グラフィカルアブストラクト推薦と評価の統合ベンチマーク

川田 拓朗[1,a)]　北田 俊輔[1,b)]　根本 颯汰[1,c)]　彌冨 仁[1,d)]

概要

Graphical Abstract (GA) は論文の要点を視覚的に伝える重要な表現手段である. 効果的な GA の作成には高度なデザインスキルが求められるため, 設計支援技術の実現が期待される. 近年, 科学論文の本文や図に加えて GA を含む論文データセット SciGA が構築され, GA 設計支援の計算基盤が整いつつある. 本研究では, GA 設計支援を目的とする 2 つの推薦タスクを定義する: 1) Intra-GA Recommendation: 同一論文内から GA に適した図を推薦するタスク. 2) Inter-GA Recommendation: ある論文の GA 作成において, 参考となる他の論文の GA を推薦するタスク. 各タスクに対し, SciGA を用いて複数手法のベンチマーク評価を行い, 性能を比較・分析する. さらに, 正解が曖昧な推薦タスクにおける新評価指標 Confidence Adjusted top-1 ground truth Ratio (CAR) を提案する. CAR は, 構築されたたランキングに対するモデルの確信度とその妥当性を考慮し, 推薦結果をインスタンスレベルで評価する. 本研究は, GA 推薦による研究成果の視覚的伝達の支援と, その評価枠組みの確立に貢献する [*1].

1. はじめに

科学的発見とその伝達は, 新たな知識の構築に不可欠であるが, その進展は研究者の限られた知識的・時間的リソースに依存している. この課題の解決に向け, 近年では AI の科学的発見への応用が注目されており, 仮説生成 [15], [16] や実験設計 [1], [22] といった取り組みがある. 同様に, AI による論文執筆 [14] や発表資料の自動生成 [8], [17], [21] などの研究成果の伝達支援も重要な研究課題となっている.

Graphical Abstract (GA) は, 科学論文の主要な知見を視覚的に要約するための重要な手段として注目されている. 特にソーシャルメディア上での共有時には, Altmetric Attention Score や読者の関心を高める効果 [3], [11] が報告されている. 近年, 正式な GA が採用されていない場合でも, 研究者は論文内の 図 1 などの視覚的素材を事実上の GA として頻繁に使用している. こうした傾向が高まるにも関わらず, GA を効果的に設計・活用する方法論は確立されていない.

本研究では, GA 設計支援を目的とし, 以下の 2 つの推薦タスクを定義する: 1) **intra-GA Recommendation**: 同一論文内から GA に適した図を推薦するタスク. 2) **inter-GA Recommendation**: ある論文の GA 作成において, 参考となる他の論文の GA を推薦するタスク. 我々は, 各タスクに対し, 複数の異なる手法でベンチマーク評価を行った. また, Intra-GA Recommendation では明確に注釈された Ground Truth (GT) 以外にも妥当な候補が複数存在しうるため, 従来の順位ベースの評価指標ではこの状況を適切に扱うことができない. この課題に対処するため, 本研究ではモデルの確信度に着目した新たなオフライン推薦指標 **Confidence Adjusted top1-GT Ratio (CAR)** を導入する (図 1).

2. GA 設計支援のためのタスク定義

我々は, N 件の論文の集合 $\mathcal{D} = \{d^{(i)} \mid i \in 1, 2, \ldots, N\}$ を対象とする 2 つの推薦タスク Intra-GA Recommendation および Inter-GA Recommendation を定義する. 各論文 $d^{(i)}$ は, abstract $T^{(i)}$, GA $I^{(i)}_{\mathrm{GA}}$, $n^{(i)}$ 枚の図 $I^{(i)}_j \mid j \in 1, 2, \ldots, n^{(i)}$, キャプション $C^{(i)}_j$ などの要素から構成される.

Intra-GA Recommendation　Intra-GA Recommendation とは, 論文 $d^{(i)}$ 内の各図 $I^{(i)}_j$ が GA として適切かを評価し, 最も適した候補を推薦するタスクである. 検索候補の集合は $\mathcal{I}^{(i)}_{\mathrm{Intra}} = I^{(i)}_j \mid j \in \mathrm{GA}, 1, 2, \ldots, n^{(i)}$ と定義され, $I^{(i)}_{\mathrm{GA}}$ が GT となる. $I^{(i)}_{\mathrm{GA}}$ の多くは 図 1 に対応するが, この位置バイアスはモデルの本質的な GA 特徴抽出能を制限する. よって, 図の参照順に依存せず, 各候補を独立に評価する必要がある.

Inter-GA Recommendation　Inter-GA Recommendation とは, ある論文 $d^{(i)}$ の GA を作成する際, 他の論文の GA がデザインアイデアとしてどれほど関連しているかを評価し, 上位の候補を推薦するタスクである. 検索候補の集合は $\mathcal{I}^{(i)}_{\mathrm{Inter}} = I^{(i')}_{\mathrm{GA}} \mid i' \in 1, 2, \ldots, N, i' \neq i$ と定義される. 図の関連性が主観的なデザインの好みや文脈的な要因に依存するため, 本タスクに明示的な GT は存在しない.

[1] 法政大学大学院理工学研究科
[a)]takuro.kawada.3g@stu.hosei.ac.jp
[b)]shunsuke.kitada.0831@gmail.com
[c)]sota.nemoto.5s@gmail.com
[d)]iyatomi@hosei.ac.jp

[*1] コードは https://github.com/IyatomiLab/SciGA にて公開されている.

図 1: Intra-GA Recommendation において推薦された上位 4 件の図とモデルが各候補に対して持つ確信度, および CAR の例 [*2]. 黄色でハイライトされた図は GT (著者によって GA として選択された図) である. 左) CAR@k が高い場合: モデルが GT に対して強い確信をもって推薦している. 中央) CAR@k が中程度の場合: モデルが複数の候補が同程度に妥当であると判断し, 特定候補への確信は弱い. 右) CAR@k が低い場合: モデルが誤った候補に対して強い確信をもって推薦している.

3. ベンチマーク手法

GA Recommendation を様々な手法を比較評価するため, 様々な基準で定義されたレリバンスに基づいて図をランク付けし, 上位 k 件を推薦するベースラインモデル群を構築する. キャプションを用いる手法においては, 位置バイアスの影響を除去するため, キャプション内のタグ (e.g., *Figure 1*, *Figure 2*) を事前に除去する (セクション 2 参照).

(i) Abstract-to-Caption Lexical Matching (Abs2Cap) 各図 $I_j^{(i)}$ のキャプション $C_j^{(i)}$ と abstract $T^{(i)}$ のテキスト類似度をレリバンスとする手法である. この手法は, 山本らによる Most Important Figure 抽出手法 [27] に対応しており, Intra-GA Recommendation のベースラインとなる.

(ii) GA/non-GA Binary Classification (GA-BC) 我々は Intra-GA Recommendation を各図 $I_j^{(i)}$ が GA であるか否かを判別する二値分類問題として定式化する. 分類モデルはクロスエントロピー損失で学習され, 入力された図が GA である確率を出力する. 各図のレリバンスはその確率である. この手法は図の視覚的特徴のみを用いるため, 文脈横断的な Inter-GA Recommendation には本質的に適用できない.

(iii) Abstract-to-Figure Retrieval (Abs2Fig) テキストエンコーダ $f(\cdot)$ とイメージエンコーダ $g(\cdot)$ からなる対照学習モデルを用いて abstract $T^{(i)}$ と各図 $I_j^{(i)}$ を同一空間に投影し, コサイン類似度 $\rho(f(T^{(i)}), g(I_j^{(i)}))$ をレリバンスとする. 対照学習モデルは以下の対照損失 InfoNCE [24] で学習され, クエリ埋め込み $\boldsymbol{z}^{\mathrm{q}}$ と正例埋め込み $\boldsymbol{z}^{+}$ との類似度を最大化し, それ以外の負例埋め込み $\boldsymbol{z}_i^{-}$ との類似度を最小化する:

$$\mathcal{L}_{\mathrm{C}}(\boldsymbol{z}^{\mathrm{q}}, \boldsymbol{z}^{+}, \{\boldsymbol{z}_i^{-}\}) = -\log \frac{e^{\frac{\rho(\boldsymbol{z}^{\mathrm{q}}, \boldsymbol{z}^{+})}{\tau}}}{e^{\frac{\rho(\boldsymbol{z}^{\mathrm{q}}, \boldsymbol{z}^{+})}{\tau}} + \sum_i e^{\frac{\rho(\boldsymbol{z}^{\mathrm{q}}, \boldsymbol{z}_i^{-})}{\tau}}}. \quad (1)$$

ここで τ は温度パラメータである. Intra-GA Recommendation では, ミニバッチ $\mathcal{B} \subset \{1, 2, \ldots, N\}$ を用い, 以下の損失でモデルを学習する:

$$\mathcal{L}_{\mathrm{Intra}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{L}_{\mathrm{C}}(f(T^{(i)}), g(I_{\mathrm{GA}}^{(i)}), \{g(I_{\mathrm{j}\neq\mathrm{GA}}^{(i)})\}). \quad (2)$$

これにより, abstract は GA と強く結びつき, 非 GA から分離される. 図の枚数 $n^{(i)}$ は論文ごとに異なるため, 無作為抽出, あるいはゼロ埋めで m 枚に揃える. Inter-GA Recommendation では, 以下の損失でモデルを学習する:

$$\begin{aligned}\mathcal{L}_{\mathrm{Inter}} =& \frac{1}{2|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{L}_{\mathrm{C}}(f(T^{(i)}), g(I_{\mathrm{GA}}^{(i)}), \{g(I_{\mathrm{GA}}^{(i'\neq i)})\}) \\ &+ \frac{1}{2|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{L}_{\mathrm{C}}(g(I_{\mathrm{GA}}^{(i)}), f(T^{(i)}), \{f(T^{(i'\neq i)})\}). \quad (3)\end{aligned}$$

これにより, abstract は同一論文の GA と強く結びつき, 他の論文の GA から分離される.

(iv) Abstract-to-Figure Retrieval with Caption (Abs2Fig w/cap) 各図の埋め込み $g(I_j^{(i)})$ にキャプションの埋め込み $f(C_j^{(i)})$ を Hadamard 積で統合し, Abs2Fig を拡張する. つまり, $g(I_j^{(i)})$ を $g(I_j^{(i)}) \odot f(C_j^{(i)})$ に置換して $\mathcal{L}_{\mathrm{Intra}}$ および $\mathcal{L}_{\mathrm{Inter}}$ を再定義する. また, コサイン類似度 $\rho(f(T^{(i)}), g(I_j^{(i)}) \odot f(C_j^{(i)}))$ をレリバンスとする.

4. Confidence Adjusted top-1 GT Ratio (CAR)

Intra-GA Recommendation において, GT (著者によって GA として選択された図) 以外にも尤もらしい候補が複数存在しうる. モデルがこのような尤もらしい図を GT よりも僅か上位にランク付けしたとしても, それは本質的に誤

[*2] 左から順に:
(1) arXiv: 2403.17859, (2) arXiv: 2402.08210,
(3) arXiv: 2403.05721, (4) arXiv: 2403.12370,
(5) arXiv: 2402.09448, (6) arXiv: 2402.09434

表 1: Intra-GA Recommendation における各手法の定量的比較. 手法 (iii) Abs2Fig w/cap は全指標で優れた推薦能を示し, abstract およびキャプションからより豊かな文脈情報を捉えていることが確認できる. 各指標の最良値は**太字**で示している.

ベンチマーク手法	実装の詳細		R@1	R@2	R@3	MRR	CAR@5
	バックボーンモデル	最大トークン数					
(i) Abs2Cap	ROUGE-L [6]	-	0.394	0.625	0.759	0.601	0.429
	METEOR [2]	-	0.353	0.589	0.737	0.571	0.404
	CIDEr [25]	-	0.277	0.489	0.653	0.500	0.374
	BM25 [20]	-	0.508	0.739	0.849	0.690	0.528
	BERTScore [30]	512	0.485	0.707	0.819	0.668	0.505
(ii) GA-BC	EfficientNetV2 [23]	-	0.449	0.674	0.797	0.643	0.486
	ViT [7]	-	0.346	0.606	0.762	0.574	0.420
	CLIP Image Encoder [18]	-	0.493	0.708	0.826	0.675	0.518
	SwinTransformerV2 [13]	-	0.494	0.712	0.823	0.675	0.516
	ConvNeXtV2 [26]	-	0.483	0.703	0.816	0.667	0.511
(iii) Abs2Fig	CLIP [18]	77	0.573	0.791	0.877	0.735	0.573
	X^2-VLM [28]	40	0.489	0.711	0.825	0.672	0.514
	OpenCLIP [5]	77	0.566	0.780	0.870	0.730	0.567
	BLIP-2 [12]	512	0.578	0.787	0.867	0.737	0.577
	Long-CLIP [29]	248	0.575	0.783	0.877	0.735	0.573
(iv) Abs2Fig w/cap	CLIP [18]	77	0.628	0.822	0.902	0.771	0.610
	X^2-VLM [28]	40	0.538	0.757	0.857	0.709	0.546
	OpenCLIP [5]	77	0.621	0.817	0.905	0.767	0.603
	BLIP-2 [12]	512	0.557	0.767	0.863	0.721	0.557
	Long-CLIP [29]	248	**0.637**	**0.826**	**0.914**	**0.778**	**0.615**

りとは言えない. しかし, Recall@k (R@k) などの従来指標は, GT の順位だけに基づいて離散値を割り当てるため, こうした妥当な推薦を全て不正解として扱ってしまう. また, Normalized Discounted Cumulative Gain (nDCG) [4] は, 段階的なレリバンスラベルの下では有効だが, 正解が二値で与えられる場合は実質的に順位依存の評価に退化する.

この課題に対し, 我々は次式で定義される評価指標 Confidence Adjusted top-1 GT Ratio@k (CAR@k) を提案する:

$$\mathrm{CAR@}k = \frac{p_{\mathrm{GT}}}{p_{\text{top-1}}}\mathcal{C}(P,k). \tag{4}$$

ここで, $P \in \mathbb{R}^k$ はモデルが予測した上位 k 件の候補のレリバンスを z-score で標準化し, Softmax 関数で確率に変換したもの, $p_{\text{top-1}}, p_{\mathrm{GT}} \in P$ はそれぞれ最上位候補と GT の中で最も上位に位置する候補の確率を示す. CAR@k は, モデルの確信度 $\mathcal{C}(P,k)$ で調整された確率比 $p_{\mathrm{GT}}/p_{\text{top-1}}$ と定義され, 曖昧だが尤もらしい推薦結果を効果的に捕捉する.

$\mathcal{C}(P,k)$ は, 区間 [0.5, 1.0] で次のように定義される:

$$\mathcal{C}(P,k) = 1 - \frac{1}{2}\max\left(0, \frac{H(P)-h}{H_{\max}(P)-h}\right). \tag{5}$$

ここで, $H(P)$ は P のエントロピー, $H_{\max}(P) = \log k$ は候補数 k に対する最大エントロピー, $h = H_{\max}(P)/2$ は閾値を表す. $H(P) \le h$ の場合, $\mathcal{C}(P,k) = 1.0$ と定義し, モデルが高い確信を持っているとみなす. 逆に $H(P) > h$ の場合, 予測の不確かを反映し, $\mathcal{C}(P,k)$ は 0.5 まで線形減衰する.

つまり, CAR@k の振る舞いは次のように解釈できる. モデルが GT を自信をもってトップに推薦している場合 ($\mathcal{C}(P,k) \approx 1.0$ かつ $p_{\mathrm{GT}}/p_{\text{top-1}} \approx 1.0$), CAR@$k$ は 1.0 に近づく. モデルの確信は低いが GT に近い妥当な候補を推薦している場合 ($\mathcal{C}(P,k) \approx 0.5$ かつ $p_{\mathrm{GT}}/p_{\text{top-1}} \approx 1.0$), CAR@$k$ は 0.5 程度となる. モデルが強い確信を持ちながら GT 以外を推薦している場合 ($\mathcal{C}(P,k) \approx 1.0$ かつ $p_{\mathrm{GT}}/p_{\text{top-1}} \approx 0.0$), CAR@$k$ は 0.0 に近づく. GT が上位 k 件に含まれていない場合, CAR@k は明示的に 0.0 と定義される. このように CAR は, モデルがランキングにどれだけ確信を持っているか, その確信がどれだけ正当かの両方を反映した指標であり, インスタンスレベルで推薦の妥当性を評価する.

5. 評価実験

SciGA-145k [10] は, 科学論文の本文, 図, メタデータに加え, GA に対する明示的なアノテーションを含む論文データセットである. 本研究では, そのうち GA を含む 20,520 件の情報科学分野の論文を対象に, 学習・検証・テストを 8:1:1 の比率で分割し, GA Recommendation を実施した.

Intra-GA Recommendation セクション 3 で述べた 4 種の手法を用いた. Abs2Cap では, テキスト類似度として ROUGE-L [6], METEOR [2], CIDEr [25], BM25 [20], BERTScore [30] を採用した. GA-BC では, バックボーンモデルとして EfficientNetV2 [23], ViT [7], CLIP Image Encoder [18], SwinTransformerV2 [13], ConvNeXtV2 [26] を採用した. Abs2Fig および Abs2Fig w/cap では, バックボーンモデルとして CLIP [18], OpenCLIP [5], Long-CLIP [29], BLIP-2 [12], X^2-VLM [28] を採用した. 各手法の推薦能は, R@k, Mean Reciprocal Rank (MRR), CAR@5 で評価した.

表 2: Inter-GA Recommendation における各手法の定量的比較. Field-P@k は分野整合性, Abs2Abs SBERT@k は意味的類似度, GA2GA CLIP-S@k は視覚的類似度を示す. いずれもスコアが高いほど, 実際に著者によって作成の GA との推薦結果の類似度が高いことを示し, 標準偏差は推薦結果の多様性を反映する. 手法 (ii) GA-BC は Inter-GA Recommendation には適用できないため, 除外している. 各指標における最良値は**太字**, 最も高い標準偏差は下線で示している.

ベンチマーク手法	バックボーンモデル	Field-P@k		Abs2Abs SBERT@k		GA2GA CLIP-S@k	
		top-5	top-10	top-5	top-10	top-5	top-10
(BL) Random Sampling	-	0.338	0.345	0.227 ± 0.111	0.228 ± 0.115	0.545 ± 0.077	0.545 ± 0.081
(i) Abs2Cap	ROUGE-L [6]	0.502	0.486	0.314 ± 0.114	0.306 ± 0.118	0.579 ± 0.066	0.578 ± 0.069
	METEOR [2]	0.421	0.417	0.268 ± 0.110	0.264 ± 0.112	0.573 ± 0.063	0.571 ± 0.064
	CIDEr [25]	0.438	0.420	0.287 ± 0.105	0.273 ± 0.108	0.579 ± 0.064	0.577 ± 0.066
	BM25 [20]	0.704	0.685	0.489 ± 0.105	0.468 ± 0.111	0.605 ± 0.072	0.601 ± 0.074
	BERTScore [30]	0.549	0.545	0.360 ± 0.107	0.351 ± 0.109	0.580 ± 0.069	0.578 ± 0.071
(iii) Abs2Fig	CLIP [18]	0.729	0.719	0.455 ± 0.105	0.444 ± 0.109	0.646 ± 0.054	0.642 ± 0.057
	X^2-VLM [28]	0.418	0.402	0.263 ± 0.116	0.257 ± 0.122	0.461 ± 0.032	0.451 ± 0.033
	OpenCLIP [5]	0.720	0.710	0.451 ± 0.106	0.440 ± 0.109	0.632 ± 0.058	0.630 ± 0.061
	BLIP-2 [12]	0.683	0.674	0.419 ± 0.110	0.410 ± 0.114	0.622 ± 0.063	0.620 ± 0.065
	Long-CLIP [29]	0.726	0.717	0.456 ± 0.108	0.445 ± 0.103	$\mathbf{0.648 \pm 0.056}$	$\mathbf{0.644 \pm 0.060}$
(iv) Abs2Fig w/cap	CLIP [18]	**0.755**	**0.742**	0.493 ± 0.098	0.479 ± 0.101	0.614 ± 0.067	0.611 ± 0.071
	X^2-VLM [28]	0.415	0.399	0.254 ± 0.114	0.250 ± 0.119	0.555 ± 0.067	0.552 ± 0.072
	OpenCLIP [5]	0.749	0.737	0.489 ± 0.097	0.475 ± 0.100	0.615 ± 0.066	0.611 ± 0.069
	BLIP-2 [12]	0.647	0.639	0.390 ± 0.105	0.382 ± 0.109	0.597 ± 0.067	0.596 ± 0.068
	Long-CLIP [29]	0.753	0.737	$\mathbf{0.498 \pm 0.098}$	$\mathbf{0.482 \pm 0.103}$	0.614 ± 0.070	0.611 ± 0.073

Inter-GA Recommendation テストセットの abstract をクエリ, 学習セットの GA を検索候補とした. GA-BC を除き, Intra-GA Recommendation と同様の手法, モデルを用いた. また, Random Sampling (検索候補から k 件の GA を無作為抽出する手法) をベースライン (BL) とした. 推薦結果の品質を評価するため, 以下の 3 種類の指標を用いた: (1) Field-Precision@k (Field-P@k): クエリ論文と推薦された論文の研究分野の一致率. (2) Abstract-to-Abstract Sentence-BERT Similarity@k (Abs2Abs SBERT@k): クエリ論文と推薦された論文の abstract の Sentence-BERT [19] 埋め込み間のコサイン類似度に基づく, 意味的類似度. (3) GA-to-GA CLIPScore@k (GA2GA CLIP-S@k): クエリ論文と推薦された論文の GA 間の CLIPScore [9] に基づく, 視覚的類似度.

5.1 Intra-GA Recommendation の結果と分析

表 1 に示すように, (iii) Abs2Fig, (iv) Abs2Fig w/cap は (1)Abs2Cap, (ii) GA-BC より一貫して高い推薦能を示した. 特に, Abs2Fig w/cap は微細な図の違いを識別するためにキャプションの文脈情報が有効であることを実証した. 中でも Long-CLIP は最良の推薦能を示し, 長文対応 (248 トークン) による詳細な文脈理解が貢献していると考えられる. 一方, Abs2Fig で高性能だった BLIP-2 は Abs2Fig w/cap で性能が低下し, Q-Former がアライメントを乱す可能性が示唆された. また, 各インスタンスに割り当てられた CAR@5 の値に基づいて定性的評価を行った結果, モデルの誤りの傾向が明らかとなった. CAR@5 が中程度の場合, モデルが複数の尤もらしい候補の中で判断に迷っている様子が観察された. 一方, CAR@5 が 0.0 付近の場合, GT が背景情報の補足を示す図であることが多く, 正しい推薦が困難なケースであることが確認された. しかし, これらの推薦を誤ったケースにおいても, アーキテクチャ概要図など, 典型的な GA デザインの図が上位に推薦されていた. CAR は定性評価と整合的な連続値を示し, モデルの振る舞いや誤りの傾向を捉える上で有用な指標であることが確認された.

5.2 Inter-GA Recommendation の結果と分析

表 2 に示すように, 各手法で関連性と多様性の間にトレードオフが見られた. (BL) Random Sampling, (i) Abs2Cap は全指標においてスコアが低かったが, 標準偏差は大きく, 推薦された k 枚の GA の多様性が大きいことが示された. 一方, (iii) Abs2Fig, (iv) Abs2Fig w/cap は一貫して高いスコアを示した. 定性的な分析では, これらの手法が単に研究分野が同じ論文の GA を推薦するだけでなく, 自動運転, 医療言語処理, 対話システム, IoT, 音声処理といった, より細かなトピックレベルで関連性のある GA を推薦することが明らかとなった. 特に, Abs2Fig w/cap における CLIP は分野整合性に優れ, Long-CLIP は意味的類似度で最良のスコアを示した. Abs2Fig における Long-CLIP は視覚的類似度で最良のスコアを示したが, 多様性は低下した.

6. おわりに

本研究では, GA 設計支援を目的とした 2 つの推薦タスクと, 推薦指標 CAR を定義した. ベンチマーク実験により, キャプション統合型の対照学習と CAR の有効性が示された. 今後は, 意外性や研究者の嗜好といった要素を考慮したより柔軟なオンライン推薦, GA の自動生成を目指す.

参考文献

[1] Baek, J., Jauhar, S. K., Cucerzan, S. and Hwang, S. J.: ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models (2024). https://doi.org/10.48550/arXiv.2404.07738.

[2] Banerjee, S. and Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *ACL* (2005).

[3] Bennett, H. and Slattery, F.: Graphical abstracts are associated with greater Altmetric attention scores, but not citations, in sport science, *Scientometrics*, Vol. 128, pp. 3793–3804 (2023).

[4] Burges, C. J., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G.: Learning to rank using gradient descent (2005).

[5] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L. and Jitsev, J.: Reproducible scaling laws for contrastive language-image learning, *CVPR* (2023).

[6] Chin-Yew Lin, F. J. O.: Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, *ACL* (2004).

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *ICLR* (2021).

[8] Fu, T.-J., Wang, W. Y., McDuff, D. and Song, Y.: DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents, *AAAI* (2022).

[9] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L. and Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning (2021).

[10] Kawada, T., Nemoto, S., Kitada, S. and Iyatomi, H.: SciGA: 学術論文における Graphical Abstract 設計支援のための統合データセット，言語処理学会 第 31 回年次大会 発表論文集 (2025).

[11] Kunze, K. N., Vadhera, A., Purbeyc, R., Singh, H., Kazarian, G. S. and Chahla, J.: Infographics Are More Effective at Increasing Social Media Attention in Comparison With Original Research Articles: An Altmetrics-Based Analysis, *Canadian Journal of Emergency Medicine*, Vol. 37, No. 8, pp. 2591–2597 (2021).

[12] Li, J., Li, D., Savarese, S. and Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (2023).

[13] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F. and Guo, B.: Swin Transformer V2: Scaling Up Capacity and Resolution, *CVPR* (2022).

[14] Lu, C., Lu, C., Lange, R., Foerste, J., Clune, J. and Ha, D.: The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery (2024). https://doi.org/10.48550/arXiv.2408.06292.

[15] Meincke, L., Girotra, K., Nave, G., Terwiesch, C. and Ulrich, K. T.: Using Large Language Models for Idea Generation in Innovation, *SSRN Electronic Journal* (2023).

[16] Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G. and Cubuk, E. D.: Scaling deep learning for materials discovery, *Nature*, Vol. 624, pp. 80–85 (2023).

[17] Pang, W., Lin, K. Q., Jian, X., He, X. and Torr, P.: Paper2Poster: Towards Multimodal Poster Automation from Scientific Papers (2025). https://doi.org/10.48550/arXiv.2505.21497.

[18] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (2022).

[19] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (2019).

[20] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M.: Okapi at TREC-3, *TREC-3* (1994).

[21] Rodriguez, J. A., Vazquez, D., Laradji, I., Pedersoli, M. and Rodriguez, P.: FigGen: Text to Scientific Figure Generation, *ICLR* (2023).

[22] Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y. and Ceder, G.: An autonomous laboratory for the accelerated synthesis of novel materials, *Nature*, Vol. 624, pp. 86–91 (2023).

[23] Tan, M. and Le, Q. V.: Efficientnetv2: Smaller Models and Faster Training (2021).

[24] van den Oord, A., Li, Y. and Vinyals, O.: Representation Learning with Contrastive Predictive Coding (2018). https://doi.org/10.48550/arXiv.1807.03748.

[25] Vedantam, R., Zitnick, C. L. and Parikh, D.: CIDEr: Consensus-based Image Description Evaluation, *CVPR* (2015).

[26] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S. and Xie, S.: ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders, *CVPR* (2023).

[27] Yamamoto, S., Fukuhara, Y., Suzuki, R., Morishima, S. and Kataoka, H.: Automatic Paper Summary Generation from Visual and Textual Information, *ICMV* (2018).

[28] Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J. and Zhou, W.: X^2-VLM: All-in-One Pre-Trained Model for Vision-Language Tasks, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 46, No. 5, pp. 3156–3168 (2023).

[29] Zhang, B., Zhang, P., Dong, X., Zang, Y. and Wang, J.: Long-CLIP: Unlocking the Long-Text Capability of CLIP, *ECCV* (2024).

[30] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y.: BERTScore: Evaluating Text Generation with BERT, *ICLR* (2020).