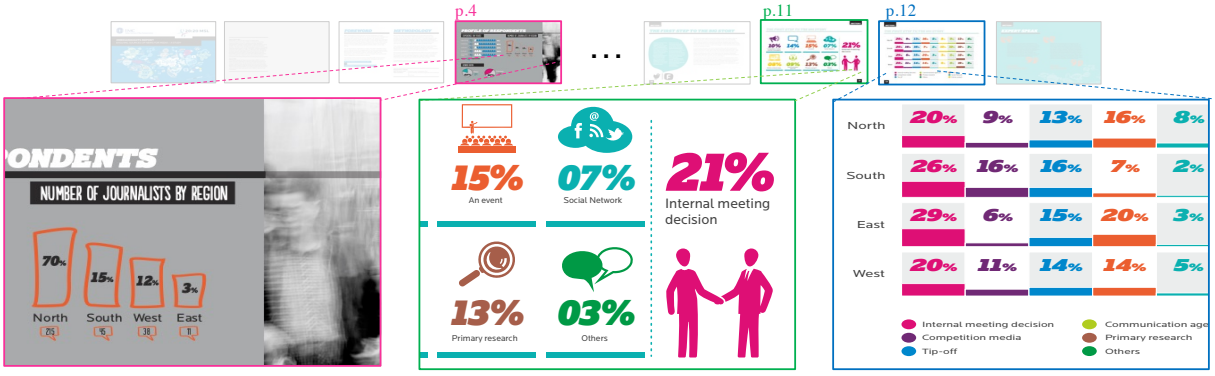
SlideVQA: 複数の文書画像に対する質問応答

田中涼太　西田京介　西田光甫　長谷川拓　斉藤いつみ　齋藤邦子
日本電信電話株式会社　NTT 人間情報研究所
{ryouta.tanaka.rg, kyosuke.nishida.rx, kosuke.nishida.ap, taku.hasegawa.ps
itsumi.saito.df, kuniko.saito.ku}@hco.ntt.co.jp



Q: What is the tip-off media percentage in the region with 70% of journalists and South?
A: 13%, 16%　**Evidence pages**: 4, 12
Answer type: Multi-Span　**Reasoning type**: Multi-hop

Q: What is the percentage of the internal meeting decision?
A: 21%　**Evidence pages**: 11
Answer type: Single-Span　**Reasoning type**: Sing-hop

Q: What is the difference in the competition media percentage between East and the region with 12% of journalists?
A: 5% (11% – 6%)　**Evidence pages**: 4, 12
Answer type: Non-Span　**Reasoning type**: Multi-hop, Numerical

図 1 SlideVQA の例．複数のスライド画像（スライドデッキ）を同時に理解し，質問に対して回答および根拠を出力する．

概要

スライドデッキに対する文書画像質問応答タスク SlideVQA を提案する．本タスクは複数の文書画像の関係性理解や数値推論などの複雑な推論能力を必要とする．さらに，回答根拠選択と質問応答を系列変換の枠組みで統一的に扱う新たな文書画像質問応答モデルを提案する．SlideVQA を用いた実験において，我々のモデルは既存の最先端モデルよりも優れた性能を示したが，依然として人間の性能に大きく劣っていることを確認した．本研究は，実世界に多数存在する視覚表現された文書を知識源として質問応答を行う人工知能の発展に貢献できる．

1　はじめに

文書を知識源とし質問に対して人間の様に回答を行う技術の実現は，AI 分野における重要課題の一つである．テキストのみで記述された文書に関する質問応答 [1, 2] では，実サービスで扱われる文書が持つテキスト情報しか扱えないことから，文書を画像として扱い質問応答を行う文書画像質問応答（DocumentVQA）[3, 4, 5, 6] が近年注目を集めている．ここで，従来のデータセットは一枚の文書画像を対象としているため，複数枚の文書内容を参照し理解するモデルは未だ実現されていない．さらに，従来モデルは視覚要素（図表など）を含む文書理解や算術演算の能力が低いことが指摘されている [5]．

本研究では，複数の文書画像で構成されるスライドデッキを読み解き，質問に対して回答および回答根拠を出力する新たなタスクおよびデータセット **SlideVQA**[1] を提案する．図 1 に示す様に，SlideVQA はマルチホップ推論，数値推論などの複雑な推論能力を必要とする．そして，視覚要素理解や算術演算の能力強化に向けた意味領域および算術式のアノテーションを提供する．

さらに本研究では，テキスト・レイアウト・視覚要素を考慮し，複数の文書画像の内容を同時理解可能な新たなモデル **M**ulti-**M**odal **M**ulti-image **D**ocument VQA model (**M3D**) を提案する．SlideVQA を用いた実験において，M3D は既存の最先端モデルよりも優れた性能を示したが，依然として人間の性能に大きく劣っており，SlideVQA が新たな未解決課題であることを確認した．

1）データセットやさらなる詳細情報は https://github.com/nttmdlab-nlp/SlideVQA および [7] で公開している．

2 SlideVQA

2.1 タスク定義

SlideVQA タスクを以下の通りに定義する．

End-to-End SlideVQA 質問 q とスライドデッキ（画像集合）$\mathbf{I} = \{I_1, \ldots, I_K\}$ $(K = 20)$ が与えられ，モデルは回答 y と関連するスライド $\hat{\mathbf{I}} = \{\hat{I}_1, \ldots, \hat{I}_{K'}\}$ を出力する．

更に，以下の二つのサブタスクに分解できる．

回答根拠選択 質問 q とスライドデッキ $\mathbf{I}$ が与えられ，モデルは回答 y を出力するために必要な画像集合 $\hat{\mathbf{I}}$ を選択する．

質問応答 質問 q とスライドデッキ ($\mathbf{I}$ 又は $\hat{\mathbf{I}}$) が与えられ，モデルは回答 y を出力する．

図 1 に示す様に，回答は画像中のテキストから一つのスパン（Single-Span），複数のスパンから抜き出す回答（Multi-Span），数値推論や見た目を問う回答（Non-Span）の 3 つのタイプに大別できる．また，数値推論を必要とする回答は $\{+, -, /, *\}$ を用いた算術演算，カウント，比較の何れかを用いる．

2.2 データ収集

スライドデッキ収集 20 ページ以上で構成されており，図表を含むスライドデッキ（英語）を対象に，SlideShare [8] から 2,619 件収集した．更に，収集したデッキの 21 ページ以降を切り捨てた．

意味領域アノテーション SPaSe [9] が定義する 9 つの意味ラベル（例: Title，Table）を用いて全画像に意味領域の特定および意味ラベルの付与を行った．

シングルホップ QA 作成 スライドデッキの中から一枚のスライドを選択し，選択したスライドに関する QA ペアを 12,466 件作成した（例：図 1 中央）．この選択スライドは回答根拠として使用される．

マルチホップ QA 作成 シングルホップ QA を編集し，マルチホップ QA を 2,018 件作成した．まず，シングルホップ質問から最大 2 件の固有表現を抽出する．次に，抽出した固有表現について説明している別のページを回答根拠として選択し，固有表現を別の表現に置換する．図 1 左では，“North”が “the region with 70% of journals”に置換されている．編集によって作成される質問の中には，非流暢な質問も含まれる．しかし，本作成手法はマルチホップ推論の保証およびデータサイズの拡張性の面で利点がある．類似手法として Wikipedia におけるマルチホップ推論を必要とする MultimodalQA [10] では，編集による QA 作成の有効性が確認されている．

算術式アノテーション 算術演算が必要な回答に対して，算術式 (例: “30 - 28”) を付与した．

表 1 データセットの比較．MI は複数文書入力，MR はマルチホップ推論，NR は数値推論，I は画像，SR は意味領域，AE は算術式を表す．

Dataset	MI	MR	NR	#QAs	#Is	#SRs	#AEs
DocVQA [3]				50k	12k	–	–
VisualMRC [4]				30k	10k	64k	–
InfographicVQA [5]			✓	30k	5k	–	–
DocCVQA [6]	✓			0.02k	14k	–	–
SlideVQA	✓	✓	✓	14.5k	52k	890k	1.7k

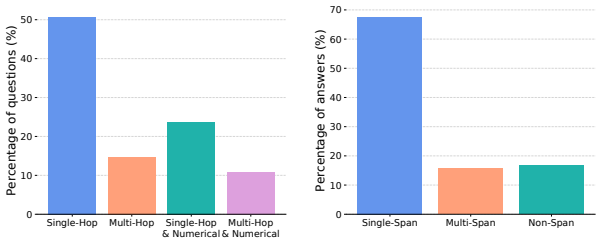


図 2 左: 推論タイプの分布．**右**: 回答タイプの分布．

2.3 統計情報および従来研究との比較

SlideVQA は 2,619 件のスライドデッキ (52,480 画像，890,945 件の意味領域) に関する 14,484 件の QA ペアを含む．また，各デッキが同一の分割のみに存在するように，訓練/開発/テストデータを 10,617/1,652/2,215 質問に分割した．

画像 表 1 で示す様に，SlideVQA は最も多くの文書画像を含むデータセットである．また，VisualMRC [4] と比べて，SlideVQA は意味領域数が 14.7 倍であり，最多である．SlideVQA タスクにおけるモデルが考慮すべき OCR 単語平均数は 1488.88 単語であり，従来研究で最も OCR 単語数の多い InfographicVQA [5] の 217.89 単語を大きく上回る．

QA 表 1 で示す様に，SlideVQA はマルチホップ推論と数値推論を必要とし，算術式付与を行った初めての研究である．また，SlideVQA は複数文書画像を入力とし，学習に資するデータ量を含む点でも初めての研究である（DocCVQA [6] は QA 数が 20 件に限られる）．図 2 左で示す様に，49.3%の質問がマルチホップ推論または数値推論を必要とする．図 2 右で示す様に，Multi-Span と Non-Span の回答が 32.4%を占めており，回答の生成・抽出能力を必要とする．

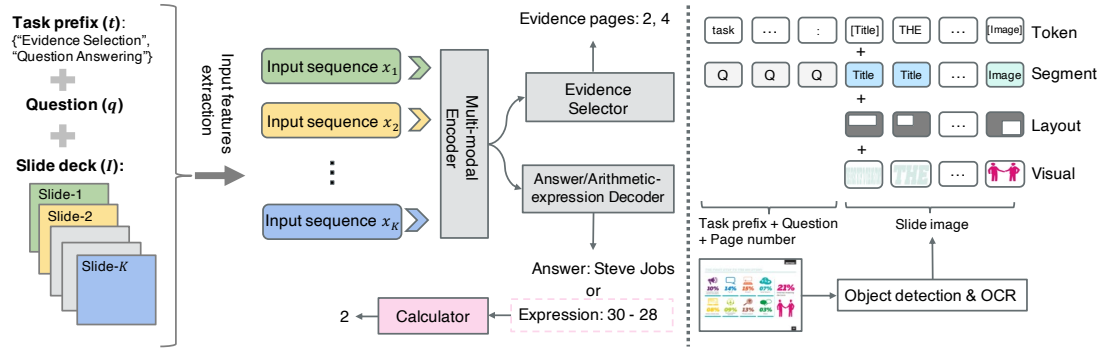


図 3 左: 提案モデル M3D のアーキテクチャ．右: 入力埋め込み系列．

3 提案モデル

提案モデル M3D を図 3 に示す．M3D は T5 [11] からなる Transformer Encoder-Decoder モデル FiD [12] をベースとする．M3D の貢献は，1) マルチモーダル情報を考慮して複数の文書画像を同時に理解できる点，2) 算術演算の能力強化を目的とした算術式の生成を行う点，3) 回答根拠選択と質問応答を系列変換の枠組みでマルチタスク学習を行う点である．

3.1 マルチモーダル入力埋め込み層

入力トークン系列 ページ k における入力トークン系列 x_k を，質問 q，Task Prefix t，ページ番号 e_k，コンテキストトークン c_k を用いて定義する．

$$x_k = (\texttt{task:}t\ \ \texttt{question:}q\ \ \texttt{page:}e_k\ \ \texttt{context:}c_k)$$

Task Prefix $t \in \{\texttt{Evidence Selection}, \texttt{Question Answering}\}$ であり，回答根拠選択と質問応答の同時学習を可能とする．コンテキストトークン c_k はページ k 中の意味領域毎の OCR トークン系列 $\mathbf{W}_k^{r_i}$ および意味ラベル $[\mathrm{R}_k^{r_i}]$ から定義される．

$$c_k = ([\mathrm{R}_k^{r_1}], \mathbf{W}_k^{r_1}, [\mathrm{R}_k^{r_2}], \mathbf{W}_k^{r_2}, \ldots, [\mathrm{R}_k^{r_N}], \mathbf{W}_k^{r_N}),$$

意味領域 (例: Title) は，画像 I_k から Faster-RCNN [13] を用いて N 個抽出する．

入力埋め込み系列 入力埋め込み系列 $\mathbf{z}$ をトークン $\mathbf{z}^{\text{token}}$，セグメント $\mathbf{z}^{\text{seg}}$，レイアウト $\mathbf{z}^{\text{lay}}$，視覚特徴埋め込み系列 $\mathbf{z}^{\text{vis}}$ を用いて以下の通りに定義する．

$$\mathbf{z} = \mathrm{LN}(\mathbf{z}^{\text{token}} + \mathbf{z}^{\text{seg}} + \mathbf{z}^{\text{lay}} + \mathbf{z}^{\text{vis}}) \in \mathbb{R}^{L\times d},$$

LN は Layer Normalization [14]，L は入力長を表す．$\mathbf{z}^{\text{token}}$ は入力トークン系列 x_k を d 次元に埋め込む．また，$\mathbf{z}^{\text{seg}}$ は各トークンが属する意味ラベルを表す．意味領域と OCR 領域に対して，1 層の FFN に入力した結果を $\mathbf{z}^{\text{lay}}$ とし，Faster-RCNN の出力を ReLU 活性化関数と 1 層の FFN に渡した結果を $\mathbf{z}^{\text{vis}}$ と表す．

3.2 マルチモーダル Endoder-Decoder

マルチモーダル Encoder FiD [12] と同様に，K 個の画像に対応する入力トークン系列 x_k を独立にエンコード後，エンコード結果 $\mathbf{x}_k \in \mathbb{R}^{L\times d}$ を結合し，$\mathbf{X} \in \mathbb{R}^{KL\times d}$ を得る．

回答/算術式 Decoder Decoder は通常回答（例: “Steve Jobs”）または算術式（例: “30 - 28”）を出力する．算術演算の必要性はモデルが判断する．出力が算術式の場合は出力系列の先頭に “Expression:”，通常の回答の場合は “Answer:”を予測する．

回答根拠選択器 Decoder と同じ構造を持ちパラメータを共有する．$\hat{e}$ を回答根拠のページ番号とした時，出力形式を $\hat{\mathbf{I}}_{\text{pages}} = (\texttt{Evidence pages:}\ \hat{e}_1, \ldots, \hat{e}_{K'})$ として，系列生成により回答根拠を選択する．

学習と推論 選択器と Decoder に関する負の対数尤度損失の和を最小化する．推論時には，出力系列の先頭トークンを削除し，最終出力とする．また，“Expression:” が出力された場合，後処理として演算を実施した（例: “30 - 28”から回答 “2”を算出）．

4 評価実験

ベースライン End-to-End/質問応答タスクにおける比較手法として，根拠選択モデルと質問応答モデルで構成されるパイプラインを用いる．根拠選択モデルとして HLayoutLMv2 を新たに導入した．HLayoutLMv2 は各画像を LayoutLMv2 [15] に渡して得られた表現を Transformer [19] によりエンコードし，上位 3 件の根拠を出力する．他に，各モーダル情報を利用した LayoutLM [18] を含む 7 つのモデルを用いた（表 2 参照）．質問応答モデルには，回答生成を行う T5 と LayoutT5 [4]，回答スパン抽出を行う LayoutLMv2，M3D のベースとなる FiD を用いた．

評価指標 HotpotQA [20] に倣い，質問応答と根拠選択の各サブタスクで Exaxt-Match (EM) と F1 を用

表 2 SlideVQA タスクにおける評価結果．(b)，(c) は SlideVQA のサブタスクである．T (テキスト)，L (レイアウト)，V (画像特徴) を表す． $M3D_{GT}$ は正解根拠のみをモデルに入力した． $\mathbf{z}^{lay}$ は入力埋め込みに対するレイアウト埋め込みを表す．

(a) End-to-End タスクの結果．

Model	Modal	JEM	JF1
T5 [11]	T	22.6	34.2
T5 + $\mathbf{z}^{lay}$ [11]	TL	23.6	35.7
LayoutT5 [4]	TLV	24.3	36.1
LayoutLMv2 [15]	TLV	16.5	26.5
M3D	TLV	**28.0**	**37.3**
$M3D_{GT}$	TLV	35.4	44.7
Human	–	88.6	91.9

(b) サブタスク: 回答根拠選択の結果．

Model	Modal	EM	F1
CLIP [16]	V	39.3	43.5
BERT [17]	T	50.3	69.2
BERT + $\mathbf{z}^{lay}$ [17]	TL	52.7	71.0
LayoutLM [18]	TL	42.0	59.9
LayoutLMv2 [15]	TLV	51.7	71.5
HLayoutLMv2 [15]	TLV	69.8	**85.6**
M3D	TLV	**75.0**	83.8
Human	–	97.7	98.0

(c) サブタスク: 質問応答の結果．

Model	Modal	EM	F1
T5 [11]	T	29.3	37.9
T5 + $\mathbf{z}^{lay}$ [11]	TL	31.0	39.7
LayoutT5 [4]	TLV	31.7	39.9
LayoutLMv2 [15]	TLV	21.4	29.3
FiD [12]	T	30.4	38.9
FiD + $\mathbf{z}^{lay}$ [12]	TL	30.6	38.9
M3D	TLV	**33.5**	**41.7**
Human	–	89.8	93.0

表 3 開発データにおける M3D の ablation 評価．

Model	E2E		Select		QA	
	JEM	JF1	EM	F1	EM	F1
M3D	**36.2**	**42.8**	**83.1**	**87.7**	**41.3**	**47.1**
↪ BinaryClass	24.7	34.8	54.5	68.5	38.8	44.8
w/o AE generation	35.7	42.3	82.9	87.7	40.5	46.3
w/o Evidence selection	–	–	–	–	40.6	46.4
w/o Layout features	35.1	42.0	82.4	87.1	40.3	46.3
w/o Visual features	34.2	40.9	81.5	86.3	39.0	44.9

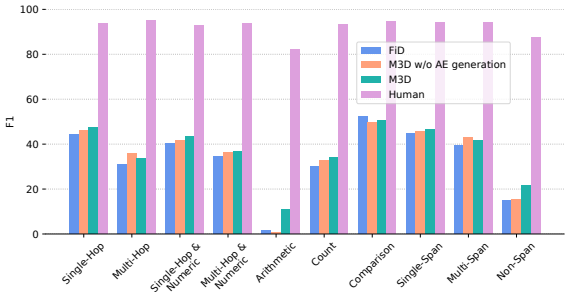


図 4 カテゴリごとの QA 性能比較．

い，Joint-EM/F1（JEM/JF1) を用いて End-to-End タスクにおける回答と回答根拠の一貫性を評価する．

4.1 評価結果と分析

提案モデルはベースラインの性能を上回るか？ 表 2a に示す様に，M3D は全てのベースラインの性能を上回った．回答根拠選択タスクでは，表 2b に示す様に，HLayoutLMv2 と M3D がその他のベースラインよりも高い性能である．これは全ての根拠候補を同時参照するモデル化が性能向上において重要であることを示唆する．質問応答タスクでは，表 2c に示す様に，M3D がベースラインの性能を上回った．これは M3D がパイプラインよりも質問に関連しない画像を排除し，回答生成できることを示唆する．$M3D_{GT}$ は正解の根拠を与えることで，大きな性能向上が確認できることから，根拠選択の性能改善の余地があることを示唆する．

提案モデルは人間の性能を上回るか？ 表 2，図 4 に示す様に，全てのタスクとカテゴリにおいて人間が大きく提案モデルの性能を上回っている．特に，視覚理解を伴う 1）マルチホップ推論，2）算術演算における性能向上は今後の課題である．

データセットの特徴は？ 表 2 に示す様に，モーダル情報の増加に伴い性能向上が全てのタスクにおいて確認できることから，SlideVQA は文書のテキスト・レイアウト・視覚情報を全て考慮する必要がある．表 2a・2c に示す様に，回答生成を行う LayoutT5 が回答スパン抽出を行う LayoutLMv2 の性能を大きく上回った．これは，Non-Span 回答を含むデータセット [21] で同様に確認される現象である．

性能向上に寄与する要素は何か？ 表 3 に示す様に，各要素を取り除くことで性能低下が確認できる．特に，根拠選択を各画像の先頭トークンに対応するエンコード表現に対して 2 層の MLP を用いた識別モデル BinaryClass に変更することで，大きな性能低下が確認できる．これは，Decoder と選択器とのパラメータ共有および系列生成としてのマルチタスク学習が性能向上に寄与していると考える．また，図 4 に示す様に，数値回答の代わりに算術式を生成すること（AE generation）により，算術演算に関する F1 において 10.4%の向上が確認できる．

5 おわりに

新たな文書画像質問応答タスク SlideVQA を提起し，データセットおよびモデルの構築を行った．SlideVQA は最高性能のモデルにおいても人間の性能に及ばない難関なタスクである．本データセットにより，実世界に多数存在する視覚表現された文書を基に QA を行う技術や，Web 検索や対話システムなど産業上重要なサービスの発展に貢献できる．

参考文献

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **EMNLP**, pp. 2383–2392, 2016.

[2] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In **ACL**, pp. 784–789, 2018.

[3] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In **WACV**, pp. 2200–2209, 2021.

[4] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In **AAAI**, pp. 13878–13888, 2021.

[5] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In **WACV**, pp. 1697–1706, 2022.

[6] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In **ICADR**, pp. 778–792, 2021.

[7] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. A dataset for document visual question answering on multiple images. In **AAAI**, 2023.

[8] Slideshare. https://www.slideshare.net/.

[9] Ziad Al-Halah Monica Haurilet and Rainer Stiefelhagen. SPaSe - Multi-Label Page Segmentation for Presentation Slides. In **WACV**, pp. 726–734, 2019.

[10] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: complex question answering over text, tables and images. In **ICLR**, 2021.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **JMLR**, Vol. 21, No. 140, pp. 1–67, 2020.

[12] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In **EACL**, pp. 874–880, 2021.

[13] Ren, Shaoqing, He, Kaiming, Girshicka, Ross, Sun, and Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. In **NIPS**, pp. 91–99, 2015.

[14] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. **arXiv:1607.06450**, 2016.

[15] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In **ACL/IJCNLP**, pp. 2579–2591, 2021.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **ICML**, pp. 8748–8763, 2021.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT**, pp. 4171–4186, 2019.

[18] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In **KDD**, pp. 1192–1200, 2020.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS**, pp. 6000–6010, 2017.

[20] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In **EMNLP**, pp. 2369–2380, 2018.

[21] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In **ACL**, pp. 2368–2378, 2019.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. **arXiv:1711.05101**, 2017.

[23] Google clould vision api. https://cloud.google.com/vision.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **CVPR**, pp. 770–778, 2016.

[25] Sebastian Ruder. An overview of gradient descent optimization algorithms. **arXiv:1609.04747**, 2016.

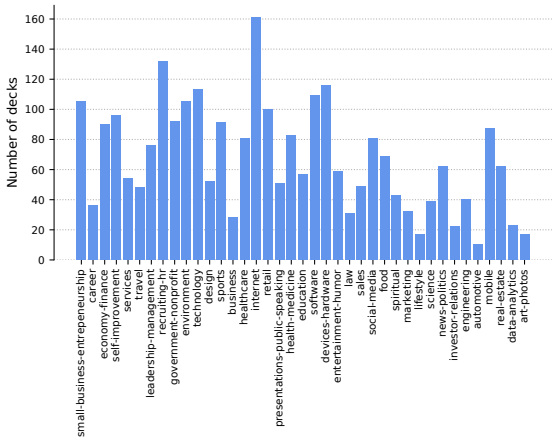


図 5 収集したスライドデッキのカテゴリ．

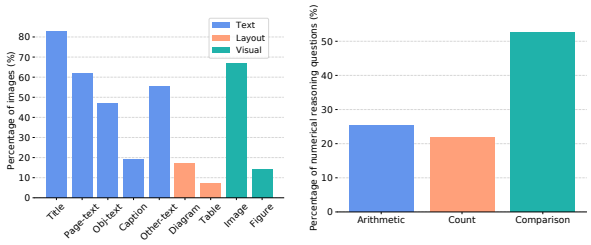


図 6 **左**: 意味領域の分布．**右**: 数値推論タイプの分布．

A　付録

スライドデッキのカテゴリ　図 5 に示す様に，収集したデッキは 39 個の SlideShare のカテゴリを網羅する．

意味領域アノテーション　図 7 に，意味領域アノテーション例を示す．下記にクラウドワーカーの指示に用いた意味ラベルの定義を示す．

- **Title**: presentation title, slide title
- **Page-text**: text in slide, bullet-point text list, text list
- **Obj-text**: text in a figure, image, diagram or table
- **Caption**: description of figure, image, diagram, or table
- **Other-text**: footnote, date, affiliation, code, URL
- **Diagram**: a graphical representation of data, a process
- **Table**: data arranged in rows and columns
- **Image**: drawing, logo, map, screenshot, realistic image
- **Figure**: graph with data points and coordinates

データ分析　図 6 左で示す様に，SlideVQA は 9 つの意味領域を広くカバーしている．一方で， DocVQA [3] や DocCVQA [6] は Visual (Image と Figure) を含む文書を対象としていない．図 2 右で示す様に，25.5%が算術を必要とする質問である．

実装　Pytorch を用いて実装し，8 台の Tesla V100 で実験した．CLIP のモデルサイズは Large 用い，その他のモデルは Base を用いた．AdamW [22] を用いて最適化し学習率は 5e-5 とした．バッチサイズは 32 とした．開発データで最も損失の小さいモデルを最終評価に用いた．入力トークン系列の最大長を 200，出力系列の最大長を 50 とした．OCR には Google Could Vision API [23] を用いた．Faster-RCNN のバックボーンは ResNet-101 [24] を使用した．また，SGD [25] を用いて最適化に用い，バッチサイズを 1，学習率 1e-3 とし，5 epoch 学習を行った．アンカーボックスのスケールを [8, 16, 32]，アスペクト比を [0.5, 1.0, 2.0] とした．

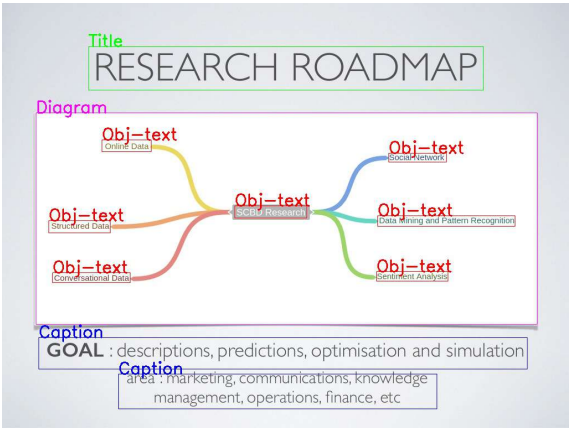
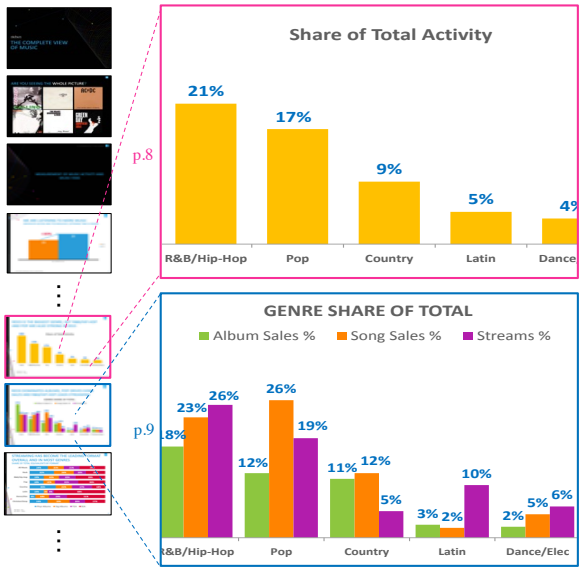


図 7 意味領域のアノテーション例．



Q: What is the combined percentage of Album Sales % and Song Sales % for the genre with a 9% Share of Total Activity?

GT	**answer**: 23%	**evidence pages**: 8, 9
FiD	**answer**: 57%	**evidence pages**: None
LayoutT5	**answer**: 68%	**evidence pages**: 8, 9
M3D	**answer**: 23% (11% + 12%)	**evidence pages**: 8, 9

図 8 出力例．(·) は生成された算術式を表す．

出力例　出力例を図 8 に示す．本例では，マルチホップ推論と算術操作を伴う回答生成が必要な例である．FiD は文書レイアウトを理解できておらず，誤った回答をしている．LayoutT5 は算術の過程を正しく理解できずに誤った回答を出力しているのに対して，提案モデル M3D は正しく情報（“11%”と “12%”）を抽出し，ground-truth と同じ回答を生成することができている．

URL 一覧　図 1・7・8 の画像ソースを以下に示す．

図 1　https://www.slideshare.net/mslgroup/mediainsights-evolving-sources-of-news-for-media
図 7　https://www.slideshare.net/andrybrewok/big-data-analytics-a-social-network-approach
図 8　https://www.slideshare.net/musicbizassoc/nielsen-2015-music-biz-presentation-final