進捗報告

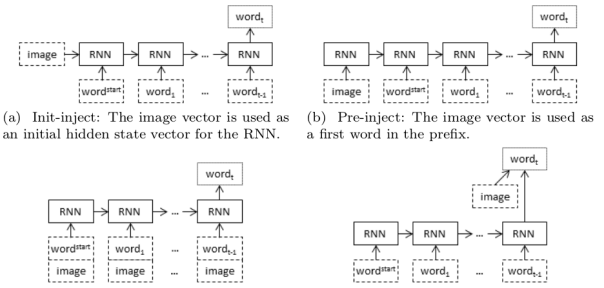
1 今週やったこと

マルチモーダルを扱う研究の準備として自然言語処理 (特に seq2seq に関わるようなところ) のリサーチをした．[1] がよくまとまっていて分かりやすかった．

- 画像キャプション生成
- transformer, BERT
- NAS セットアップ

1.1 画像キャプション生成

画像と自然言語をつなぐ代表的なタスクとして画像キャプション生成があるので調べた．図 1 に LSTM を用いた代表的な 4 種類のモデルを示す．



(a) Init-inject: The image vector is used as an initial hidden state vector for the RNN.

(b) Pre-inject: The image vector is used as a first word in the prefix.

(c) Par-inject: The RNN accepts two inputs at once in every time step: a word and an image.

(d) Merge: The image vector is merged with the prefix outside of the RNN.

図 1: 画像キャプション生成における代表的なモデル

- a:init-inject)LSTM の隠れ層の初期状態として画像の分散表現を用いる
- b:pre-inject) 画像を入力データの 1 時系列目として用いる
- c:par-inject) 各時系列に対して画像を concat して用いる (やや冗長)
- d:merge)LSTM の出力と画像の分散表現をマージする

先行研究によると init-inject と merge が良い精度を出している [2].

4 コマ漫画をマルチモーダルに LSTM でやろうとするとこのあたりをベースに組むのかなという感じ．ひとまずは去年の話題にも上がっていた LSTM を用いないセリフマッチング問題に取り組んでみる．

1.2 transformer, BERT

そのうちセリフの分散表現をとってくるときに BERT を使うと思うので，原著と論文解説等を読んだ．背景と特徴を押さえつつ従来の Encoder-Decoder モデルと対比しつつ読み，ふんわりとした理解だがアイデアの概要を掴んだ．

2 来週の予定

- 4 コマ漫画ストーリーデータセットの自然言語データのフォーマット
- 画像と自然言語の分散表現のマッチング試行
- DCAI の原稿作成 (1/31 〆) にとりかかる

参考文献

[1] AI LAB NLP. https://ai-lab.lapras.com/nlp/text-generation-2019/.

[2] Marc Tanti, Albert Gatt, and Kenneth P. Camilleri. Where to put the image in an image caption generator. CoRR, abs/1703.09137, 2017.