Attention Weight の操作による注目領域に着目したグラフ図のキャプション生成

ER20080 増田大河　　指導教授：藤吉弘亘

1. はじめに

研究開発における生産性向上のため，学術論文等の文献から研究内容を理解して説明する AI の実現が期待されている．学術論文のグラフ図は多くの文字情報を含んでおり，研究内容を理解するための重要な要素である．画像を言語化する Vison-Language の研究が盛んに取り組まれているが，多くは一般物体画像を対象としてる．そのため，グラフ図を言語で説明することは困難である．そこで本研究では，大量のグラフ図を用いて，グラフ図の理解に必要な知識を獲得した基盤モデルをファインチューニングして，グラフ図のキャプションを生成するモデルを提案する．また，グラフ図の注目領域を明示することでより詳細なキャプションの生成を目指す．

2. MatCha

MatCha[1] は，グラフ図に対する多くのタスクに適応できる基盤モデルとして提案されている．MatCha は，グラフ図を理解する上で必要な (i) グラフ図からその描画プログラム，テーブルデータへの逆変換，(ii) 数学的問題推論の 2 つのタスクによって事前学習している．MatCha のモデル構造を図 1 に示す．事前学習によりグラフ図の理解や数学的推論に特化しているが，グラフ図に対する質問応答を行う ChartQA などの下流タスクに適応するにはファインチューニング (FT) が必要となる．

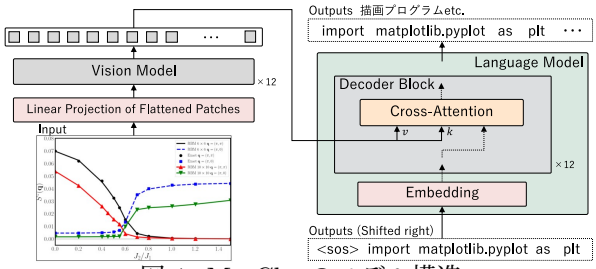

図 1: MatCha のモデル構造

3. 提案手法

グラフ図のキャプション生成を行うモデルを構築し，グラフ図の注目領域を強調することで詳細なキャプションを生成する方法を提案する．

モデルの構築　MatCha を FT することでグラフ図のキャプションを生成するモデルを構築する．MatCha によるグラフ図の表現を活用してキャプションを生成するために，画像エンコーダ部分は凍結して言語生成モデル部分のみを FT する．

注目領域の強調方法　人間が指定した任意の領域に対する詳細なキャプションを生成するために，指定した注目領域に対応するパッチの Attention Weight の値に対して γ を加算する．これにより注目領域に着目した特徴抽出を行う．注目領域を強調した Attention を式 (1) に示す．

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \left(\mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) + \mathbf{M}\times\gamma\right)\mathbf{V} \tag{1}$$

ここで，$\mathbf{M}$ は注目領域に対応するパッチ部分が 1，それ以外のパッチ部分が 0 になっているマスクである．注目領域の強調方法を図 2 に示す．これにより，画像全体としての特徴を捉えつつ，注目領域にも着目したキャプションが生成できる．

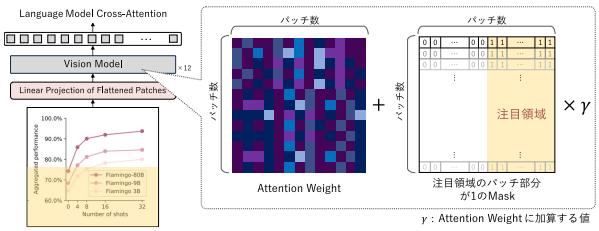

図 2: 注目領域の強調方法

4. 評価実験

提案手法の有効性を示すために，従来手法とのキャプション生成性能の比較を行う．また，指定領域の強調による生成文の変化について確認する．従来手法には M4C-Captioner を用いる．提案手法と従来手法はグラフ図とキャプションなどから構成される SciCap+データセットを用いて FT した．

4.1. キャプション生成性能の比較

キャプション生成の評価指標による比較結果を表 1 に示す．これより，ROUGE-L で 5.7pt，CIDEr で 26.1pt の精度の向上を確認した．BLEU-4 の評価が低下しているが，CIDEr の評価が向上していることから，重要な単語を抜き出すことが出来ていることがわかる．

表 1: キャプション生成の評価指標による比較

モデル	BLEU-4	ROUGE-L	CIDEr
M4C-Captioner †	**1.5**	15.4	4.6
Ours	1.3	**21.1**	**30.7**

†:SciCap+論文値

4.2. 注目領域の強調による生成文の変化

生成したキャプションに対応するグラフ図の領域を定量的に評価する．生成文中に評価領域に存在する単語が含まれる割合を評価する．x，y 軸周辺にそれぞれ注目することを期待して，強調する領域を画像の下半分と左半分とした．加算する値 γ を 12.5×10^{-5} とした時の結果を表 2 に示す．ここで，γ の値はあらかじめグリッドサーチによって探索した値である．表 2 より，強調なしと比較して，画像の下半分を強調した場合は下半分の数値が増加している．一方で，画像の左半分を強調した場合は左半分の数値が減少し，下半分の数値が増加した．

表 2: 生成文中に評価領域の単語が含まれる割合 (%)

強調する領域	評価領域		
	全体	左半分	下半分
強調なし	21.5	**22.6**	21.6
左半分	22.0	21.8	**22.8**
下半分	**22.1**	22.2	**22.8**

注目領域の強調による生成文の変化を図 3 に示す．これより，修正なしの場合と比べて，注目領域として強調した部分に含まれる単語を生成文に多く含んでいることが確認できる．これらの結果より，注目領域に対応する画像エンコーダの Attention Weight に値を加算することで，注目領域に合わせたキャプションの生成が可能である．

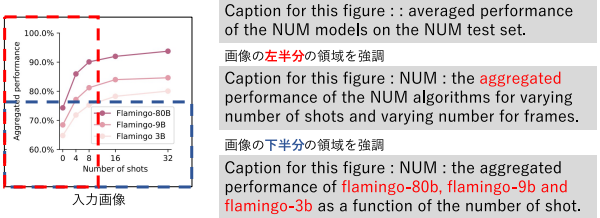

図 3: 注目領域の強調による生成文の変化

5. おわりに

本研究では，グラフ図を詳細に説明するモデルの構築と，その注目領域の強調方法を提案した．評価実験より，構築したモデルは既存手法と比較してグラフ図のキャプション生成精度が高いことを確認した．また，注目領域の強調によって，より詳しい文章生成が可能になることを確認した．今後はモデルに関連文も入力することで，キャプションの生成精度の向上を図る．

参考文献

[1] Fangyu Liu, *et al.*, “MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derenderingn”, In ACL, 2023.