時間はそれぞれ 1.7×10^{-2} s/サンプルであった. early stopping の条件として, [Prechelt 98] に示される generalization を用いた.

4.2 定量的結果

RFCM 及び Memory-Augmented Recurrent Transformer (MART [Lei 20]) について比較を行った. MART は transformer を用いた動画キャプション生成タスクにおける代表的な手法の一つであり, Future captioning タスクへの適用も可能であったことからベースライン手法とした. 表 1 に YouCook2-FC データセットにおける定量的結果を示した. 各手法につき実験を 5 回行い, 表にはその平均値及び標準偏差を示した.

生成文の評価は, 動画キャプション生成タスクにおける標準尺度である BLEU4, ROUGE-L [Lin 04], METEOR [Banerjee 05], 及び CIDEr-D [Vedantam 15] により行った. 主要尺度は CIDEr-D とした.

表より, CIDEr-D が 0.46 ポイント向上した. また, BLEU4 及び METEOR もそれぞれ 0.18 ポイント及び 0.29 ポイント向上した. この結果より, Future captioning タスクにおいて RFCM はベースライン手法よりも適切な文生成が可能であることがわかった.

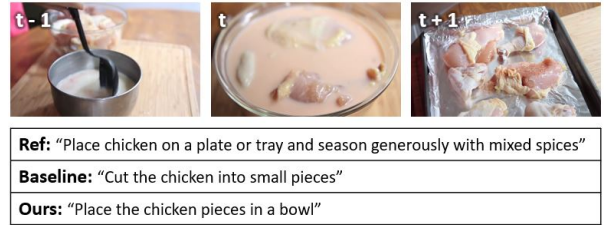


図3 YouCook2-FC データセットにおける成功例.

4.3 定性的結果

図 3 に定性的結果を示した. 各図において, 上段はイベントを示す. 各イベントは時系列順に示されている. 紙面の都合上いくつかのイベントについて省略した. また, 下段は正解文及び各手法による生成文を示す.

図 3 に YouCook2-FC データセットにおける成功例を示した. 図において, “rub flour” 及び “coat with breadcrumbs” という 2 つのイベントが起きた. ベースライン手法は “breadcrumbs” について言及しなかった. そのため, 生成文は不足していた. 一方で, 提案手法は “coat the chicken with flour and bread crumbs” と適切に記述した. 以上の結果から, 提案手法は次時刻の調理ステップを適切に予測し, キャプションを生成できたといえる.

4.4 Ablation study

各モジュールについて ablation study を行った. 表 1 に定量的結果を示した. どのモジュールが最も性能向上に寄与しているかを調査するため, ablation 条件を以下の 2 条件とした. (a) w/o RSA: RSA エンコーダにおいて, RSA 層の代わりに MHA 層 [Vaswani 17] を用いた. (b) w/o transformer decoder: transformer デコーダを取り除いた.

提案手法と条件 (a) 及び (b) を比較すると, CIDEr-D はそれぞれ 0.06 及び 1.57 ポイント低下した. これより,

transformer デコーダが最も性能向上に寄与していることがわかった。

5 結論

本稿では、将来イベントについての説明文を生成するタスクの一つである、Future captioning タスクを扱った。特に、料理タスクにおける Future captioning タスクを扱った。

本稿の主要な貢献は以下である。

- 将来イベントについての説明文を生成可能なクロスモーダル言語生成モデル、RFCM を提案した。
- RSA エンコーダの導入により、効果的にイベント間の関係性を抽出できる。
- RFCM は YouCook2-FC データセットにおいてベースライン手法を上回った。

謝辞

本研究の一部は、JSPS 科研費 20H04269, JST CREST, NEDO の助成を受けて実施されたものである。

参考文献

- [Banerjee 05] Banerjee, S., et al.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in *the ACL Workshop on IEEM for MTS*, pp. 65–72 (2005)
- [Ging 20] Ging, S., et al.: COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning, in *NeurIPS*, pp. 22605–22618 (2020)
- [Hossain 19] Hossain, Z., Sohel, F., et al.: A Comprehensive Survey of Deep Learning for Image Captioning, *ACM CSUR*, Vol. 51, No. 6, pp. 1–36 (2019)
- [Hosseinzadeh 21] Hosseinzadeh, M. and Wang, Y.: Video Captioning of Future Frames, in *WACV*, pp. 980–989 (2021)
- [Kambara 21] Kambara, M. and Sugiura, K.: Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions, *IEEE RA-L*, Vol. 6, No. 4, pp. 8371–8378 (2021)
- [Kim 21] Kim, M., Kwon, H., Wang, C., et al.: Relational Self-Attention: What’s Missing in Attention for Video Understanding, in *NeurIPS* (2021)
- [Krishna 17] Krishna, R., et al.: Dense-Captioning Events in Videos, in *CVPR*, pp. 706–715 (2017)
- [Lei 20] Lei, J., et al.: MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning, in *ACL*, pp. 2603–2614 (2020)
- [Lin 04] Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries, in *Text summarization branches out*, pp. 74–81 (2004)
- [Magassouba 21] Magassouba, A., et al.: Predicting and attending to damaging collisions for placing everyday objects in photo-realistic simulations, *Advanced Robotics*, Vol. 35, No. 12, pp. 1–13 (2021)
- [Mahmud 21] Mahmud, T., Billah, M., et al.: Prediction and Description of Near-Future Activities in Video, *CVIU*, Vol. 210, p. 103230 (2021)
- [Mori 21] Mori, Y., et al.: Image Captioning in Near Future from Vehicle Camera Images and Motion Information, in *IEEE IV*, pp. 1378–1384 (2021)
- [Prechelt 98] Prechelt, L.: Automatic early stopping using cross validation: quantifying the criteria, *Neural Networks*, Vol. 11, No. 4, pp. 761–767 (1998)
- [Vaswani 17] Vaswani, A., et al.: Attention Is All You Need, in *NeurIPS*, pp. 5998–6008 (2017)
- [Vedantam 15] Vedantam, R., et al.: CIDEr: Consensus-based Image Description Evaluation, in *CVPR*, pp. 4566–4575 (2015)
- [Wang 18] Wang, X., et al.: Video Captioning via Hierarchical Reinforcement Learning, in *CVPR*, pp. 4213–4222 (2018)
- [Zhou 18] Zhou, L., Xu, C., and Corso, J.: Towards Automatic Learning of Procedures From Web Instructional Videos, in *AAAI*, pp. 7590–7598 (2018)