

NLPとVision-and-Languageの 基礎・最新動向(2)

NTT人間情報研究所

西田京介，西田光甫，田中涼太，斎藤いつみ

2023年3月8日

DEIM2023 第15回データ工学と情報マネジメントに関するフォーラム
チュートリアル講演

講師紹介



西田京介

専門：機械読解（質問応答）, Vision-and-Language

NLP'18 最優秀賞, '19優秀賞(筆頭), '21最優秀賞, '20'22優秀賞(共著), DBSJ上林
奨励賞(2017)など. MS MARCO投稿時首位獲得. JSAI'22, NLP'19など 招待講演.



西田光甫

専門：自然言語理解・説明

NLP'22 優秀賞, NLP'20 若手奨励賞(筆頭), HotpotQA, FEVERにて投稿時首位獲得.



田中涼太

専門：視覚的機械読解

NLP'21最優秀賞, NLP'22若手奨励賞, ICDAR 2021 Infographics VQA runners-up
(筆頭), AI王ワークショップ'21 招待講演.



斎藤いつみ

(※本日欠席)

専門：文書要約

JSAI'20優秀賞, NLP'19最優秀ポスター(筆頭), 言語処理セミナー'20 招待講演.

目次



西田京介

1. 深層学習による自然言語処理技術の基礎 (40min)

タスクとモデルの分類, Transformerとセルフアテンション, BERTとGPT-3, 大規模言語モデル（LLM）に関する知見



西田光甫

2. 大規模言語モデルの革新 (40min)

InstructionチューニングとRLHF, ChatGPT, ChatGPTの評価と言語モデルの今後の課題

質疑応答 (10min)



西田京介

3. Vision-and-Languageへの派生 (40min)

Vision Transformer, 視覚・言語の基盤モデルCLIP, CLIP+LLMによるモデル：Flamingo, KOSMOS-1など



田中涼太

4. 文書や画面の理解 (40min)

処理・タスク・モデルの紹介, 文書画像理解の発展：SlideVQA, LiLT, Pix2Struct, ReActなど

質疑応答 (10min)



自然言語処理から Vision-and-Languageへ

Vision-and-Languageとは

- 画像処理と自然言語処理の融合領域
- TransformerやBERTの成功が、視覚と言語を結びつけた理解にも派生し、急速に発展している

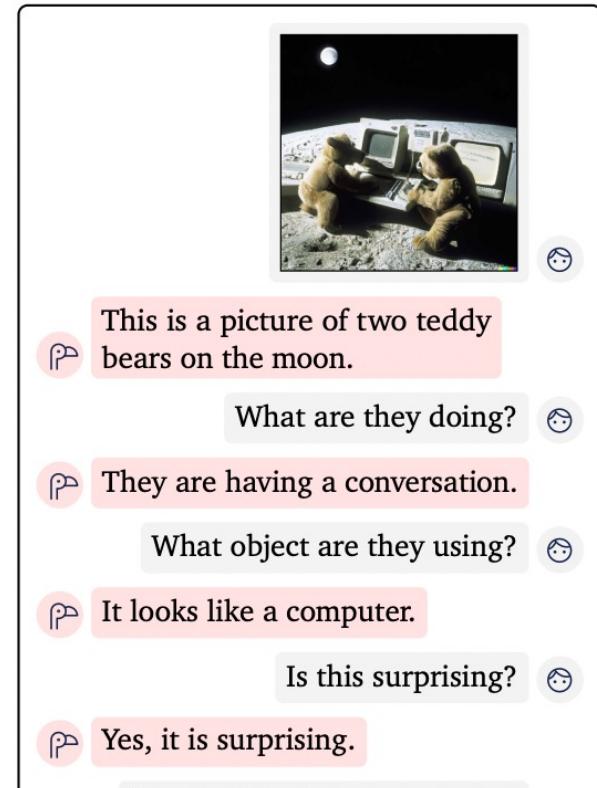
“DALL-E 2”によりテキストから生成された画像
[Ramesh(OpenAI)+, 2022/04/13]



vibrant portrait painting of Salvador Dalí with a robotic half face

<https://cdn.openai.com/papers/dall-e-2.pdf>

“Flamingo”による画像の内容に基づく対話
[Alayrac(Deepmind)+, 2022/04/29]



<https://arxiv.org/abs/2204.14198>

Vision-and-Languageの主なタスク

- 自然言語処理と同様に、**理解／生成**で大別可能

視覚・言語の融合理解

画像・言語のマッチング（検索）



+ チョコレート
シロップが掛けた
ワッフル2つとアイス
→ Score: 0.98

VQA（回答選択型）



+ ベッドの上に
子供は何人いますか？
→ 1 / 2 / 3 / ...

視覚・言語の融合理解に基づく生成

Document VQA（回答生成型）



+ ソーシャルメディア
の利用率は？
→ 16% (100-84)

ロボットナビゲーション



+ 階段を登り、次に
ピアノの横を...
→ forward,
left, ..., stop

Image-to-text / Text-to-image



ベレー帽と
タートルネック
を着た柴犬

(初期の) V&L事前学習方法の概要

1. Webなどから大量の画像とキャプションのペアを収集
2. 画像を系列データとして捉えて、テキストの単語系列と併せてTransformerエンコーダに入力
3. 穴埋め・マッチングタスクなどで事前学習



画像を系列データとして扱う

- Faster R-CNN [Ren+, NIPS15]などの物体検出技術により画像からオブジェクトを検出し画像からオブジェクト（およびそのベクトル表現）を検出し、1つの系列に並べる
- VQAなどのタスクではオブジェクトの種類や個数が問われる事が多いため、初期のV&Lモデルではこのアプローチが一般的であった



画像



N個のオブジェクトの特徴ベクトル系列

$$\mathbb{R}^{N \times d}$$

事前学習タスク: Masked Language/Region Modeling

- BERTと同様に単語をマスクして穴埋め問題を解く
- 画像側のトークンをマスクした穴埋め問題（特徴量の再現やオブジェクトのクラス分類）を解く場合もあり



事前学習タスク: Image-Text Matching

- データセットの中にあるペア（正例）とランダムサンプリングで作った負例のペアを見分けるタスク
- このタスクも大量にデータを用意できる

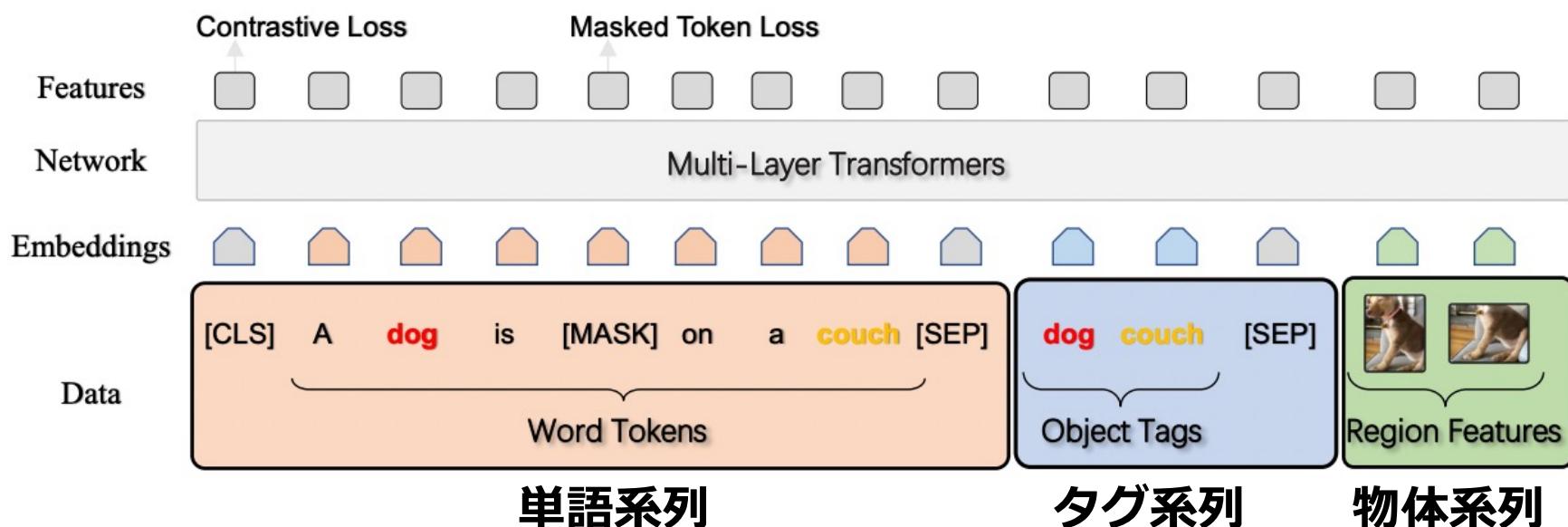
Matching/ Not-matching



VinVL [Zhang(Microsoft)+, CVPR'21]

- 物体検出を1848クラス/524属性まで強化した事前学習モデル
- さらに、検出された物体の名前（タグ）を同時に入力することで、言語・画像を結びつけた事前学習を強化

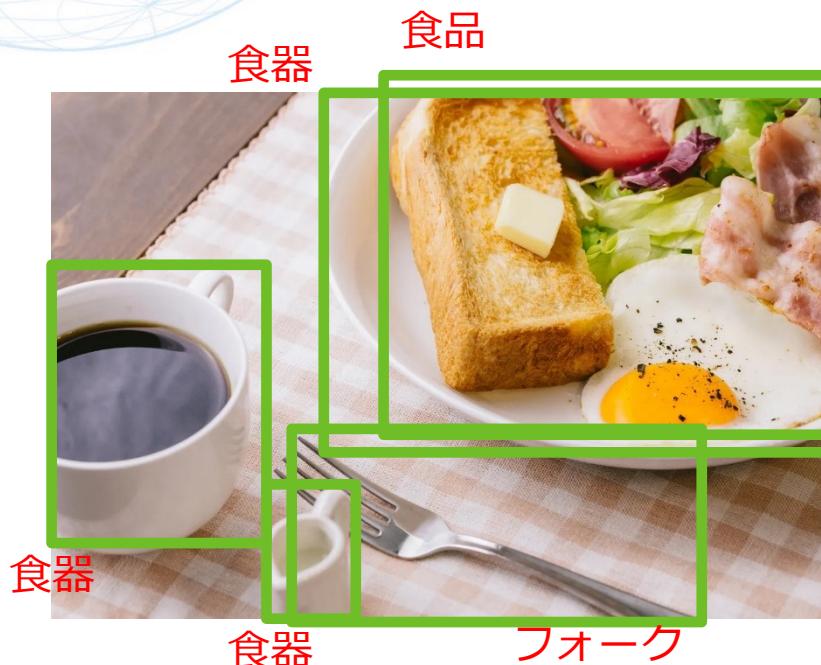
マッチングタスク トークン穴埋め



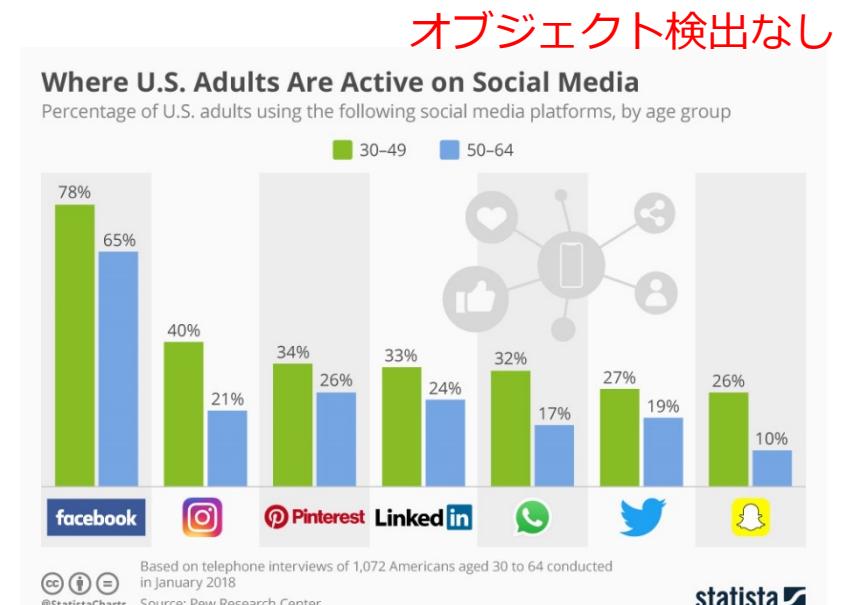
物体検出による画像表現抽出の限界

- 物体検出器がカバーする概念の広さと検出精度にV&Lモデルの性能が大きく依存してしまう
→ 物体検出に依存しない、汎用の画像エンコーダが得られないか？

商用の物体検出APIの結果



食品や食器の細かい違いが認識できていない
マットやコーヒーについて認識漏れ

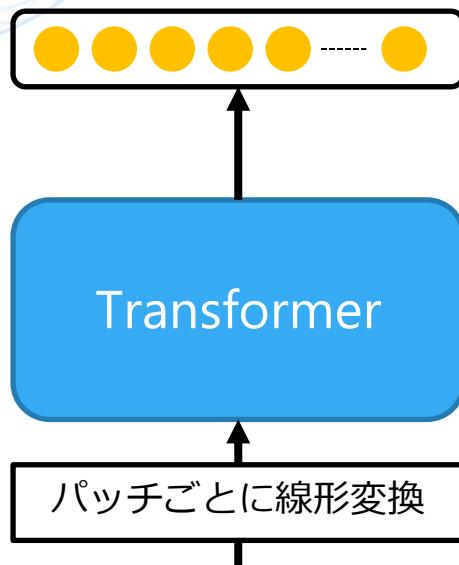


各種Webサービスのアイコンを
認識できず

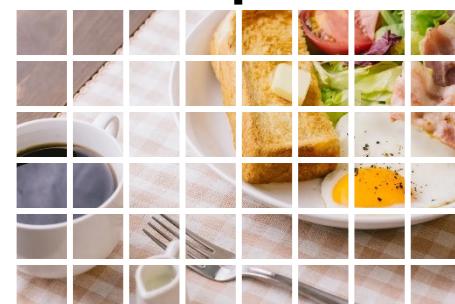
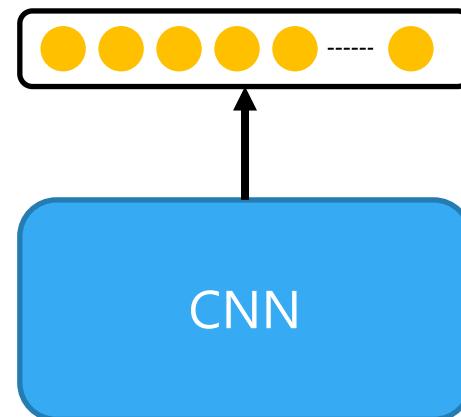
パッチ／グリッド方式による系列化

- ・ パッチ：画像を小さい領域に分割して埋め込みトークン化
- ・ グリッド：画像をそのまま入力しグリッド毎の特徴量を出力

画像特徴（パッチ数）



画像特徴（グリッド数）



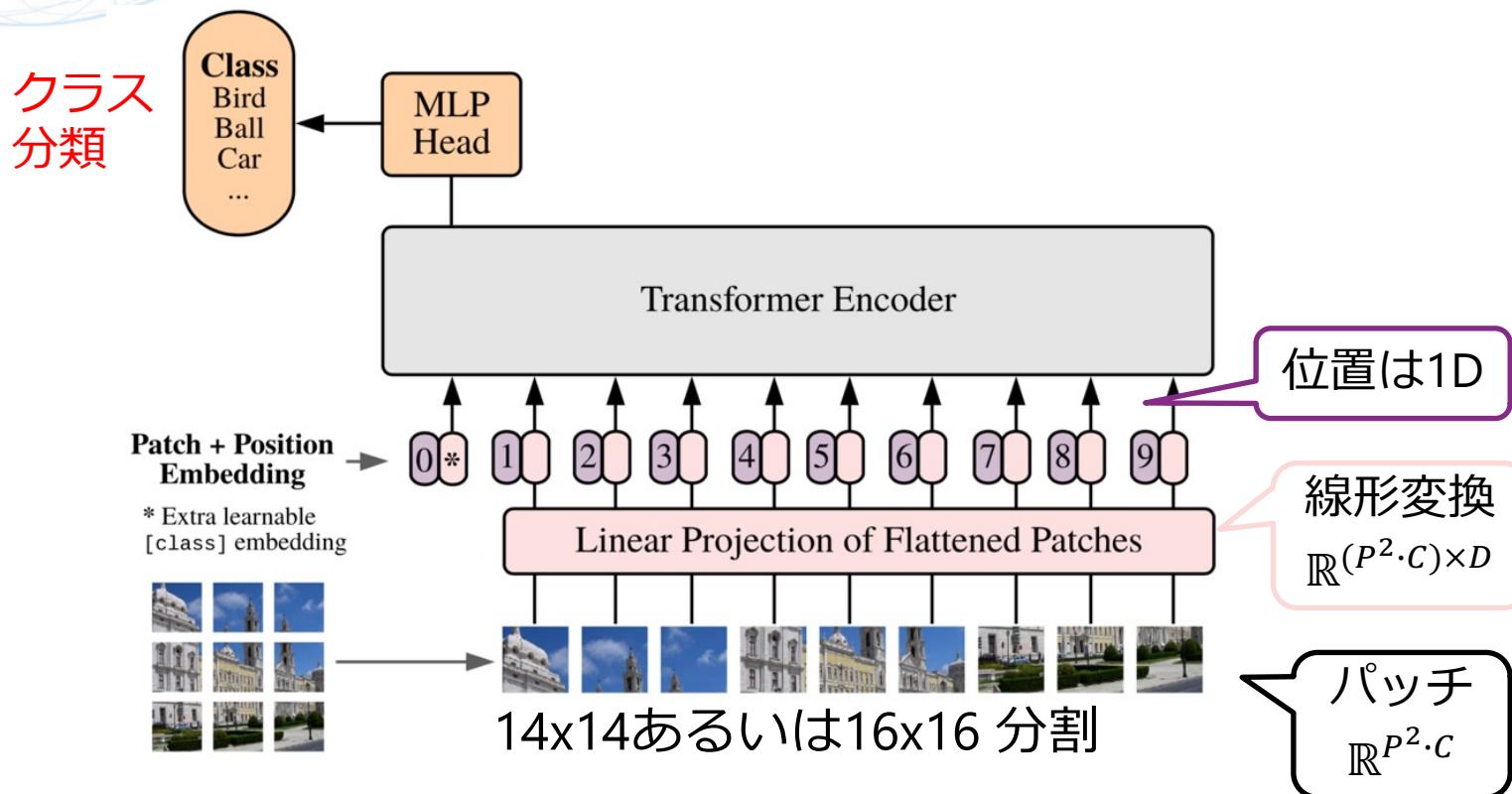
パッチ分割



Vision Transformer (ViT) [Dosovitskiy(Google)+, ICLR'21]

NTT

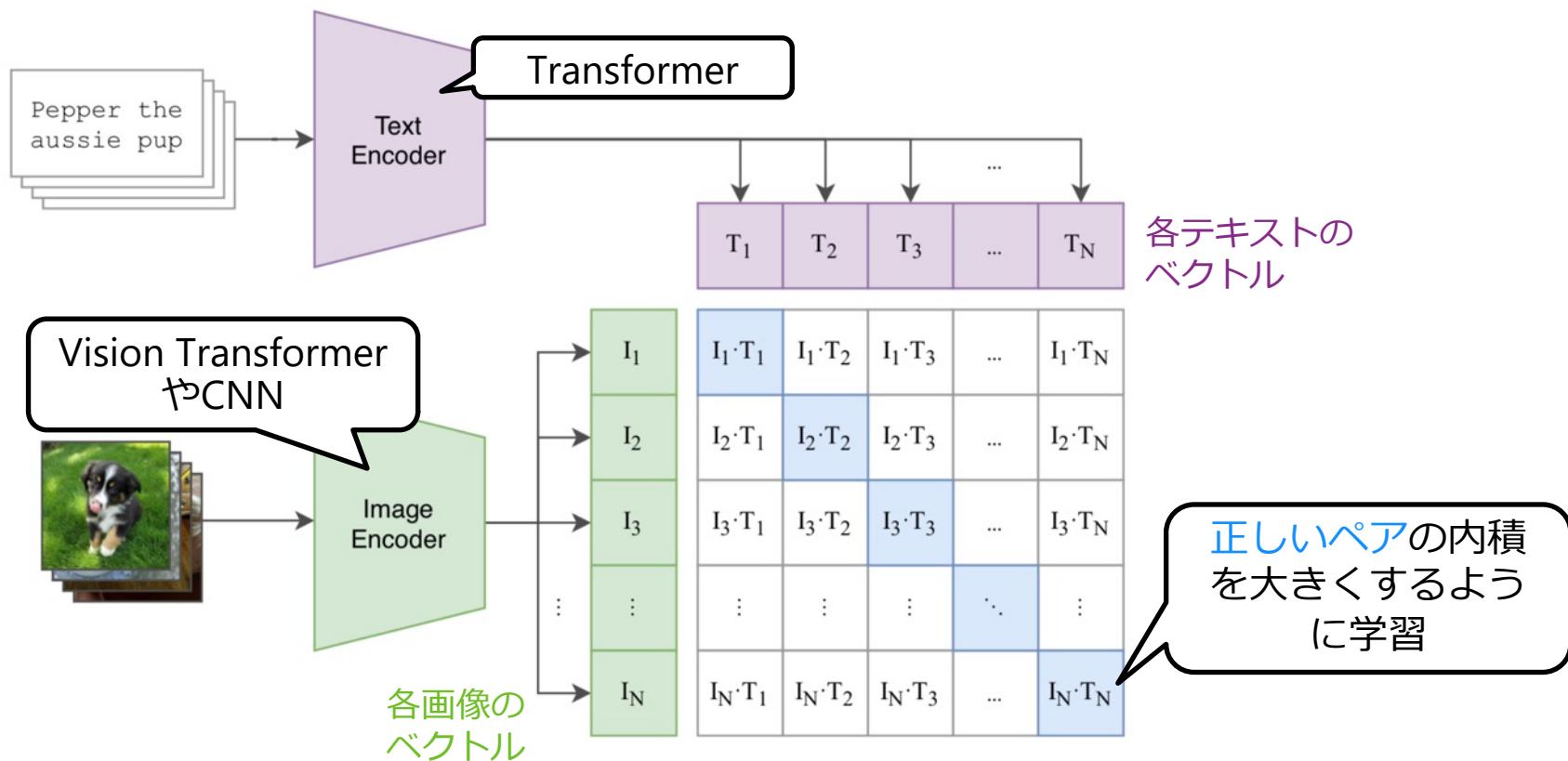
- 画像パッチを単語とみなすパラメータ数632MのTransformerエンコーダ
- 3億枚以上の画像分類で事前学習し、画像分類タスクで最高精度を達成
- ただし、従来のCNNを圧倒する程の精度では無かった
→ 自然言語を用いない画像分類タスクでの事前学習に限界がある？



CLIPの対照学習アプローチ

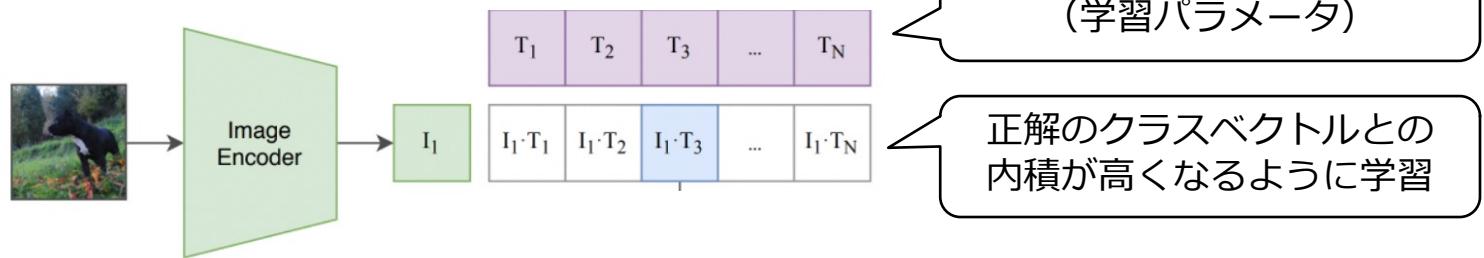
CLIP [Radford(OpenAI)+, 2021/01, ICML'21]

- Webから収集した4億件の画像とテキスト（画像の説明文）のペアから事前学習された視覚・言語の基盤モデル
- 画像エンコーダとテキストエンコーダが独立に入力をベクトル化
- 正しい画像とテキストのペアの内積が大きくなるように対照学習

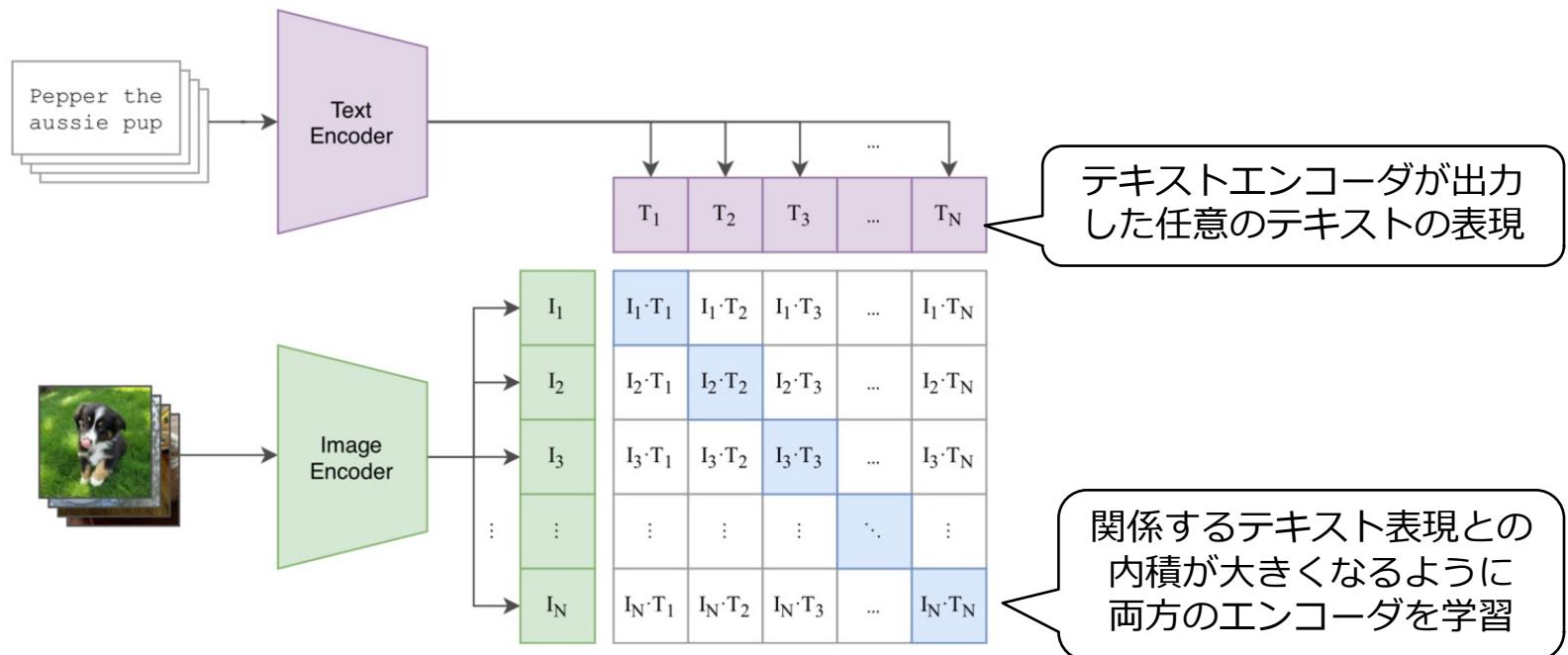


CLIP [Radford(OpenAI)+, 2021/01, ICML'21]

通常の画像分類

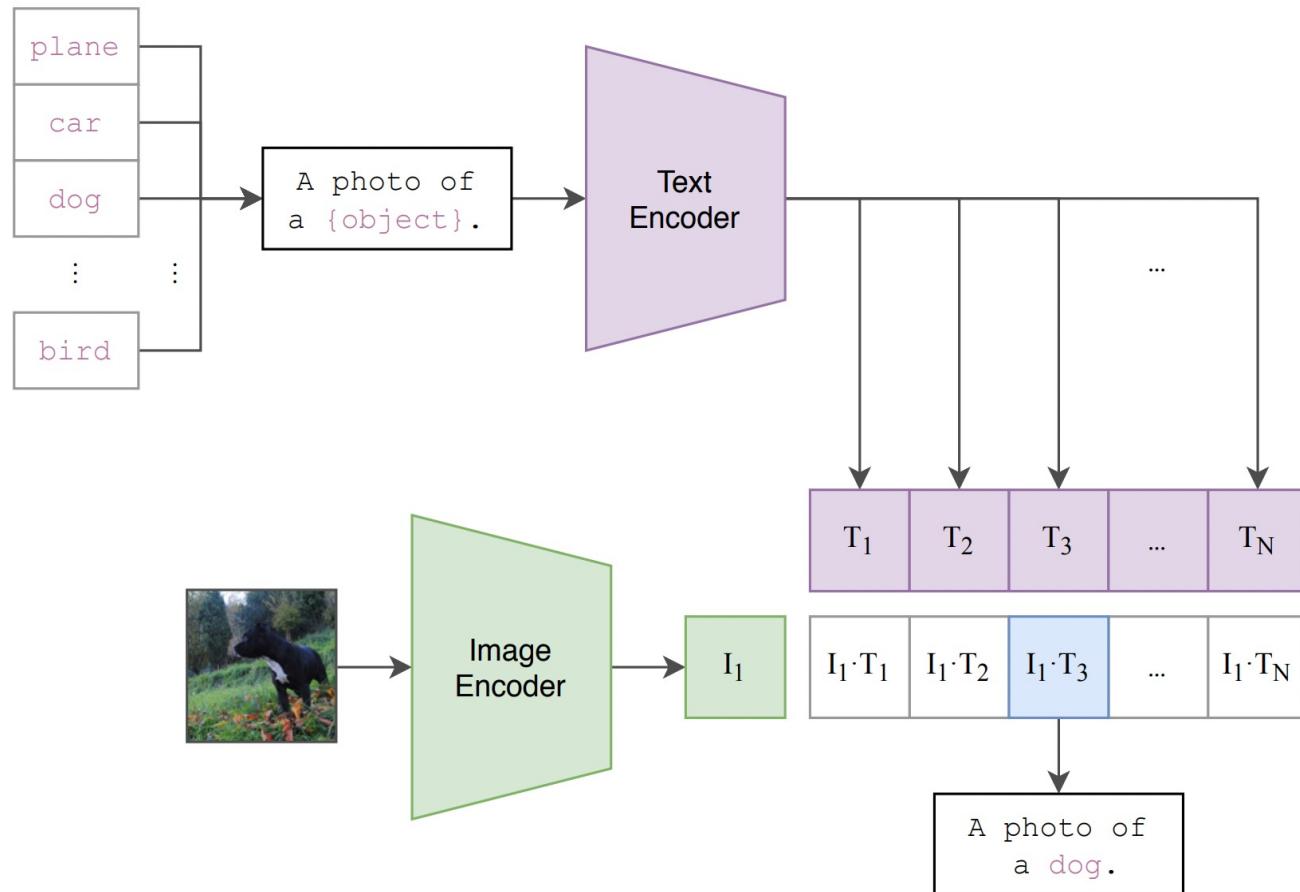


CLIP (Contrastive Language-Image Pre-training)



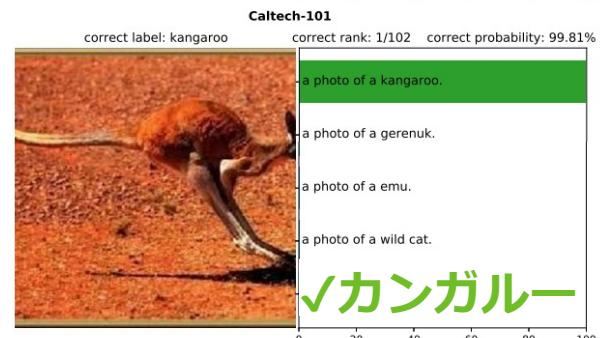
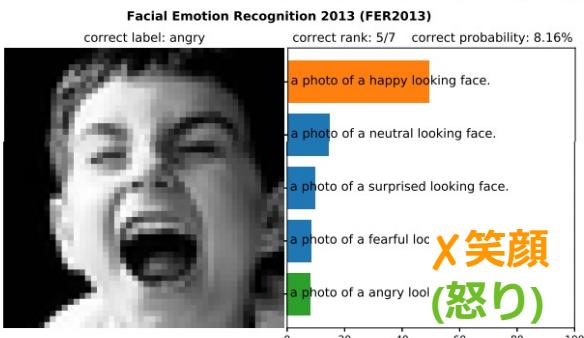
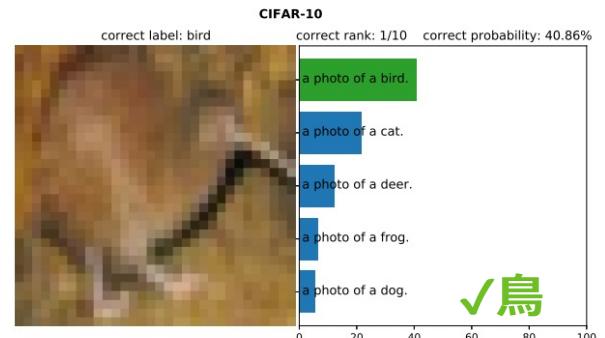
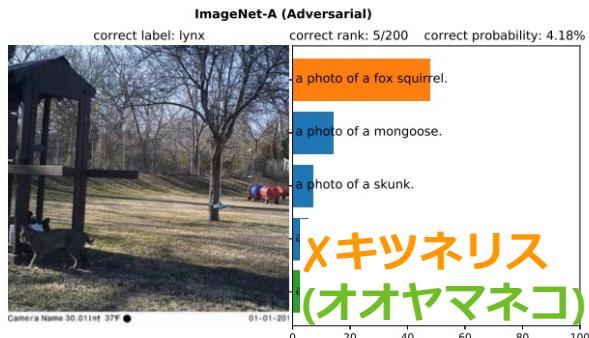
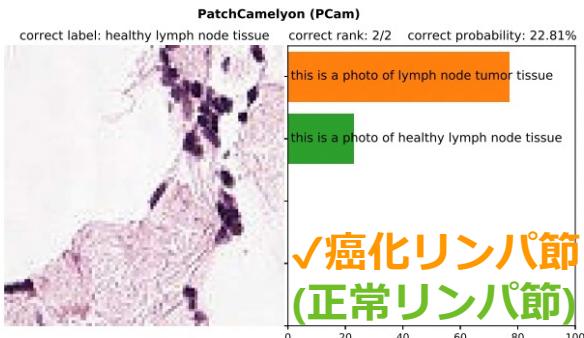
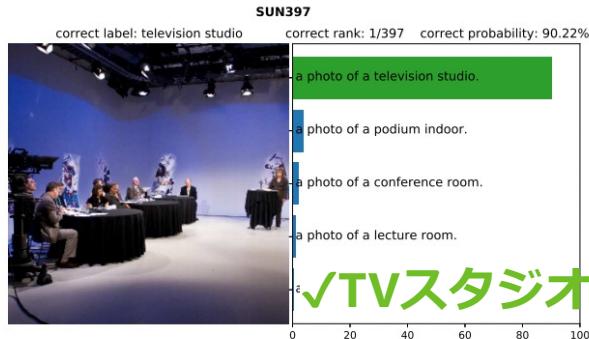
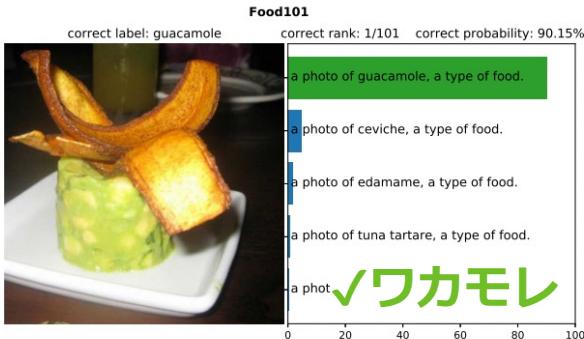
CLIP [Radford(OpenAI)+, 2021/01, ICML'21]

- 事前学習後は、追加の学習無し (zero-shot) で任意のテキスト候補の中から画像にマッチするものを選ぶことが可能



Zero-shot分類の例

- 従来の画像分類や物体検出ではカバーできていない概念も理解



CLIPは何が凄いのか(1)

- これまでの画像分類や物体検出で事前学習された画像エンコーダは、視覚と「有限個の記号」の結びつけであった
- CLIPにより、**視覚と「任意の言語」を結びつけた理解**が革新的に進み、最近の急速なモデルの進化につながった



画像分類・物体検出の場合



Class: 11

画像エンコーダはこのクラスが
何を意味しているのか理解していない



CLIPの場合



“赤、黄、白、黒などで鮮やかに
彩られた美しい小鳥が木に止まっている”

言語と結びつけた理解は
様々な概念の共通理解に通づる

CLIPは何が凄いのか(2)

- 従来のImageNetデータセットでは、25,000人が1,400万枚の画像に対して22,000クラスをWordNetの名詞と紐付けてアノテーション
- CLIPはWebから収集したデータを用いて、（ノイズはあるが）比較的低コストで遙かに広い概念を言語と結びつけて学習可能

ImageNetの “Tench” (コイ科の淡水魚) カテゴリの写真の例



本来不要な男性も含めてTenchが学習されてしまう

CLIP: 「釣り上げたTenchを抱えている男性」と学習

ImageNetの “Siberian Husky” カテゴリの写真の例

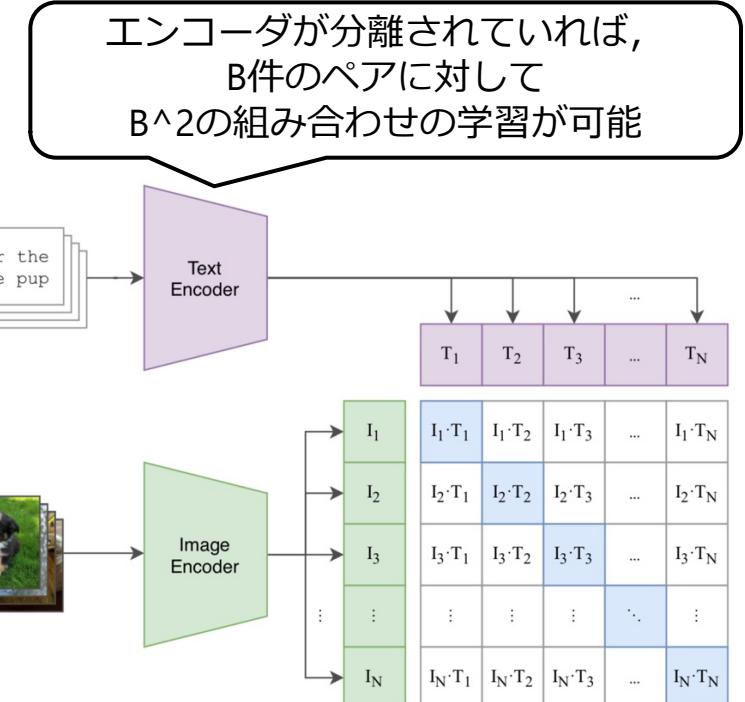


犬の数や動作については無考慮

CLIP: 「3匹のシベリアンハスキーが芝生に座る」と学習

CLIPは何が凄いのか(3)

- 初期のV&Lモデルでも画像とテキストのマッチングの学習は行っていたが、画像とテキストを同時に入力するので学習効率を上げにくい
- CLIPではエンコーダが分離されているので対照学習を行い易い（一度に多くのペアの類似度を計算可能）
 - テキスト検索においては“In-batch Negatives”として知られる技術 [Karpukhin, EMNLP’20]



CLIPが可能にした技術

StyleCLIP [Patashnik+, ICCV'21]

- CLIP空間上のテキストの変化方向を、StyleGAN [Karras+, CVPR'20] 空間に画像チャネル毎に射影し、自然言語による画像編集を可能にした
- 各空間における画像の変化の関係を予め前処理として獲得しておくことで、画像編集時は学習なしで編集可能



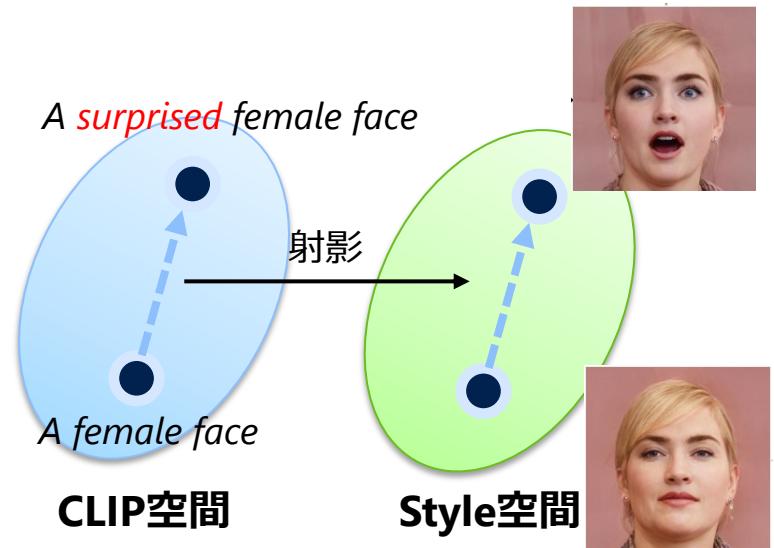
“Without makeup”



“Cute cat”



“Gothic church”



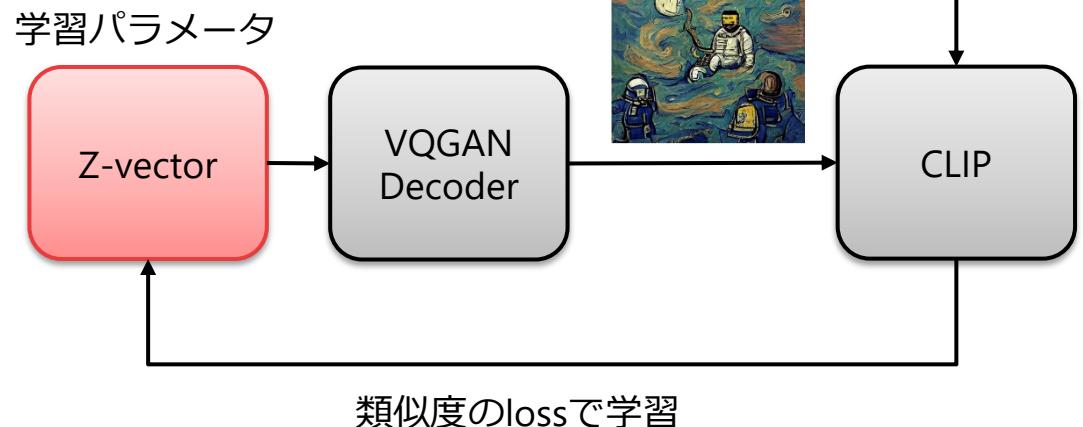
VQGAN + CLIP [Crowson(EleutherAI)+, 2021/07]

- VQGAN [Esser+, CVPR'21] で生成した画像がCLIP空間でターゲットテキストと類似するように潜在変数 (Z-vector) を最適化する
- “AIアート”としてSNS・ニュースサイト等で話題に

blue whales swimming through neon city



*an astronaut
in the style
of van Gogh*

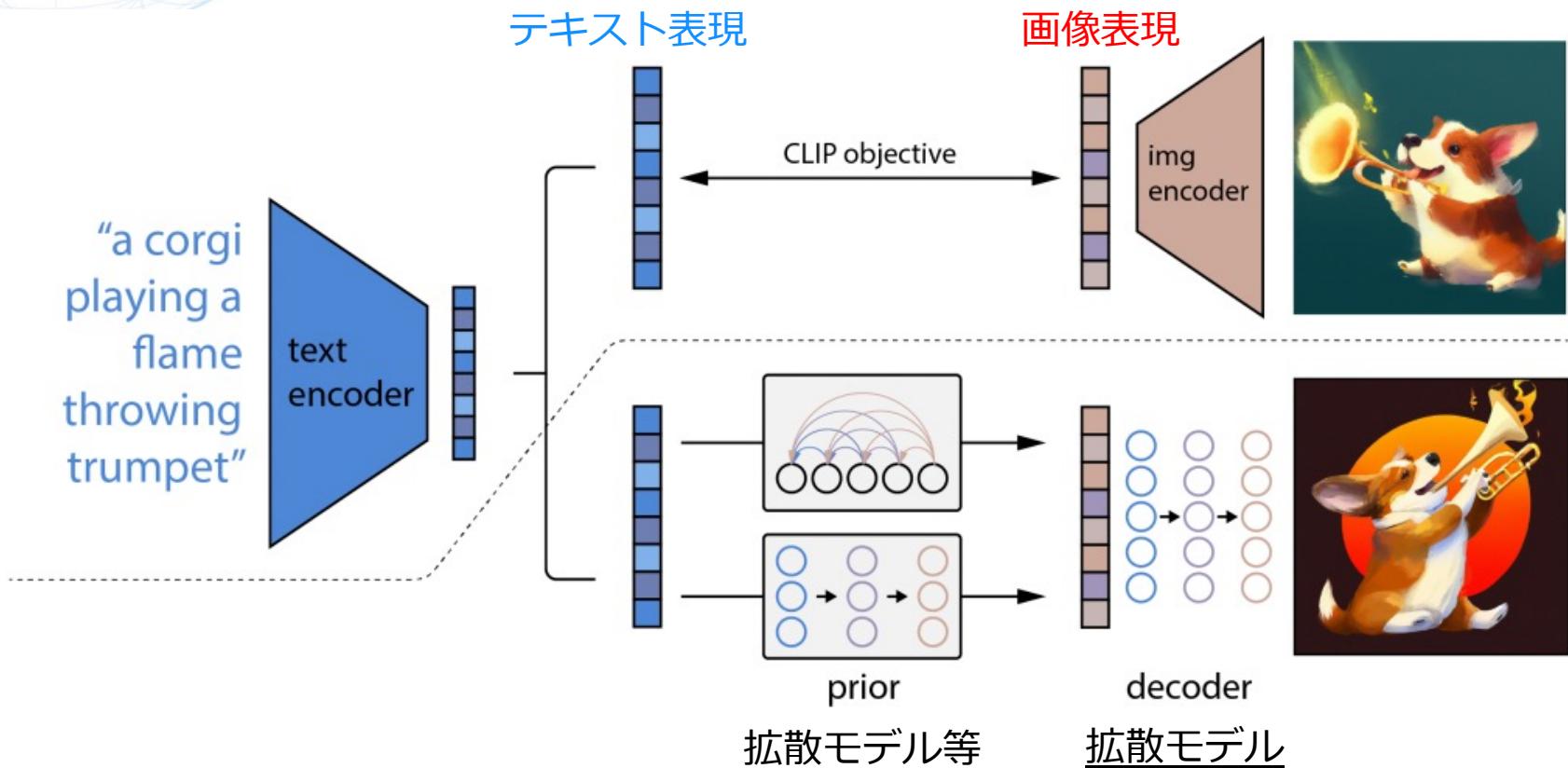


<https://twitter.com/ak92501/status/1413360535685435396>

<https://arxiv.org/abs/2204.08583>

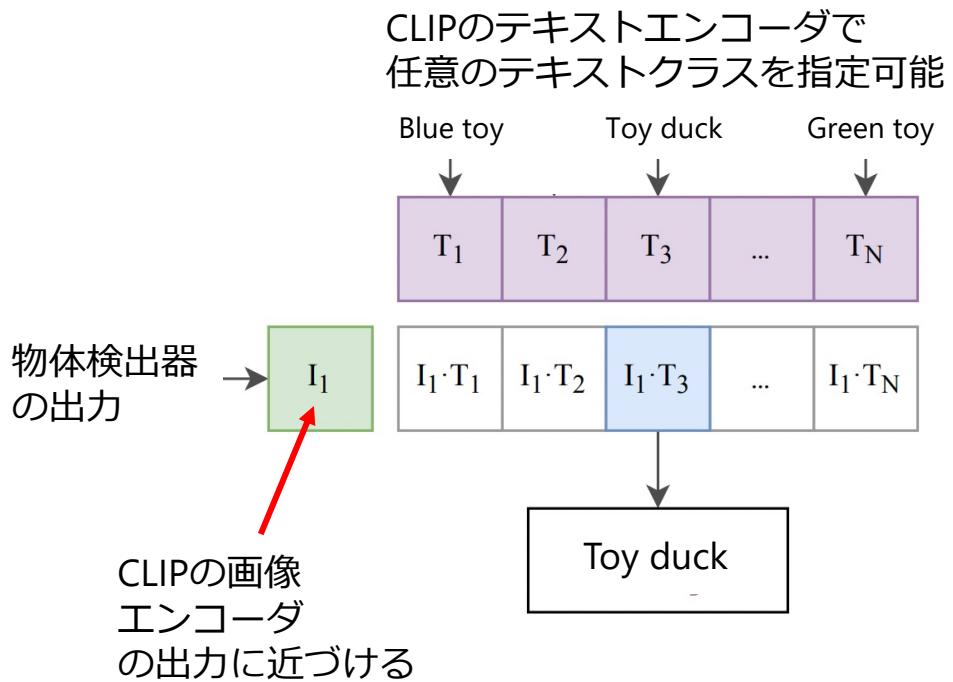
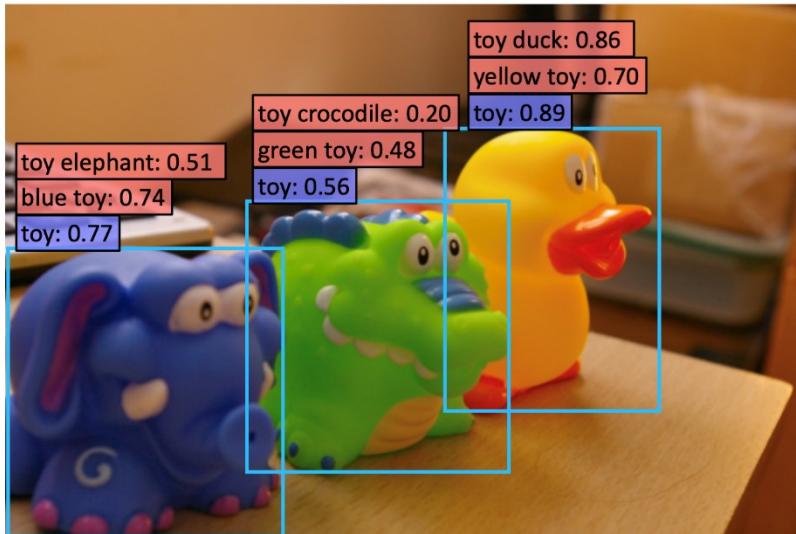
DALL-E 2 [Ramesh(OpenAI)+, 2022/04/13]

- CLIP + 拡散モデル [Ho+, NeurIPS'20] によるテキストからの画像生成
- (1) CLIPのテキスト表現→画像表現を生成するpriorと, 2) CLIPの画像表現→画像を生成するdecoderにて画像を生成



ViLD [Gu(Google)+, ICLR'22]

- 学習時に、物体検出モデルの出力する領域の画像表現を、CLIP空間の画像表現に近づける（蒸留）ようにすることで、任意のテキスト（"open vocabulary"）で物体検出可能なモデル

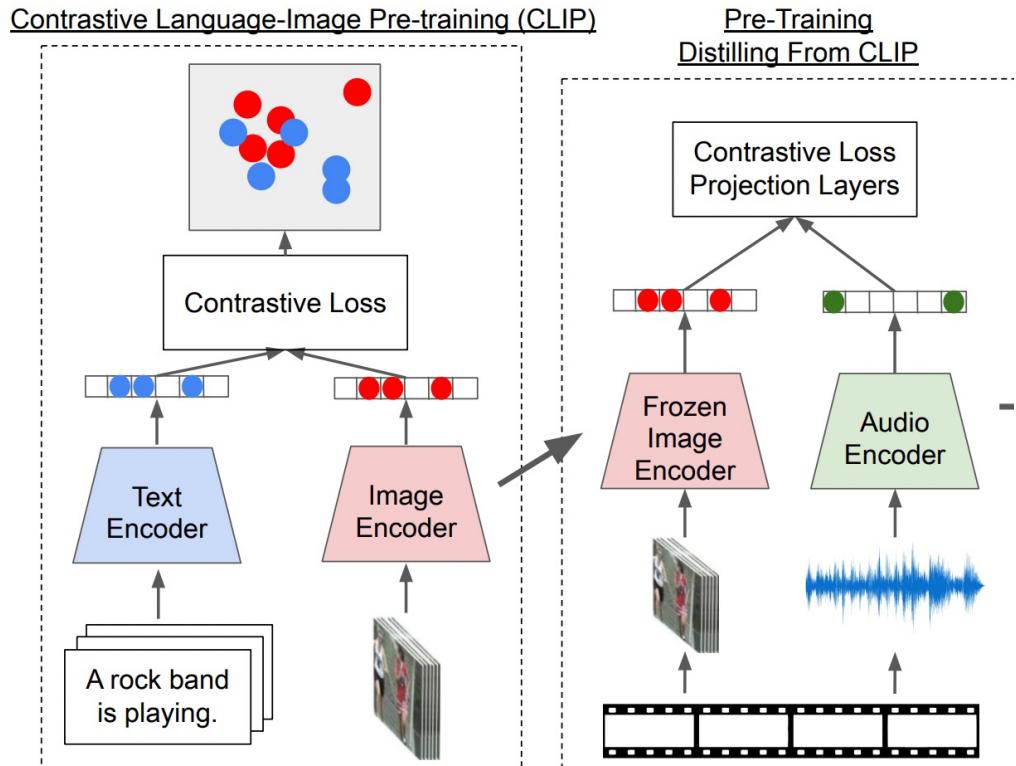


Wav2Clip [Wu(Descript)+, ICASSP'22]

- ・ 視覚・言語で学習したCLIPを動画データを用いて音と結びつける
- ・ VQGAN+CLIPを利用して、音声に関連する画像を生成することが可能



<https://descriptinc.github.io/lyrebird-wav2clip/>



<https://arxiv.org/abs/2110.11499>

CLIPasso [Vinker(EPFL)+, SIGGRAPH'22]

- 画像からスケッチに変換. 元画像とスケッチ画像がCLIP空間（画像エンコーダのみ利用）上で近くなるように、n本の曲線を最適化していく



Plants



Objects



Faces

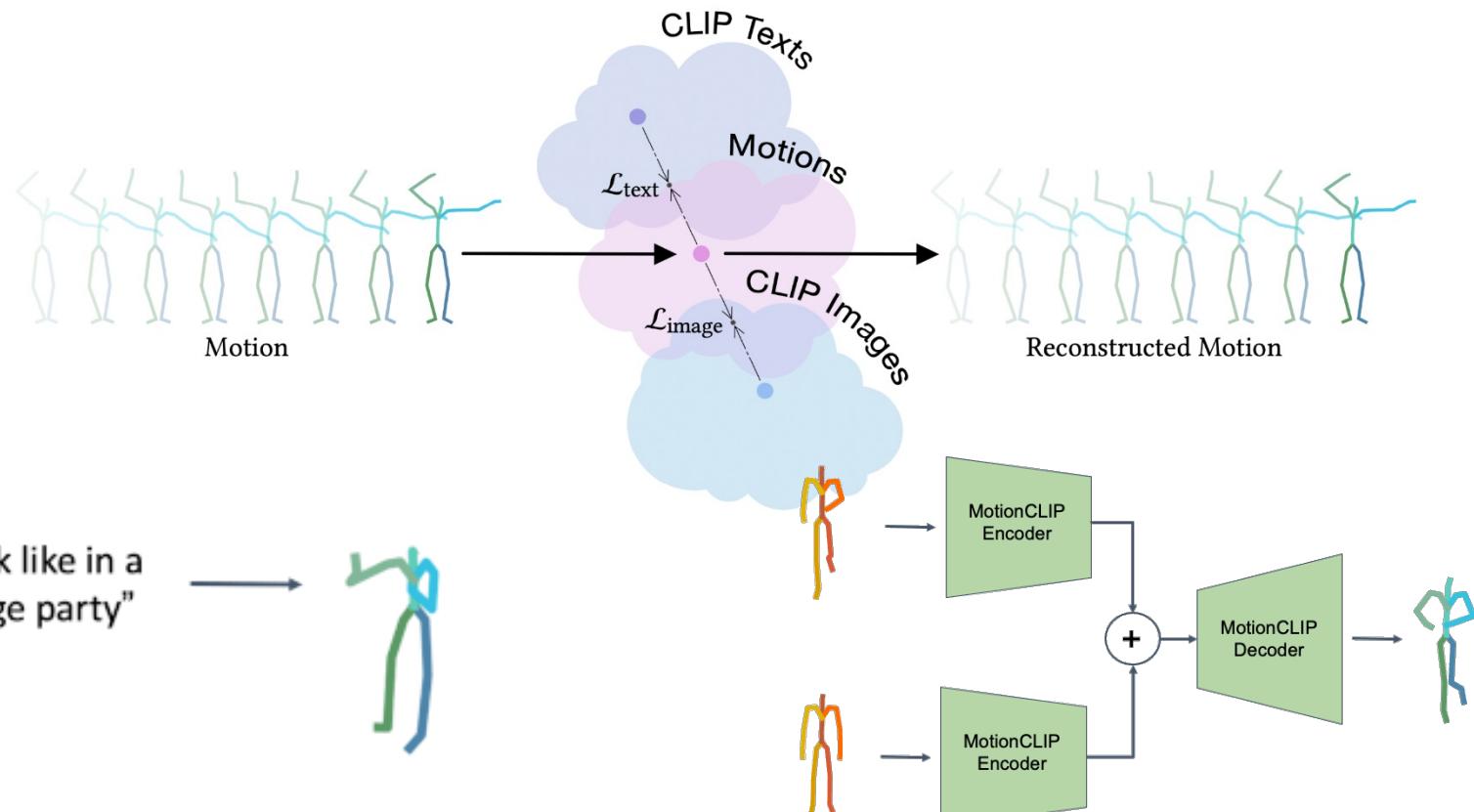


2D & 3D Artwork



MotionCLIP [Tevet(Tel Aviv U.)+, ECCV'22]

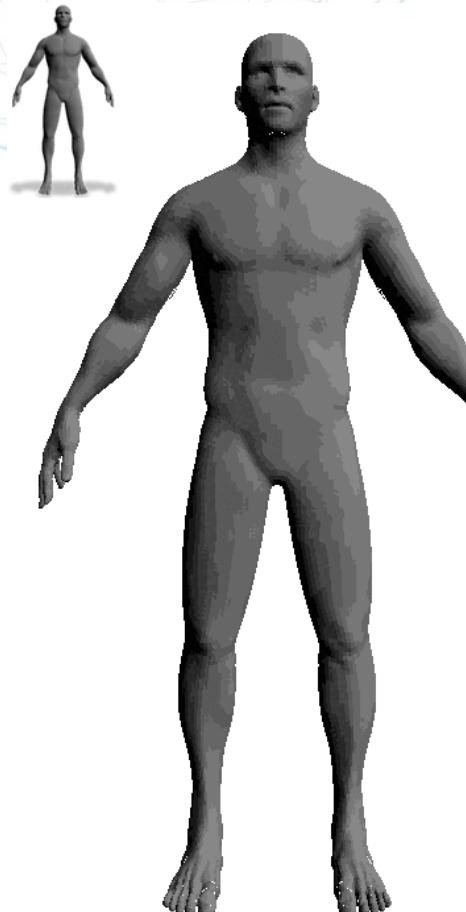
- 運動（モーションキャプチャの系列）のオートエンコーダの潜在変数を CLIP空間の言語・視覚と結びつけて学習
- モーションの生成、任意の運動の認識、運動の合成などが可能に



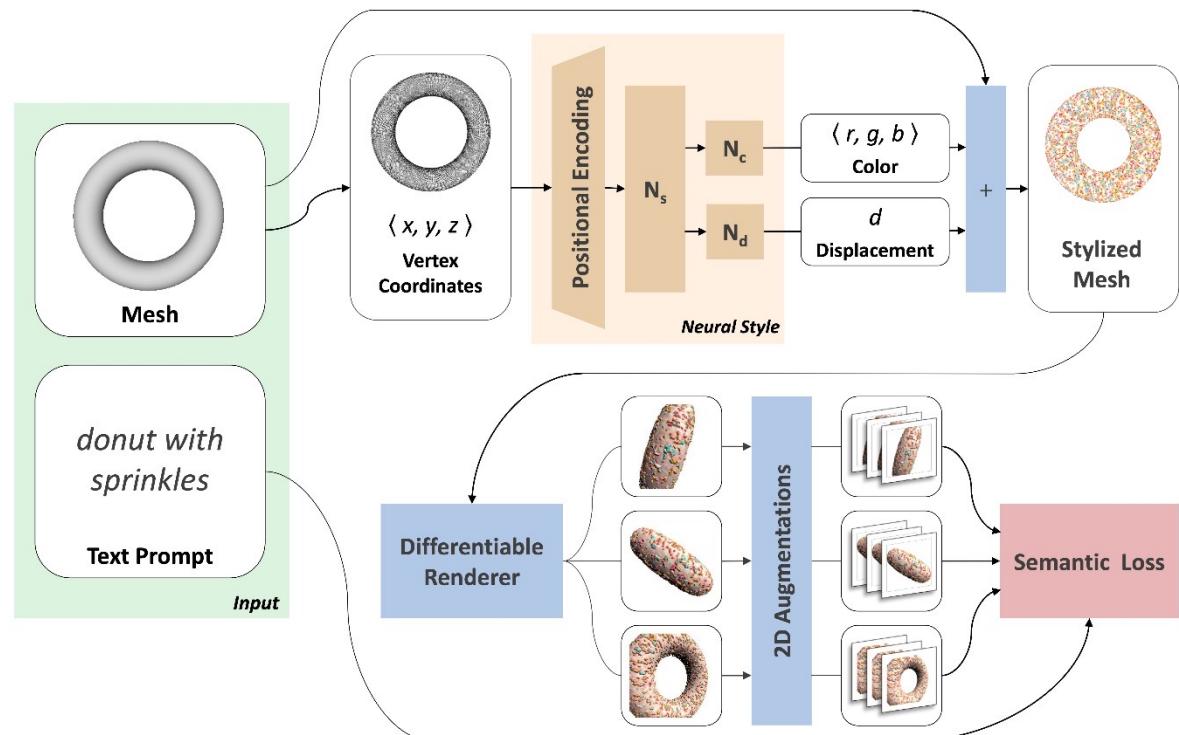
https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136820349.pdf

Text2Mesh [Michel(Chicago U)+, CVPR'22]

- 3Dオブジェクト（メッシュ）とテキスト指示を受け取ると色や形状を生成してくれるモデル

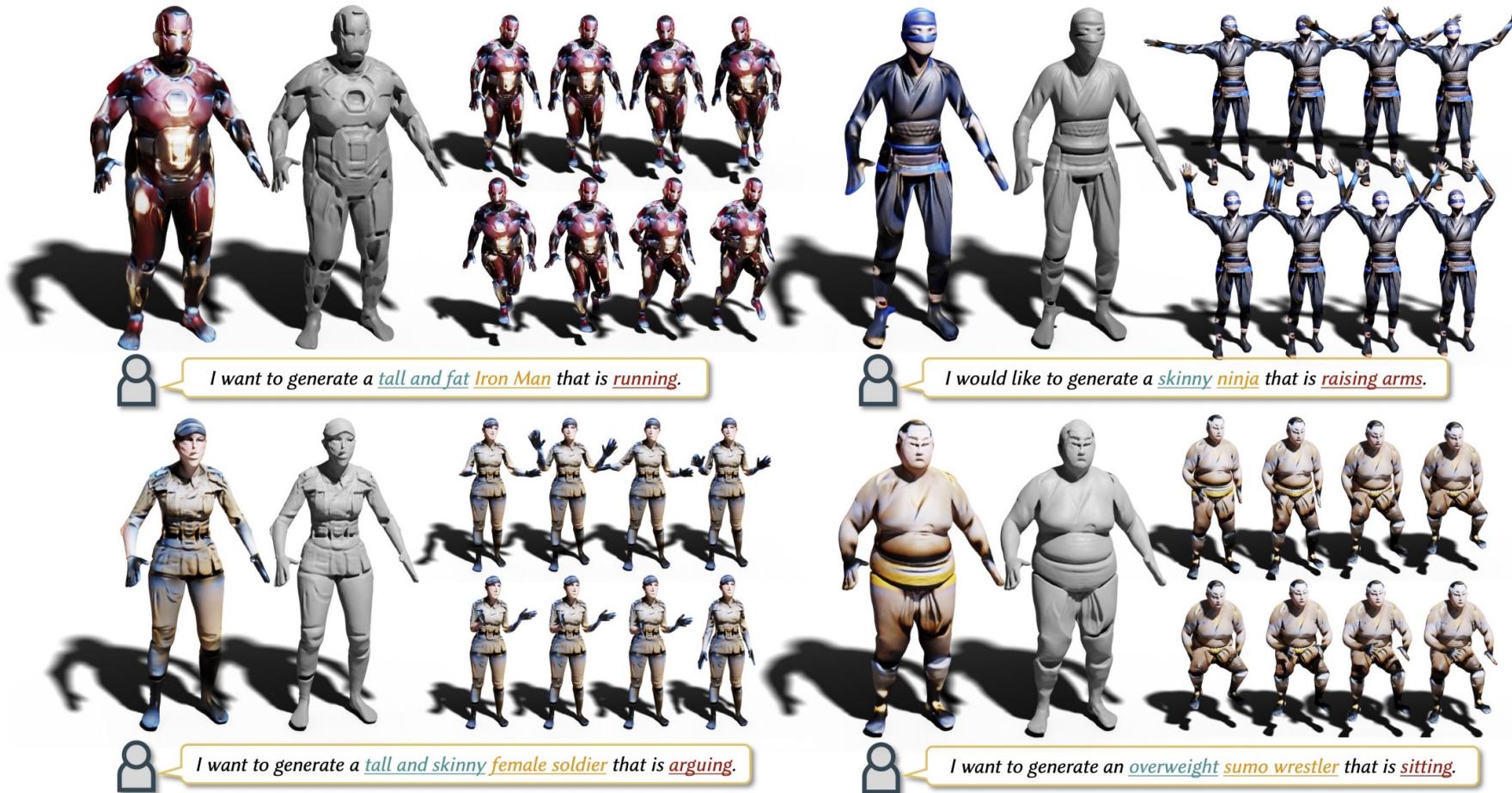


Iron Man



https://openaccess.thecvf.com/content/CVPR2022/html/Michel_Text2Mesh_Text-Driven_Neural_Stylization_for_Meshes_CVPR_2022_paper.html

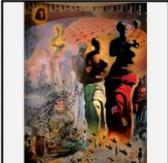
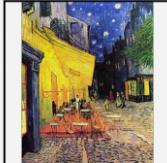
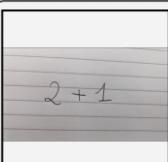
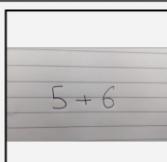
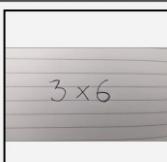
- 言語から3Dオブジェクトおよびモーションを生成する
- オブジェクトの形状やテクスチャ，ポーズなどを2D画像としてCLIP lossを使って生成



CLIPの画像エンコーダ + 大規模言語モデルの発展

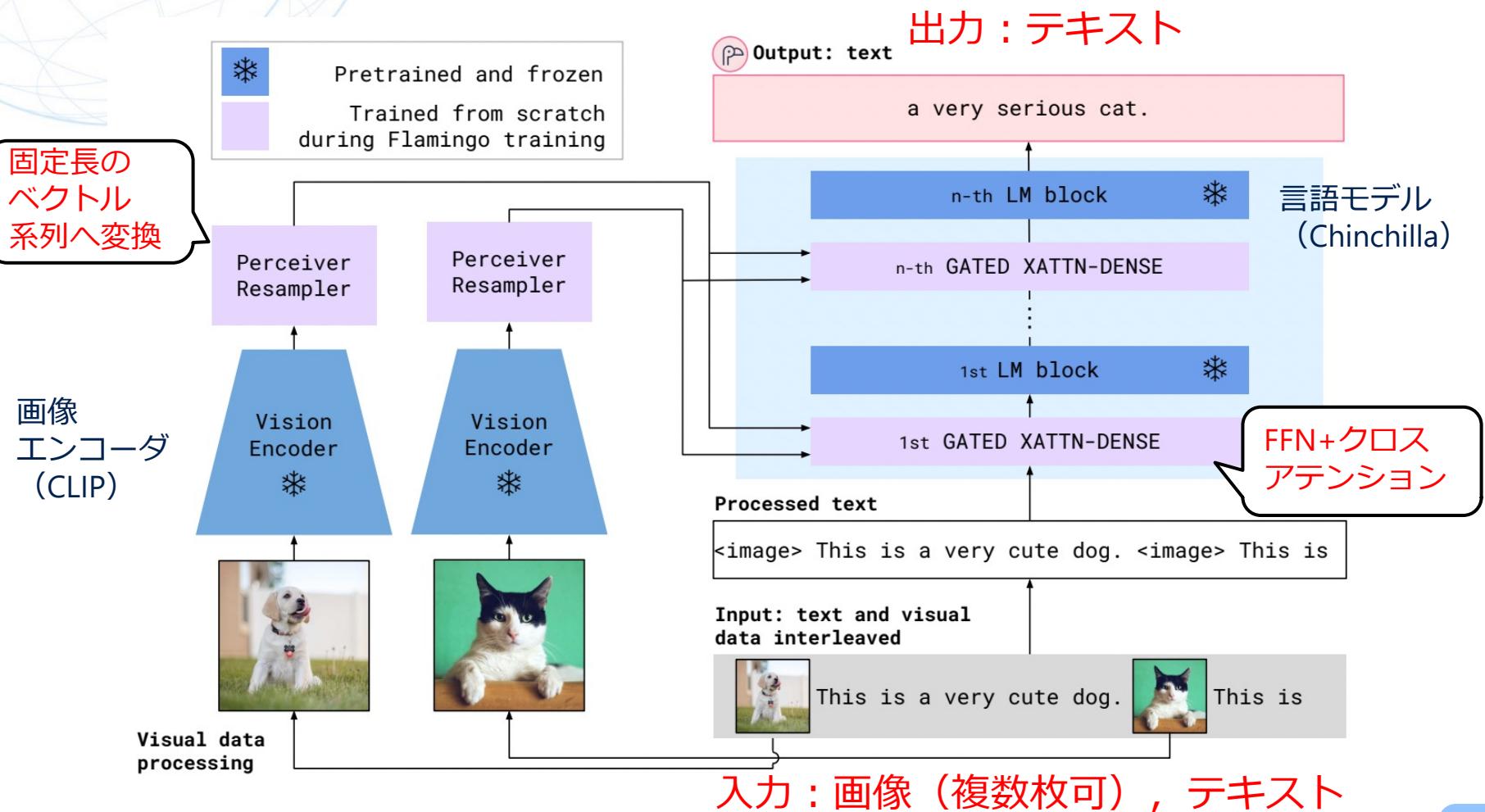
Flamingo [Alayac(DeepMind)+, 2022/04/29]

- 例示として「画像とテキスト」を与えてテキストを生成
- 追加の学習無しで汎用なタスクを解くことが可能

Input Prompt				Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.	 This is a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.	 What is the name of the city where this was painted? Answer: Arles.
	Output: "Underground"		Output: "Congress"	 Output: "Soulomes"
	2+1=3		5+6=11	 3x6=18

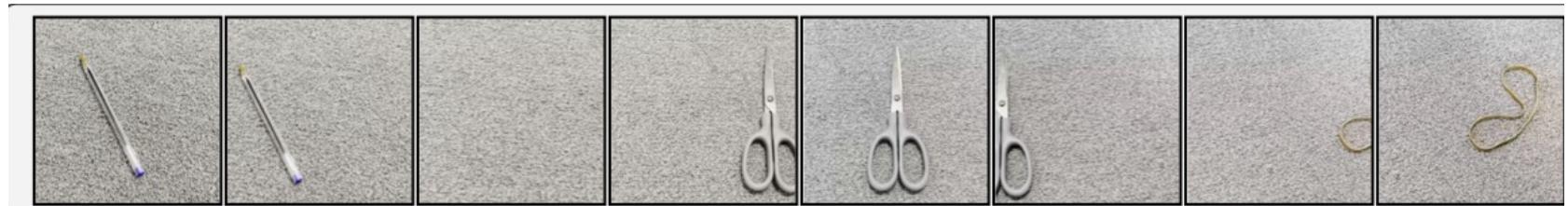
Flamingo [Alayac(DeepMind)+, 2022/04/29]

- CLPの方式で学習した画像エンコーダ（433M）と、事前学習済みの大規模言語モデル（70B）から構成。それぞれは固定し、適応用のネットワークを（194M／10B）を追加することで結合



Flamingo [Alayac(DeepMind)+, 2022/04/29]

- 動画に基づく質問応答の例



Question: What are the
three objects in this
video? Answer:



A scissors, a pen, and
a rubber band.



Flamingo [Alayac(DeepMind)+, 2022/04/29]

- 画像に関する対話の例



What do you think the robot is doing here?

P The robot is picking up a ball.

What color is the ball?

P It's red.

I want the ball to be hidden. Is there something the robot can use for that?

P The robot can use a cloth to hide the ball.

What color is the cloth?

P The cloth is blue.

知識を
用いた対話



Which city is this?

P This is a street sign in New York City.

Which street corner is it?

P It is the intersection of Madison Avenue and East 42nd Street.

Which subway station is nearby?

P The closest subway station is Grand Central Station.

How about the nearest park?

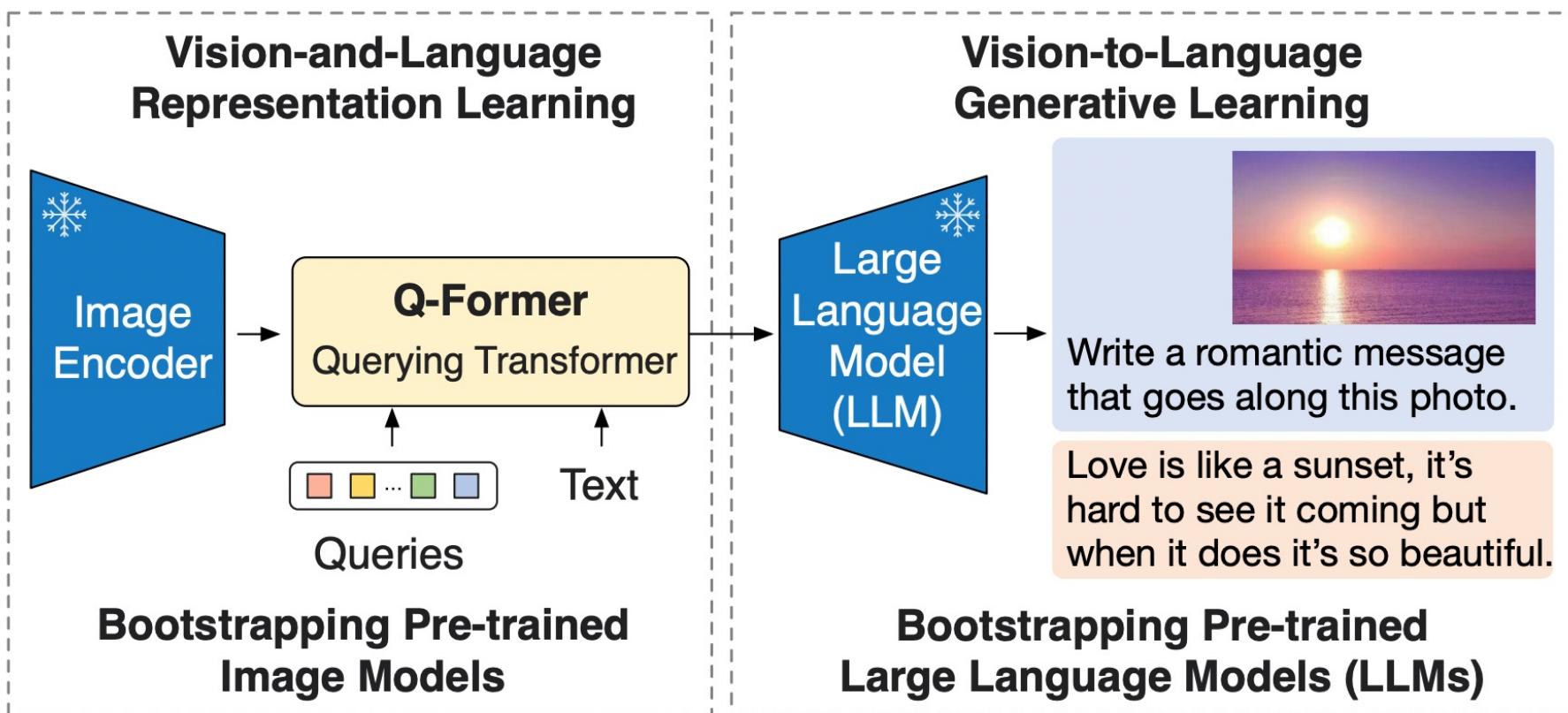
P The nearest park is Bryant Park.

There is another sign. What does it say?

P It says "One way".

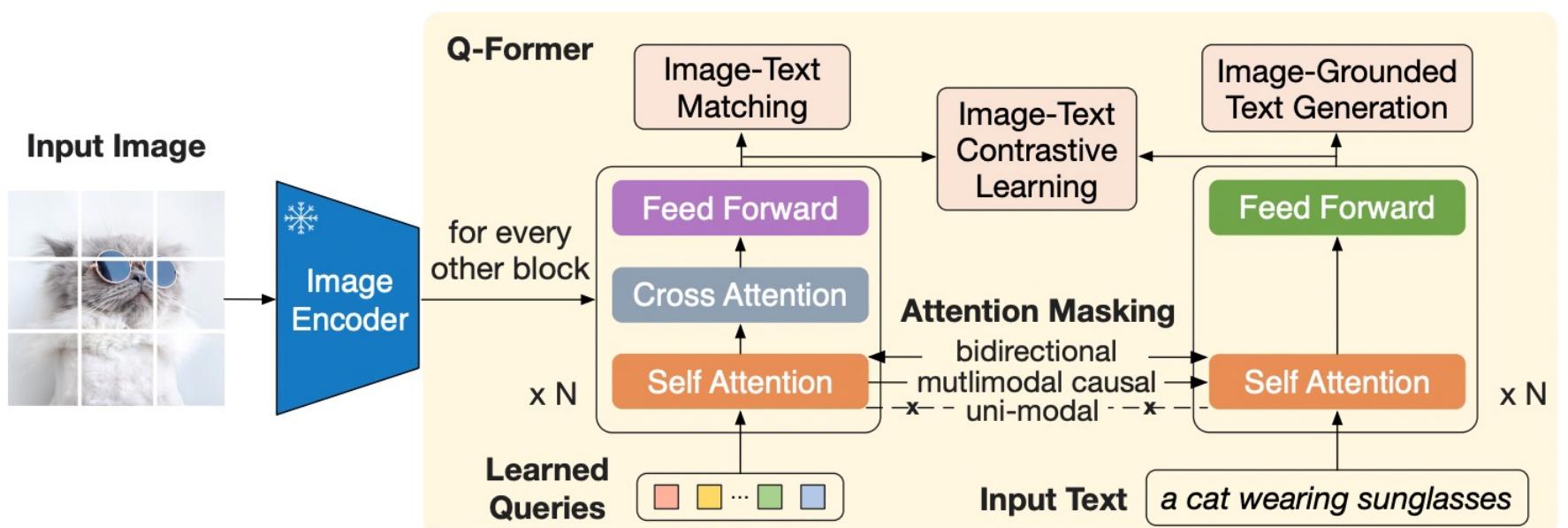
BLIP-2 [Li(Salesforce)+, 2023/01]

- CLIPのImage Encoderと大規模言語モデルを固定して、画像からLLMへ入力する「視覚的なプロンプト」を出力する適応用モジュールを追加し、2段階の事前学習を行う
- VQAなど一部ベンチマークではFlamingoを上回る性能



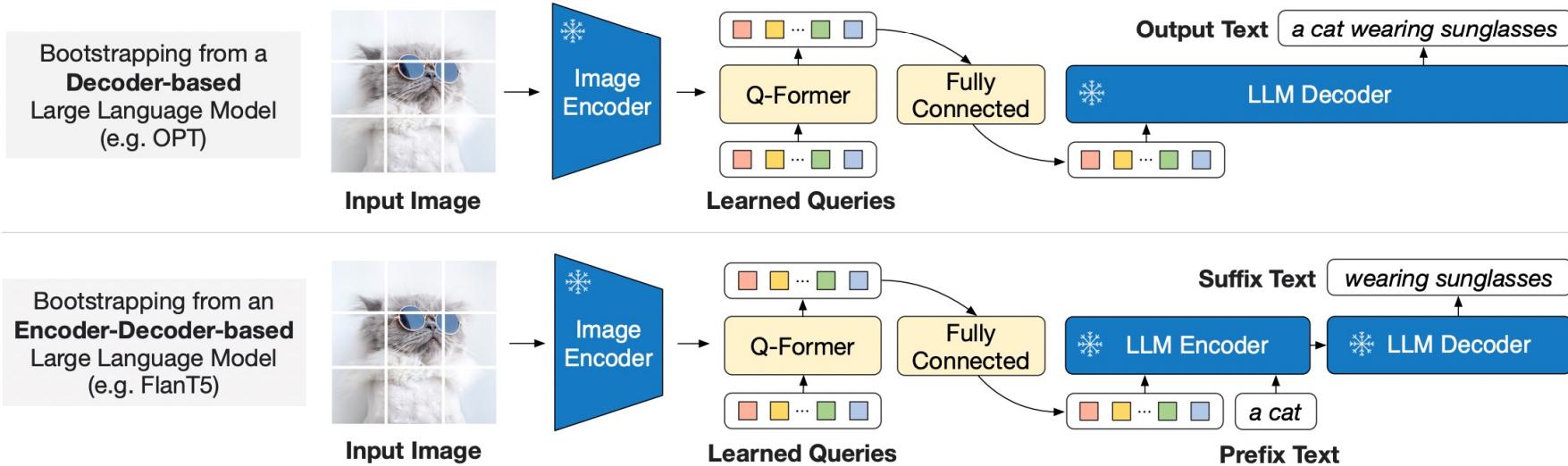
BLIP-2 [Li(Salesforce)+, 2023/01]

- 第一段階：Image Encoderを改善する
- 既存のCLIP（固定）のImage Encoderを直接LLM（固定）につなげるではなく、接続用ネットワークとしてQ-Formerを利用する
- Q-Formerは再度CLIPの対照学習を行ライ IMAGEだが、Image-encoderへの入力は“Learned Queries”となる



BLIP-2 [Li(Salesforce)+, 2023/01]

- 第二段階：LLMをQ-Formerの出力に適応させる
- Q-Formerの画像エンコーダ部分が“Learned Queries”に対する出力を、LLMへ入力し、対応するテキスト出力を得る
- どちらも<画像, キャプション>の組で学習するが、LLMの能力により画像に基づく対話が可能になる



BLIP-2 [Li(Salesforce)+, 2023/01]

8
8

Explain the advantages of this product.

The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.

8
8

Tell me something about the history of this place.

The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.

8
8

Write down the facts that you know about this flower.

It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.

8
8

Is this photo unusual?



Yes, it's a house that looks like it's upside down.

How could someone get out of the house?



It has a slide on the side of the house.

8
8

What are shown in the photo?



A man and a chicken.

What does the man feel and why?



He is scared of the chicken because it is flying at him.

8
8

What are the ingredients I need to make this?



Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

8

What is the first step?



Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

FROMAGe [Koh(CMU)+, 2023/01]

- Image EncoderとLLMを線形層だけで接続し、画像キャプショニングと対照学習でシンプルに学習
- [RET]トークンを画像表現と一致させることでテキストと画像の両方をLLMが生成可能になる
- CLIP ViT-L/14とOPT 6.7Bを接続して画像に対する対話や検索を実現

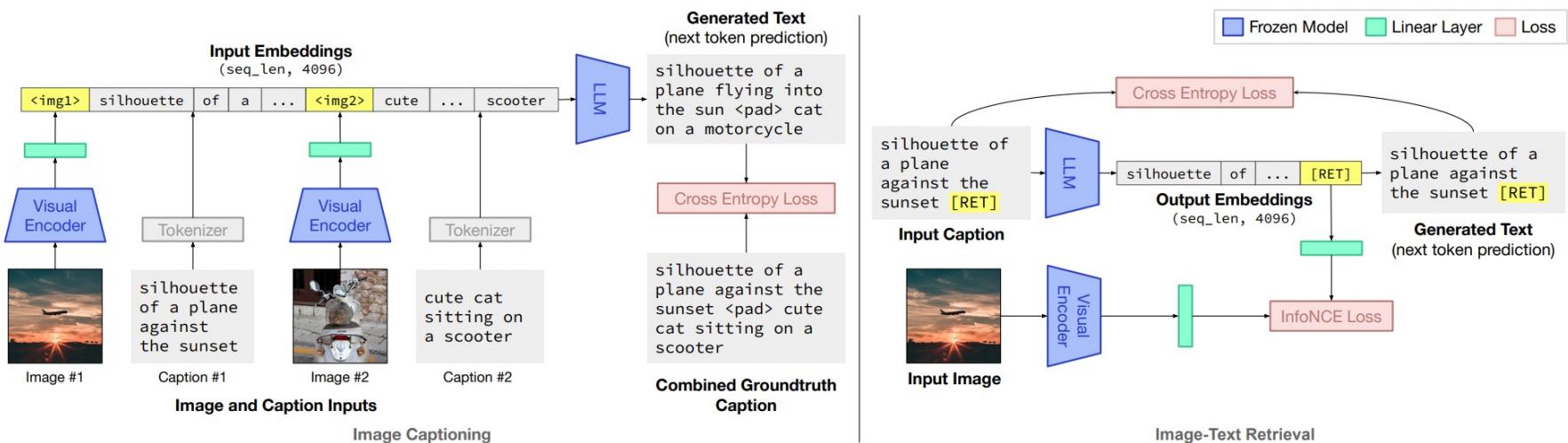


Figure 2. Overview of the FROMAGe architecture. FROMAGe is a model trained on image-text pairs for image captioning and image-text retrieval. It is capable of processing arbitrarily interleaved image and text inputs, and producing interleaved images and text as outputs.

FROMAGe [Koh(CMU)+, 2023/01]

- FROMAGeはテキスト→画像検索ができるので、モデルが対話中に適切な画像を選択して対話することが可能

I am thinking of an animal. It is brown and furry, and has a tail. What is it?

A squirrel.

No, the animal I am thinking of lives in the river.

A beaver.

That's right, what do beavers look like?

They have a tail.

Show me an illustration of a beaver



A beaver.

Yes, what about a pencil drawing of one?



What about a photo of one?



What is this?

This is a house with a unique design.

What is its shape?

The house is shaped like a dome.

Can you show me more houses like this?

The house is designed by architect.

That's nice, any others?

Yes, there are many houses like this in the world.





Provide instructions on how to cook.

How do I make macarons?

Macarons are a French dessert that are made by combining egg whites and sugar. The egg whites are beaten until stiff and then folded into the sugar.

What does the stiff egg white look like?

The egg white should be stiff and glossy.

Show me a picture of what that looks like



After this, what do I do?

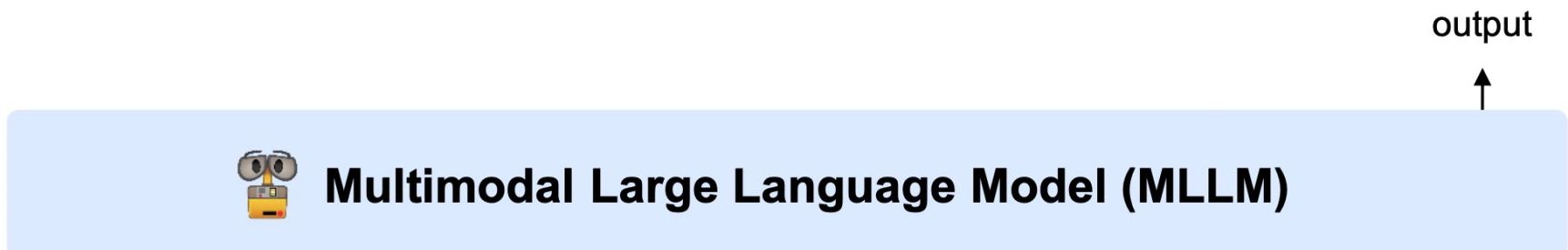
The macaron is then baked in a hot oven.

Multimodal Dialogue

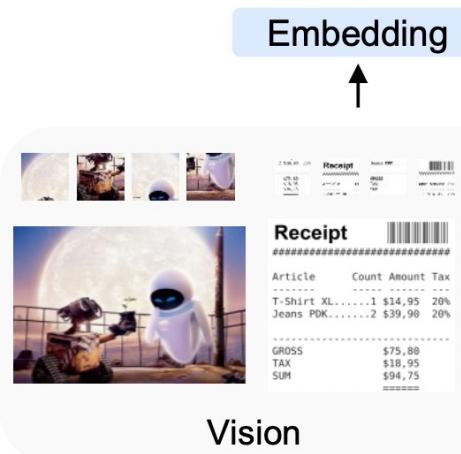
Our model can generate multimodal dialogue. Green bubbles indicate model generated outputs, grey bubbles indicate user provided prompts.

KOSMOS-1 [Huang(Microsoft)+, 2023/02]

- 画像入力を受け付け可能なマルチモーダルLLM（パラメータ数1.6B）
- CLIPのImage Encoder（ViT-L/14）の最終層以外を固定して画像をテキストと合わせて入力する
- 言語コーパス、画像キャプション、画像-テキストの綴じ込み（interleaved）データで学習



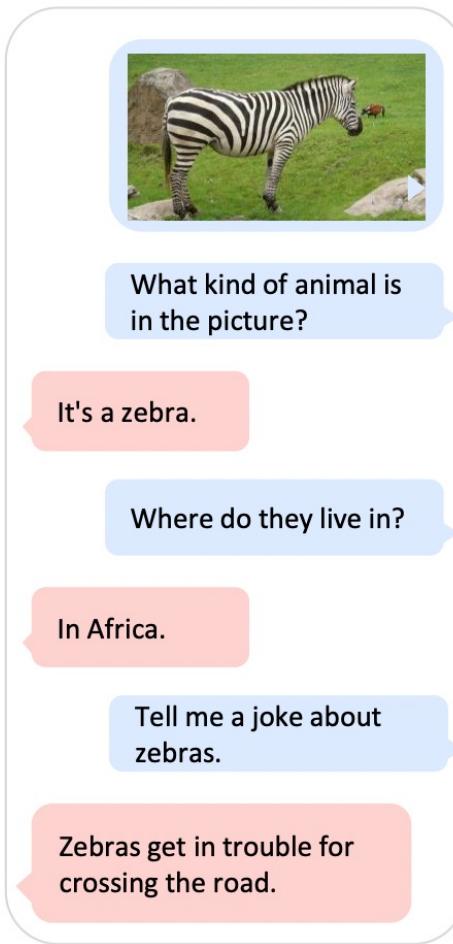
Kosmos-1 can perceive both language and   , learn in context , reason, and generate



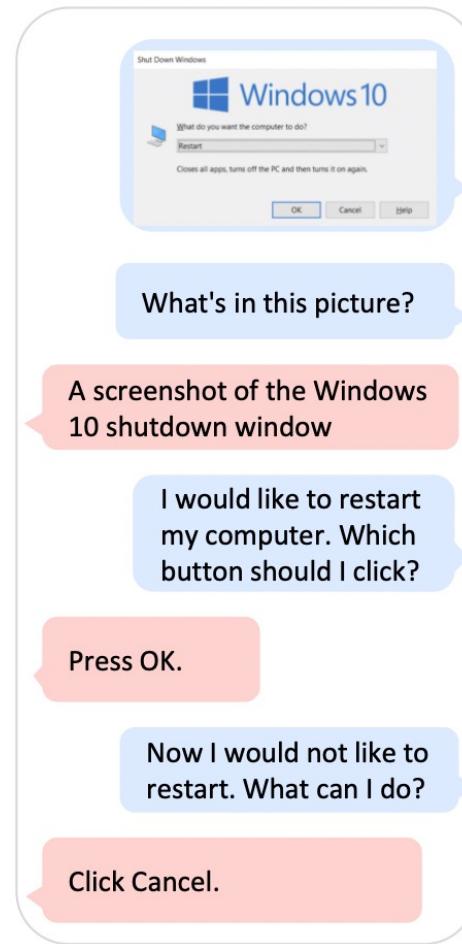
KOSMOS-1 [Huang(Microsoft)+, 2023/02]



(9)



(10)



(11)

Figure 3: Selected examples generated from KOSMOS-1. Blue boxes are input prompt and pink boxes are KOSMOS-1 output. The examples include (1)-(2) image captioning, (3)-(6) visual question answering, (7)-(8) OCR, and (9)-(11) visual dialogue.

ここまでまとめ

- Vision-and-Languageタスクも理解・生成に大別される
- BERT以降，画像を物体検出して系列化し，言語の系列と合わせてTransformerでモデリングする方式の検討が進んできた
- 2021/01の対照学習を用いるCLIPにより，画像分類や物体検出に依存した事前学習の限界を越え，視覚と言語の融合が大きく進んだ
- 近年では，画像・テキスト検索のみならず，テキストからの画像生成，3dオブジェクト・モーション生成など，様々なタスクでCLIPの導入が進んでいる
- CLIPの優れた画像エンコーダと大規模言語モデルを結合する取り組みも活発化してきている

目次



西田京介

1. 深層学習による自然言語処理技術の基礎 (40min)

タスクとモデルの分類, Transformerとセルフアテンション, BERTとGPT-3, 大規模言語モデル（LLM）に関する知見



西田光甫

2. 大規模言語モデルの革新 (40min)

InstructionチューニングとRLHF, ChatGPT, ChatGPTの評価と言語モデルの今後の課題

質疑応答 (10min)



西田京介

3. Vision-and-Languageへの派生 (40min)

Vision Transformer, 視覚・言語の基盤モデルCLIP, CLIP+LLMによるモデル：Flamingo, KOSMOS-1など



田中涼太

4. 文書や画面の理解 (40min)

処理・タスク・モデルの紹介, 文書画像理解の発展：SlideVQA, LiLT, Pix2Struct, ReActなど

質疑応答 (10min)

テキストの読解から 視覚的な文書読解へ

我々はどういった文書を扱っているのか？ NTT

フォーム

STOUT INDUSTRIES, INC.

10675
6425 W. FLORISSANT AVE., ST. LOUIS, MO. 63136 (314) 365-0290

PROPOSAL

To: Gorillard Corporation
Address: 666 Fifth Avenue
City: New York
State: NY 10103
Attn: Mr. Robert Kennedy

For: Metal "Pack" Plaque
Date: October 16, 1987
Mr. A. D. Stelberg
Your Representative

It is our pleasure to propose the following:

ITEM: Harley Davidson Metal Plaque	SIZE: 17½ x 23¾"	
MATERIAL: Steel <input checked="" type="checkbox"/>	OTHER: Aluminum <input type="checkbox"/>	GAUGE: .025"
COLORS: Transparent gold, opaque black, white and orange	SINGLE FACE <input type="checkbox"/>	DOUBLE FACE <input type="checkbox"/>
BASE COLOR: Aluminum	Tooling: Form die, brass emboss die to achieve detail on eagle.	
POLES: 58" <input type="checkbox"/>	NUMBER OF CORNERS: ROUND <input type="checkbox"/>	SQUARE <input type="checkbox"/> ANGLE <input type="checkbox"/> CUT TO SHAPE <input type="checkbox"/>
EDGES: ROUNDED <input type="checkbox"/>	CURLED <input type="checkbox"/>	EMBOSSED <input checked="" type="checkbox"/> BEADED BORDER <input type="checkbox"/>
STAMP FRAME <input checked="" type="checkbox"/>	RIGHT ANGLE BEND <input type="checkbox"/>	BACK FRAME <input type="checkbox"/>
PACKING: PER CARTON: 10 <input type="checkbox"/>	PER CRATE: <input type="checkbox"/>	PER BUNDLE: <input type="checkbox"/>
OTHER: Price is based on reproduction of customer supplied "Pack" box. Tooling: Form die, brass emboss die to achieve detail on eagle.		
QUANTITIES: 500 Plaques	One time tooling \$ 23,015.00	
PRICE: \$9.18 each	Steel tips \$1,045.00	
<i>5000</i>		
BILLING: BILL AS MANUFACTURE <input checked="" type="checkbox"/>	BILL AS SHIP FOR: 12 MOS. <input type="checkbox"/>	
WAREHOUSING: SHIP IMMEDIATELY. <input checked="" type="checkbox"/>	6 MOS. WHSE. <input type="checkbox"/>	
DROP SHIPPING: PEI SHIPMENT <i>PEIAC 1850</i>	12 MOS. WHSE. <input type="checkbox"/>	
<i>57528360</i>		
CONDITIONS: THIS PRICE IS FOB ST. LOUIS. PRICES ARE BASED ON CURRENT COSTS AND ARE FOR IMMEDIATE ACCEPTANCE. These prices will be updated periodically. First day of each, to reflect changes in material and labor costs. These prices do not include shipping, insurance or handling charges. These prices do not include taxes. Taxes will be added by the home office of this company in St. Louis, MO. No contracts for delivery are completed upon sale, unless an order contract is made with the home office of this company in St. Louis, MO. No contracts for delivery are completed upon sale, unless an order contract is made with the home office of this company in St. Louis, MO. Any taxes or governmental charges which the seller may be required to pay on or collect again from the producer, sale, storage, or delivery, under any existing or future laws or regulations, shall be paid by the seller and will be added to the price of the article upon demand. F.O.B. ST. LOIS, MO. MINIMUM FREIGHT: Minimum flat rate charged freight collect. F.O.B. ST. LOIS, MO. MINIMUM FREIGHT: Minimum flat rate charged freight collect. F.O.B. ST. LOIS, MO. MINIMUM FREIGHT: Minimum flat rate charged freight collect.		
TERMS: -NET 30 DAYS A service charge of 1% per month will be applied to any unpaid balances over 30 days.		
STOUT INDUSTRIES, INC. 		

CONFIDENTIAL
THIS DOCUMENT AND ITS CONTENTS ARE BASED ON CURRENT COSTS AND ARE FOR IMMEDIATE ACCEPTANCE.
These prices will be updated periodically. First day of each, to reflect changes in material and labor costs.
These prices do not include shipping, insurance or handling charges. These prices do not include taxes. Taxes will be added by the home office of this company in St. Louis, MO. No contracts for delivery are completed upon sale, unless an order contract is made with the home office of this company in St. Louis, MO. No contracts for delivery are completed upon sale, unless an order contract is made with the home office of this company in St. Louis, MO. Any taxes or governmental charges which the seller may be required to pay on or collect again from the producer, sale, storage, or delivery, under any existing or future laws or regulations, shall be paid by the seller and will be added to the price of the article upon demand.
F.O.B. ST. LOIS, MO. MINIMUM FREIGHT: Minimum flat rate charged freight collect.
F.O.B. ST. LOIS, MO. MINIMUM FREIGHT: Minimum flat rate charged freight collect.
F.O.B. ST. LOIS, MO. MINIMUM FREIGHT: Minimum flat rate charged freight collect.

<https://guillaumejaume.github.io/FUNSD/>

レシート

STARBUCKS Store #10208
11302 Euclid Avenue
Cleveland, OH (216) 229-0749

CHK 664290
12/07/2014 06:43 PM
1912003 Drawer: 2 Reg: 2

Vt Pep Mocha	4.95
Starb Card	4.95
XXXXXXXXXXXX3228	
Subtotal	\$4.95
Total	\$4.95
Change Due	\$0.00

Check Closed
12/07/2014 06:43 PM

Starb Card x3228 New Balance: 37.45
Card is registered

<https://rrc.cvc.uab.es/?ch=13&com=tasks>

インフォグラフィック



<https://rrc.cvc.uab.es/?ch=17>

Webページ

2007 Ig Nobel Prize winners announced

Friday, October 5, 2007

The winners of the 2007 Ig Nobel Prize have been announced. The awards, given out every early October since 1991 by the *Annals of Improbable Research*, are a parody of the Nobel Prize, which are awards given out in several fields. The awards are given to achievements that, "first make people laugh, and then make them think." They were presented at Harvard University's Sanders Theater.

Ten awards have been presented, each given to a different field. The winners are:

- **Medicine:** Brian Witcombe, of Gloucestershire Royal NHS Foundation Trust, UK, and Dan Meyer, who studied the health consequences of sword swallowing.
- **Physics:** A team from the USA and Chile, who made a study about how cloth



The 2007 Ig Nobel Prize in aviation went to a team from an Argentinian university, who discovered that impotence drugs can help hamsters

https://en.wikinews.org/wiki/2007_Ig_Nobel_Prize_winners_announced

スライド

テキストベース質問応答（機械読解）

NTT

- テキストで記述された文書を知識源とし、質問に対して回答を行う
 - Input: 質問、テキストベースの文書
 - Output: 回答



テキストベースの文書



回答

言語処理AI

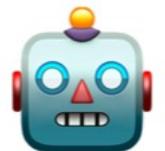
保險金のお支払い対象となる費用に関しては、当社の同意を得たうえで支出された費用に限りります。また、事故の相手が不明である場合など、**相手の方に法律上の損害賠償請求を行うことができないときは**、本特約は対象外となりますのでご注意ください。

など..

テキストベースの限界

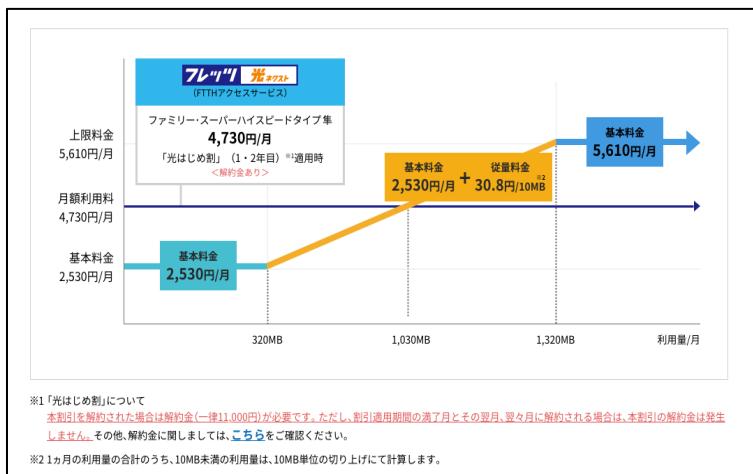
- BERTをはじめとする従来の読解モデルは、我々が普段扱う文書が持つ視覚的な情報を自然言語と併せて理解できない

図・表やグラフ、文字の見た目、配置等の視覚的な情報を読み取ることができない



言語処理AI

文書中のテキスト
情報のみを扱う



テキスト抽出
(OCR)

HTMLやPDF形式等の文書（契約書やマニュアル等）

フレッツ光ネクスト（FTTHアクセスサービス） ファミリー・スーパー・ハイスピードタイプ隼 上限料金 5610円/月 4730円/月
「光はじめ割」(1・2年目)※1 適用時 <解約金あり> 基本料金 2530円/月 320MB 基本料金
従量料金 2530円/月 30.8円 / 10MB 1030MB
1329MB 基本料金 5610円/月 利用量/月 「光はじめ割」について本割引を解約された場合は解約金（一律11,000円）が必要です。ただし、割引適用期間の満了月とその翌月、翌々月に解約される場合は、本割引の解約金は発生しません。その他、解約金に関しましては、こちらをご確認ください。1ヶ月の利用量の合計のうち、10MB未満の利用量は、10MB単位の切り上げにて計算します。

テキストデータ

文書を“視覚的に”読み解くAIを目指して

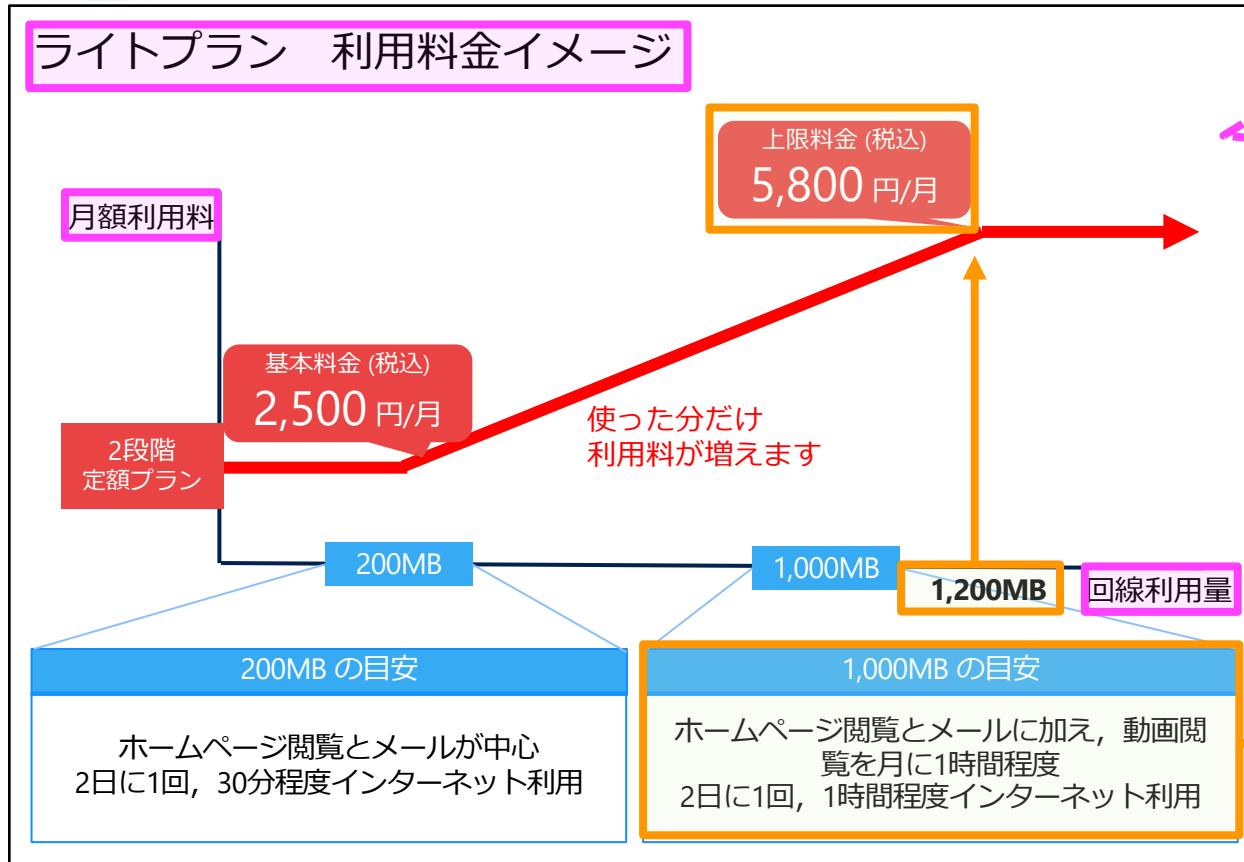
NTT 

- 人間が見る文書の情報をそのまま理解可能なAIの実現を目指す
- 言語情報、レイアウト、図やグラフ・表など多岐に渡る理解が必要



動画を毎日1時間は見るのですが、ライトプランの料金は月額幾らになりますか？

上限料金の税込月額5,800円になります



このページに答え
がありそうだな

動画を見るな
ら1200MB/月は
確実に超えそう

文書画像理解における一般的な解法

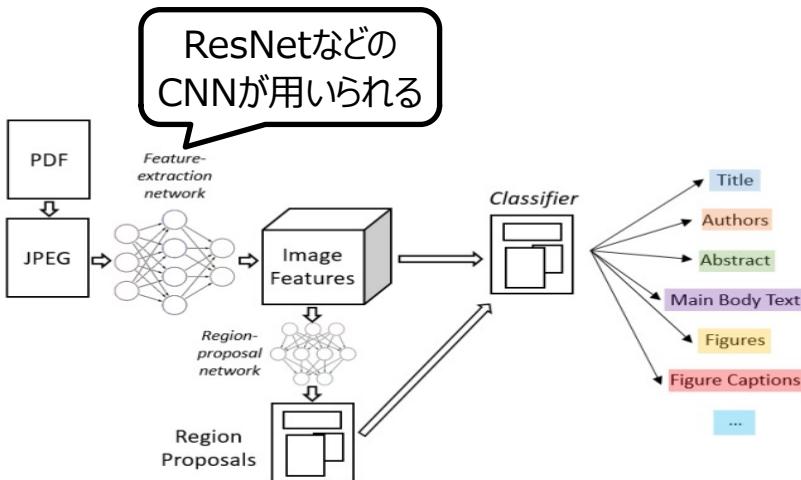
1. 文書レイアウト解析: 関心領域の検出および分類
2. OCR: テキスト領域の検出とテキストの認識
3. その他: 読み順予測, 一般物体認識など
4. 文書画像読解: 文書表現の獲得・理解

前処理
行わない／一部のみ行う
ケースがある



1. 文書レイアウト解析

- 文書画像内の関心領域を検出および分類するタスク
 - CNNベースのFaster-RCNNが一般的に用いられる
- 正解領域やラベルが自動的に入手可能なデータ (LaTeXソースやXML/HTMLを基に入手可能) を扱うことが多い



Faster-RCNNを用いたレイアウト解析
[Soto(Upton)+, EMNLP19]

<https://aclanthology.org/D19-1348/>

XML

```
<table-wrap id="tbl-003" orientation="portrait" position="float">
<label>Table 3.</label>
<caption>
  <p>Further hypercoagulability workup is summarized in Table 3</p>
</caption>
<alternatives>
  <graphic xlink:href="1757-1626-0002-000007170-007">/>
  <table frame="hsides" rules="groups">
    <caption>
      <p>Further hypercoagulability workup is summarized in Table 3</p>
    </caption>
    <thead>
      <tr>
        <th>Parameter</th>
        <th>Normal range</th>
      </tr>
    </thead>
    <tbody>
      <tr>
        <td>High-sensitivity fibrinopeptide B</td>
        <td>≤ 1.8</td>
      </tr>
      <tr>
        <td>Wintrobe's cell count</td>
        <td>10.0–11.0</td>
      </tr>
      <tr>
        <td>Neutrophils</td>
        <td>1.0%–1.5%</td>
      </tr>
      <tr>
        <td>Lymphocytes</td>
        <td>4.0%–6.0%</td>
      </tr>
      <tr>
        <td>Hemoglobin</td>
        <td>11.0–13.0 g/dL</td>
      </tr>
      <tr>
        <td>Platelets</td>
        <td>202–302 × 103/μL</td>
      </tr>
      <tr>
        <td>RDW</td>
        <td>14.4–14.8</td>
      </tr>
      <tr>
        <td>Prothrombin time</td>
        <td>12.2–13.5 seconds</td>
      </tr>
      <tr>
        <td>INR</td>
        <td>0.99–1.0</td>
      </tr>
      <tr>
        <td>PTT</td>
        <td>31 seconds (reference normal < 40)</td>
      </tr>
      <tr>
        <td>ANA</td>
        <td>Negative</td>
      </tr>
      <tr>
        <td>Antithrombin factor</td>
        <td>100% (normal)</td>
      </tr>
      <tr>
        <td>ESR</td>
        <td>10–20 (normal)</td>
      </tr>
      <tr>
        <td>Serum protein electrophoresis</td>
        <td>Normal</td>
      </tr>
      <tr>
        <td>Lipid profile</td>
        <td>Normal</td>
      </tr>
      <tr>
        <td>Hemoglobin electrophoresis</td>
        <td>Normal</td>
      </tr>
      <tr>
        <td>VTE RL</td>
        <td>Negative</td>
      </tr>
      <tr>
        <td>PTA-ABG</td>
        <td>Negative</td>
      </tr>
    </tbody>
  </table>
</div>
```

XMLのパースにより得られたデータ

Cases Journal 2009, 2:710
<http://casesjournal.com/casesjournal/article/view/710>

Table 2. Laboratory tests include hemostatic/ability, coagulation	
Coagulation	90
High-sensitivity fibrinopeptide B	1.8
Wintrobe's cell count	10
Neutrophils	8.0%
Lymphocytes	1.0%
Hemoglobin	11.0
Platelets	202
RDW	14.4
Prothrombin time	12.2
INR	0.99
PTT	31 (reference normal < 40)
ANA	Negative
Antithrombin factor	100% (normal)
ESR	10 (normal)
Serum protein electrophoresis	Normal
Lipid profile	Normal
Hemoglobin electrophoresis	Normal
VTE RL	Negative
PTA-ABG	Negative

Figure 3. Fluorescent Angiogram of the right eye shows blocked venous fluorescence from the retinal hemorrhage, extensive areas of capillary non-perfusion, and vessel wall leakage.

Figure 4. An example of acute hypertension retinopathy, which is one of the differentials for CRVO. Figures showing arteriovenous nicking, copper wire arterial changes, hemorrhages, cotton wool spots, disc edema bilaterally (left more than right), and macula that dominate in the capillary area.

Table 3 of 4
(page number not for citation purposes)

Publaynet [Xu(IBM)+, ICDAR19]

<https://arxiv.org/abs/1908.07836>

53

2. OCR (光学文字認識)

- テキスト領域の検出および領域内のテキストを認識
 - OCRはオープンソースツールを使うことが一般的。フリーで使えるものとしては、Tesseractが有名
 - 回転・湾曲・手書き文字、低資源言語の認識精度は比較的低い

テキスト

2007, Ig, Nobel, Prize, Winners, announced

テキスト領域: Bounding box

2007	Ig	Nobel	Prize	winners	announced	
------	----	-------	-------	---------	-----------	--

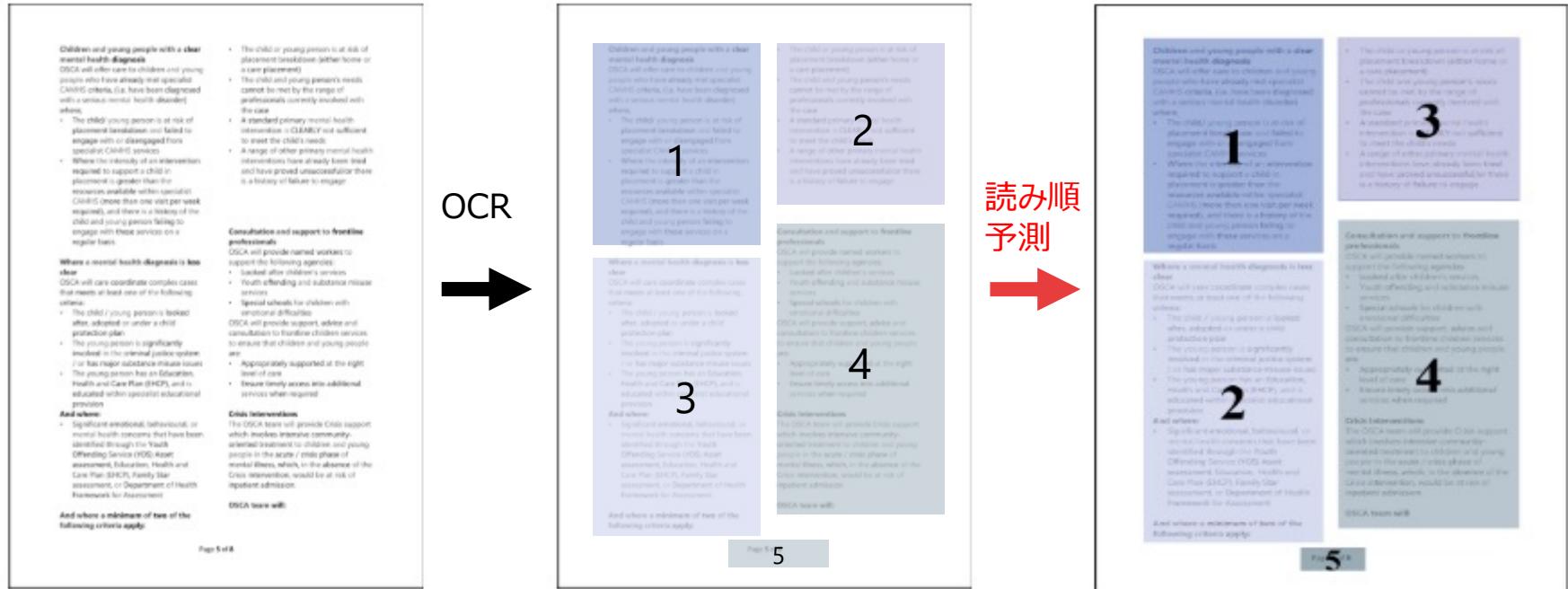
Friday, October 5, 2007

The winners of the 2007 Ig Nobel Prize have been announced. The awards, given out every early October since 1991 by the Annals of Improbable Research, are a parody of the Nobel Prize, which are awards given out in several fields. The awards are given to achievements that "first make people laugh and then make them



3. その他: 読み順予測 [Wang(UCS)+, EMNLP21] NTT

- OCRで出力される系列を、人間が読むような系列順に修正
 - 通常、OCR系列はLeft-to-right Top-to-downで出力されるため、複数カラムやテキストが複雑に配置されている場合、読み順を誤る

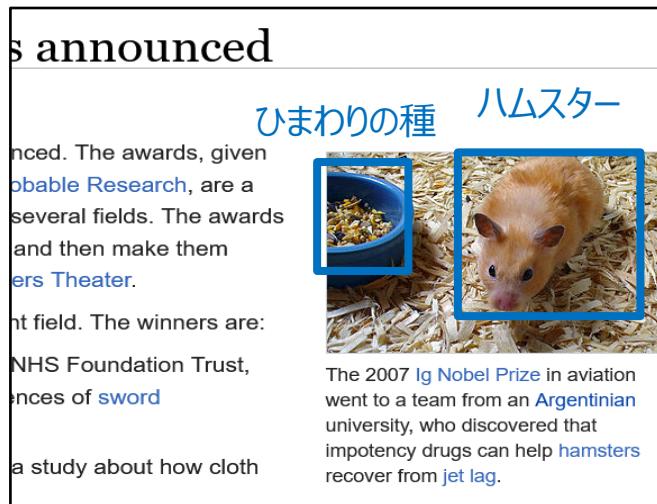


ReadingBank dataset

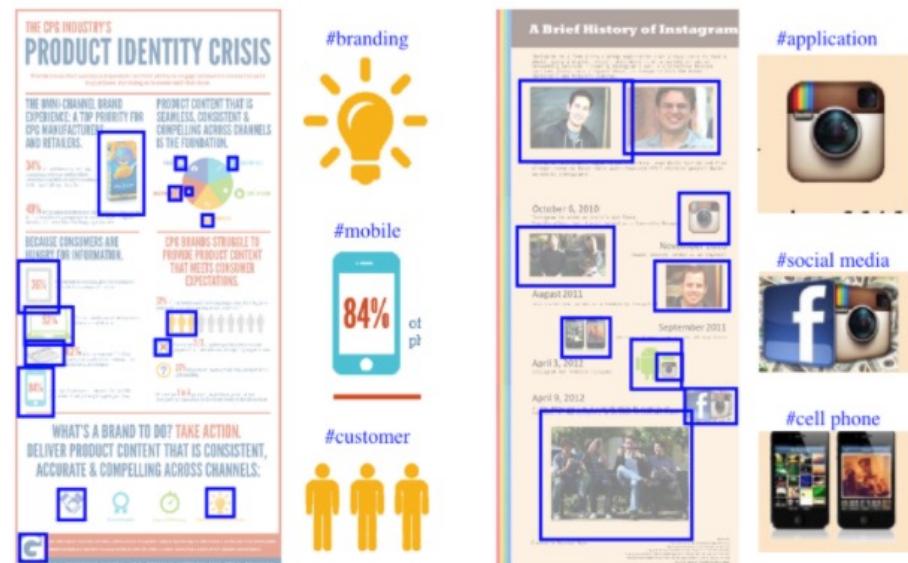
<https://aclanthology.org/2021.emnlp-main.389/>

3. その他: 一般物体認識

- 文書に含まれる一般物体(写真中の物体やアイコンなど)の領域検出とカテゴリを認識
 - 文書レイアウト解析と同様にFaster-RCNNなどの物体検出器を用いて、領域と意味を出力
 - 近年の文書画像QAデータセットは、一般物体についても問われる



Webページのスクショに対する一般物体検出



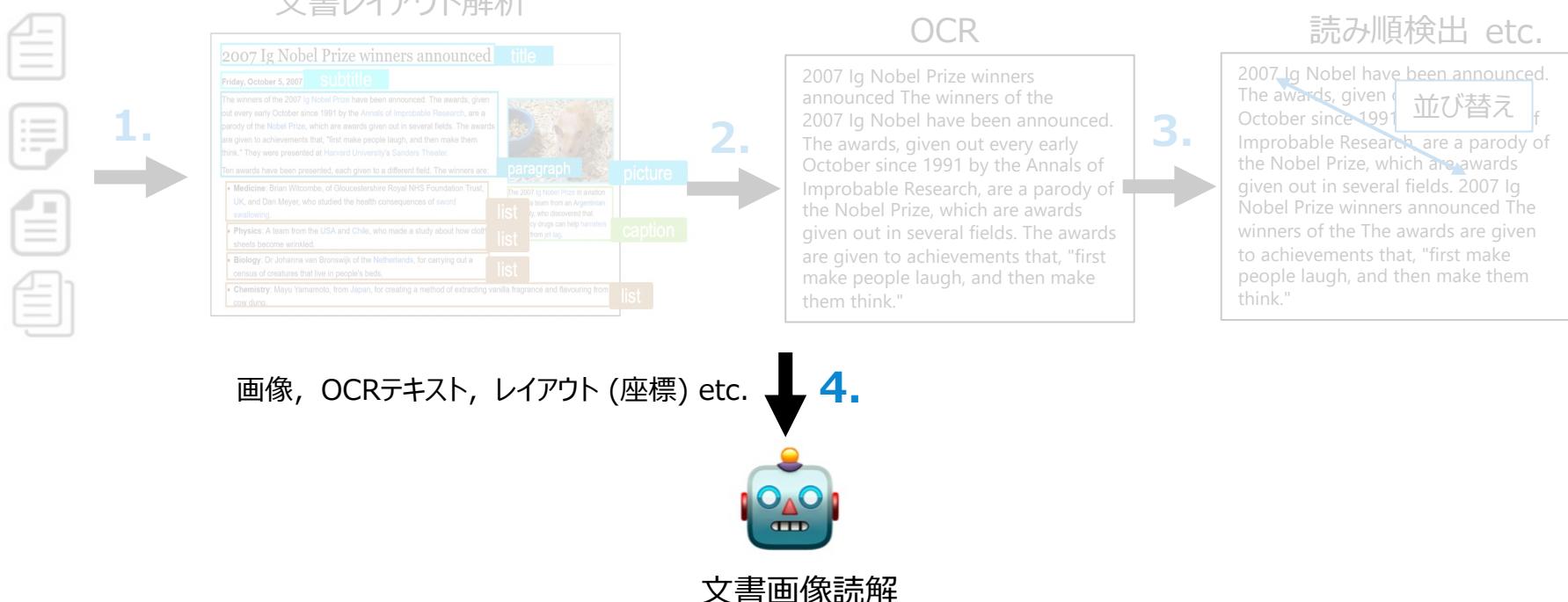
インフォグラフィックに対するアイコン検出

<http://visdata.mit.edu/>

文書画像理解における一般的な解法

1. 文書レイアウト解析: 関心領域の検出および分類
2. OCR: テキスト領域の検出とテキストの認識
3. その他: 読み順予測, 一般物体認識など
4. 文書画像読解: 文書表現の獲得・理解

前処理
行わない/一部のみ行う
ケースがある



文書画像読解タスク

シーケンスラベリング [Jaume(ETH)+, ICDAR-OST'19] NTT

- 文書画像中の全てのテキスト領域に対して、KeyやValueとなるラベルなどを付与するタスク
 - 例: Key=Supplier, Value=AUDIENCE STUDIES

MR 1909 (3-69) 100

BROWN & WILLIAMSON TOBACCO CORPORATION
FILTER SCORES

Brands: RALSTON (RELAIR portion not tested) **Project #:** 1969-105

Commercial: LAKE - NEW PACK :40 (with RELAIR Badain-
ton :20) **Sample:** 336 **PMS Base:** (234)

Code #: RW-KT-69-98

Supplier: AUDIENCE STUDIES

TEST DATES:

L. Angeles: 8/5 and 6
Chicago: 8/8

PMS SCORES

Overall	1.7
CITY	0.0
Los Angeles	3.3
Chicago	3.3
SEX	0.0
Hale	3.3
Female	3.3
AGE	0.0
26-35	0.0
36-45	0.0
46 & Over	9.3
35 & Under	0.0
35 & Over	5.0

COMMENTS: This commercial was tested in color.

465607116 p

SPORTS MARKETING ENTERPRISES
DOCUMENT CLEARANCE SHEET

Date Routed:	January 11, 1994	Contract No.	4011 00 00
Contract Subject:	Joe's Place Exhibits		
Company	SPEVCO, INC.	Brand(s)	Camel/Winston
Total Contract Cost	\$1,340,000.00	Current Year Cost	1994-1995
Brief Description	2 Joe's Place Exhibits for use at Winston Cup, Winston Drag and Camel Super Bike Events.		
G/L Code:	Program Budget Code	NAME	SIGNATURE
Originator:	Michael Wright	DATE	
Manager:	John Powell	B. J. Powell	1-11-94
REVIEW ROUTING	SIGNATURE	DATE	
Insurance			
Law			
FS - Marketing			
REVISIONS TO SHELL (Other than Term, Compensation or Job)	PAGE(S)	SECTION(S)	
APPROVAL ROUTING			
* Sr. Manager (B. J. Powell)			
* Director - (G. L. Littell)			
** Sr. VP T. W. Robertson			
Return To:	MARY SEAGRAVES Ext. 1485	SME	13 Plaza

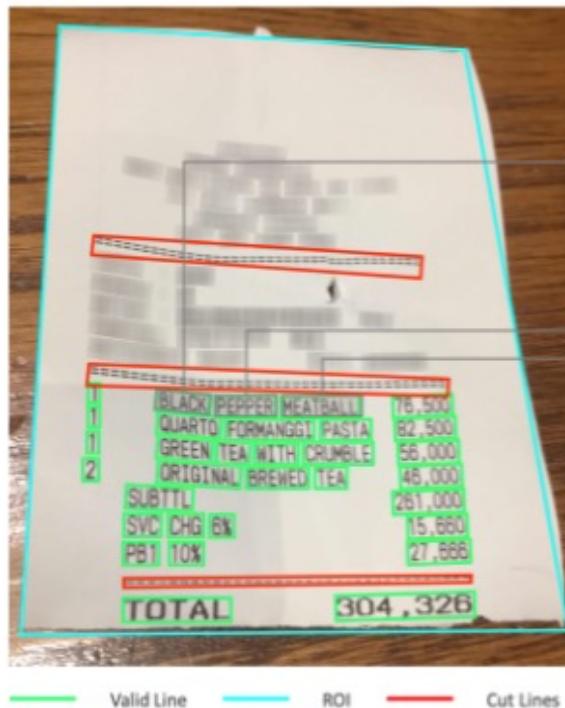
* UP TO AND INCLUDING \$25,000
** OVER \$25,000

Revised 10/26/92

51569 8130

情報抽出 [Park(NAVER)+, NeurIPS workshop'19]

- 文書画像から特定の情報 (menu, priceなど) を抽出するタスク
 - 例: menu=BLACK PAPPER MEATBALL



Image

```
...
  "words": [
    {
      "quad": {"x2": 837, "y3": 2260, "x3": 829, "y4": 2238,
               "x1": 625, "y1": 2156, "x4": 617, "y2": 2177},
      "is_key": 0,
      "row_id": 1631316,
      "text": "BLACK"
    },
    {
      "quad": {"x2": 1105, "y3": 2261, "x3": 1105, "y4": 2261,
               "x1": 857, "y1": 2180, "x4": 857, "y2": 2180},
      "is_key": 0,
      "row_id": 1631316,
      "text": "PEPPER"
    },
    {
      "quad": {"x2": 1441, "y3": 2264, "x3": 1443, "y4": 2269,
               "x1": 1125, "y1": 2184, "x4": 1126, "y2": 2179},
      "is_key": 0,
      "row_id": 1631316,
      "text": "MEATBALL"
    }
  ],
  "category": "menu_nm",
  "group_id": 3
}
...
```

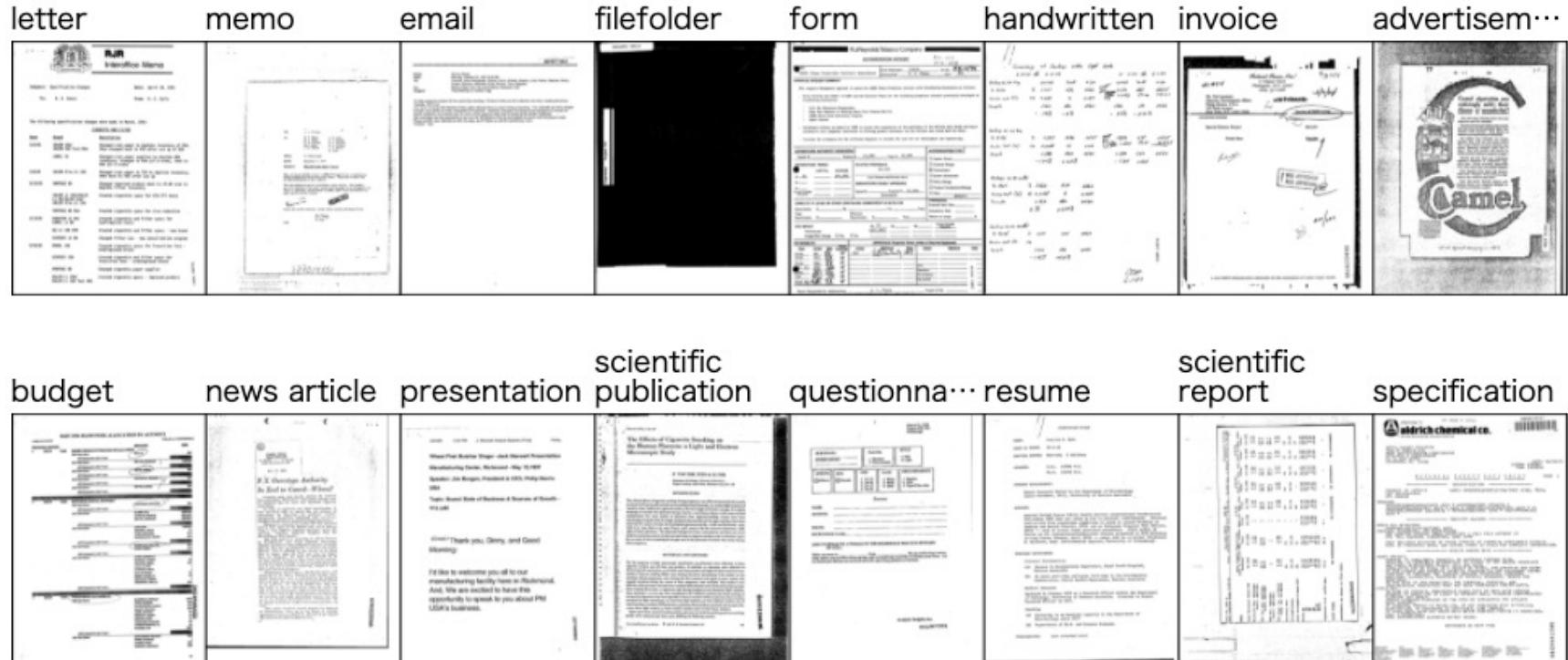
JSON

<https://github.com/clovaai/cord>

文書画像分類 [Adam(Ryerson Univ.)+, ICDAR'15]

NTT

- 文書画像を複数のクラスに仕分けるタスク
 - RVL-CDIPでは、16クラス、40万サンプル用意されている
 - 最新のモデルだと、95%程度の精度で識別可能

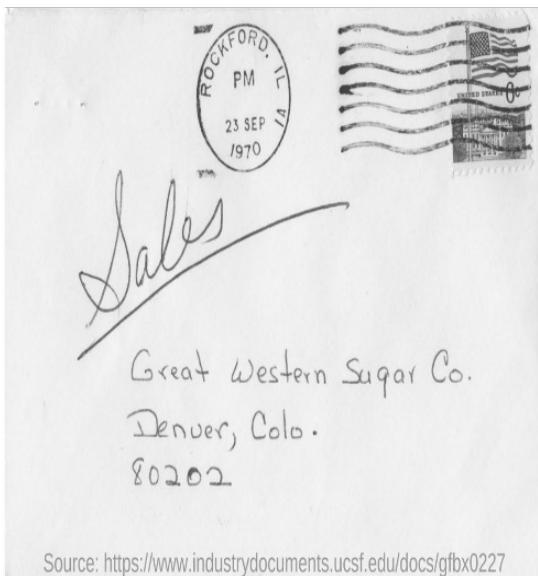


<https://www.cs.cmu.edu/~aharley/rvl-cdip/>

視覚的読解

- 文書画像に関する質問に対して、回答テキストを文書内のテキストから抽出 or 自由な形式で生成するタスク
 - 一般物体(アイコンなど)や図表理解、算術演算が課題

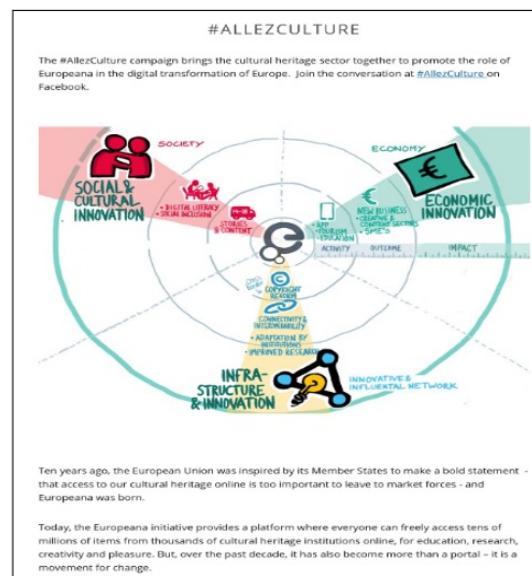
DocVQA [Mathew+, 21]



Q: Mention the ZIP code written?

A: 80202

VisualMRC [Tanaka&Nishida+, 21]

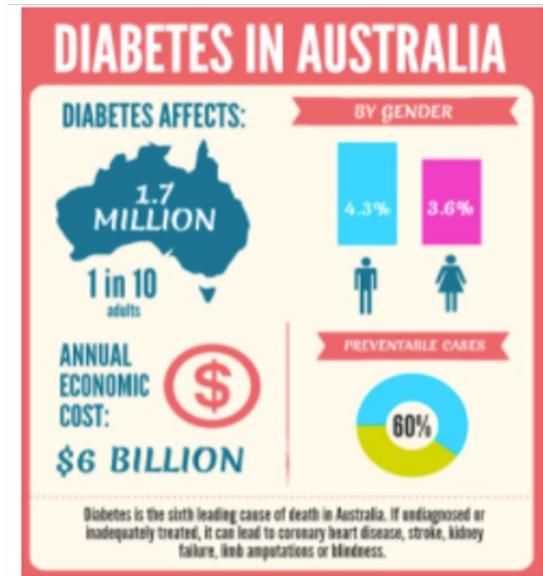


Q: In the graphic, what is the color of economy?

A: the color is green

<https://arxiv.org/abs/2007.00398>

InfographicVQA [Mathew+, 22]



Q: How many females are affected by diabetes?

A: 3.6%

<https://arxiv.org/abs/2104.12756>

- Webページのスクリーンショットに基づく視覚的読解
- レイアウト理解と生成型機械読解を統合した問題設定とモデルの提案

2007 Ig Nobel Prize winners announced

Friday, October 5, 2007

The winners of the 2007 Ig Nobel Prize have been announced. The awards, given out every early October since 1991 by the Annals of Improbable Research, are a parody of the Nobel Prize, which are awards given out in several fields. The awards are given to achievements that, "first make people laugh, and then make them think." They were presented at Harvard University's Sanders Theater.

Ten awards have been presented, each given to a different field. The winners are:

- **Medicine:** Brian Witcombe, of Gloucestershire Royal NHS Foundation Trust, UK, and Dan Meyer, who studied the health consequences of sword swallowing.
- **Physics:** A team from the USA and Chile, who made a study about how cloth sheets become wrinkled.
- **Biology:** Dr Johanna van Bronswijk of the Netherlands, for carrying out a census of creatures that live in people's beds.
- **Chemistry:** Mayu Yamamoto, from Japan, for creating a method of extracting vanilla fragrance and flavouring from cow dung.



The 2007 Ig Nobel Prize in aviation went to a team from an Argentinian university, who discovered that impotency drugs can help hamsters recover from jet lag.

約10,000枚の画像を収集

全ての領域をアノテーション：
9つの意味クラスに分類

Q: Who were the winners of the Ig Nobel prize for Biology and Chemistry?

A: The winner of the Ig Nobel prize for biology was Dr Johanna van Bronswijk, and the winner for Chemistry was Mayu Yamamoto.

約30,000件の生成型QAを作成

SlideVQA [Tanaka(NTT)+, AAAI'23]

- 複数枚の画像集合 (プレゼン資料) に跨る質問応答タスク
- 演算が必要な回答に対して、演算過程 (算術式) を人手で付与
- 画像選択と質問応答タスクを統合的に解くモデルを提案



画面指示遂行

- モバイル/PC画面に対して、自然言語による指示を行い、エージェントが画面操作を行う
- 画面遷移を伴う指示遂行データセットが多く取り組まれており、指示が実行可能か否かを問うタスクも存在する

Clicking on a canvas

Create a line that bisects the angle evenly in two, then press submit.

Dragging

Drag Nisse right by one.

Yolane	Bertie	Meaghan
Margie	Constanta	Gusta
Sibella	Nisse	Julieta

Multi-step

Use the textbox to enter "Tula" and press "Search", then find and click the 5th search result.

Tula

Annis
https://www.et.hk
Sit interdum.
Cierra
https://www.sempervulputate.gov
Semper mi enim.
Annis
https://www.bibendumauktor.se
Felis odio tellus.

1 2 3 ≥

Task

Open settings and clear search history

App Environment

Spotify

Settings

ohgo93nvdctg55v6w...
Delete cache
Data Saver
Playback
Crossfade
Shuffle
Autoplay
Allow Explicit Content
Show unplayable songs
Third-party software
Terms and Conditions

Follow Up Questions

Do you mean "open settings and clear cache"?
Do you mean delete cache?
Should this be 'go to settings and delete cache'?

キーとマウスの操作を行って、目的のタスクを遂行

操作が実行可能な場合は、Follow-up質問を出力

MiniWob++: Webブラウザに対する指示遂行タスク [Peter(DeepMind)+, ICML'22]

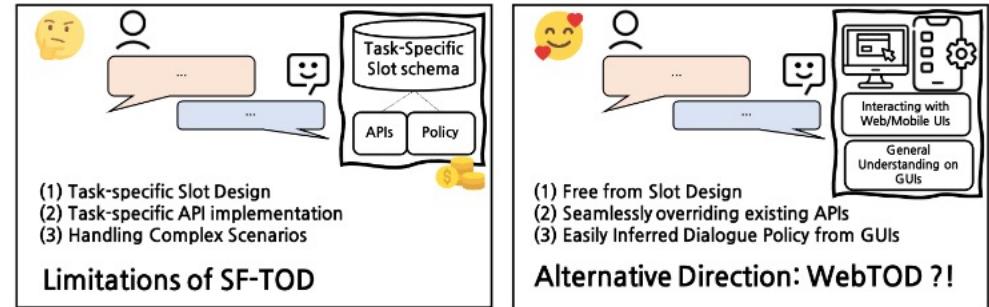
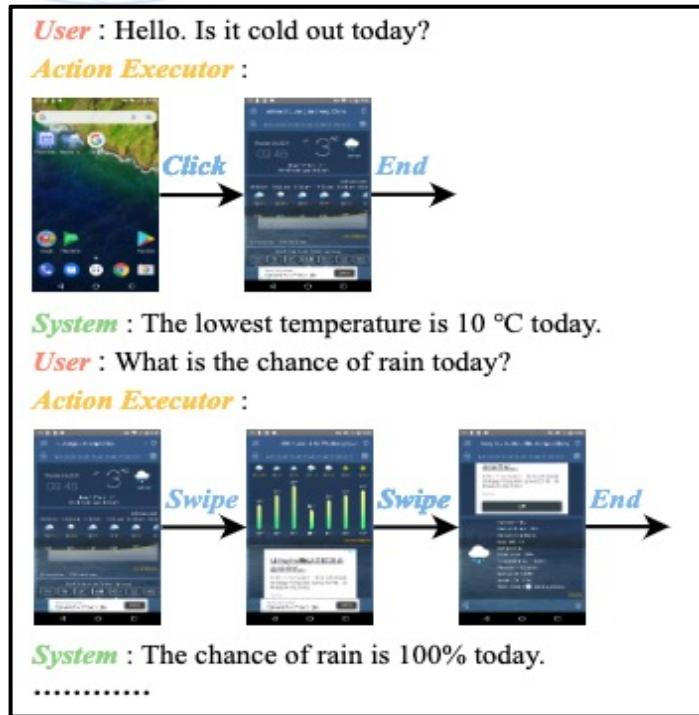
<https://proceedings.mlr.press/v162/humphreys22a/humphreys22a.pdf>

MOTIF: モバイルアプリに対する指示遂行タスク [Burns(Boston Univ.)+, ECCV'22]

https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136680304.pdf

画面指示遂行 + 対話

- モバイル/PC画面に対して、自然言語による指示(対話)を行い、画面操作と返答を返す
- 大規模な事前学習データを取集することが困難
→ 汎用的な画面に対する理解・操作が課題



WebTOD: GUIを用いたタスク指向対話のコンセプト [Lee(NAVER)+, arXiv'22]

<https://arxiv.org/abs/2212.10504>

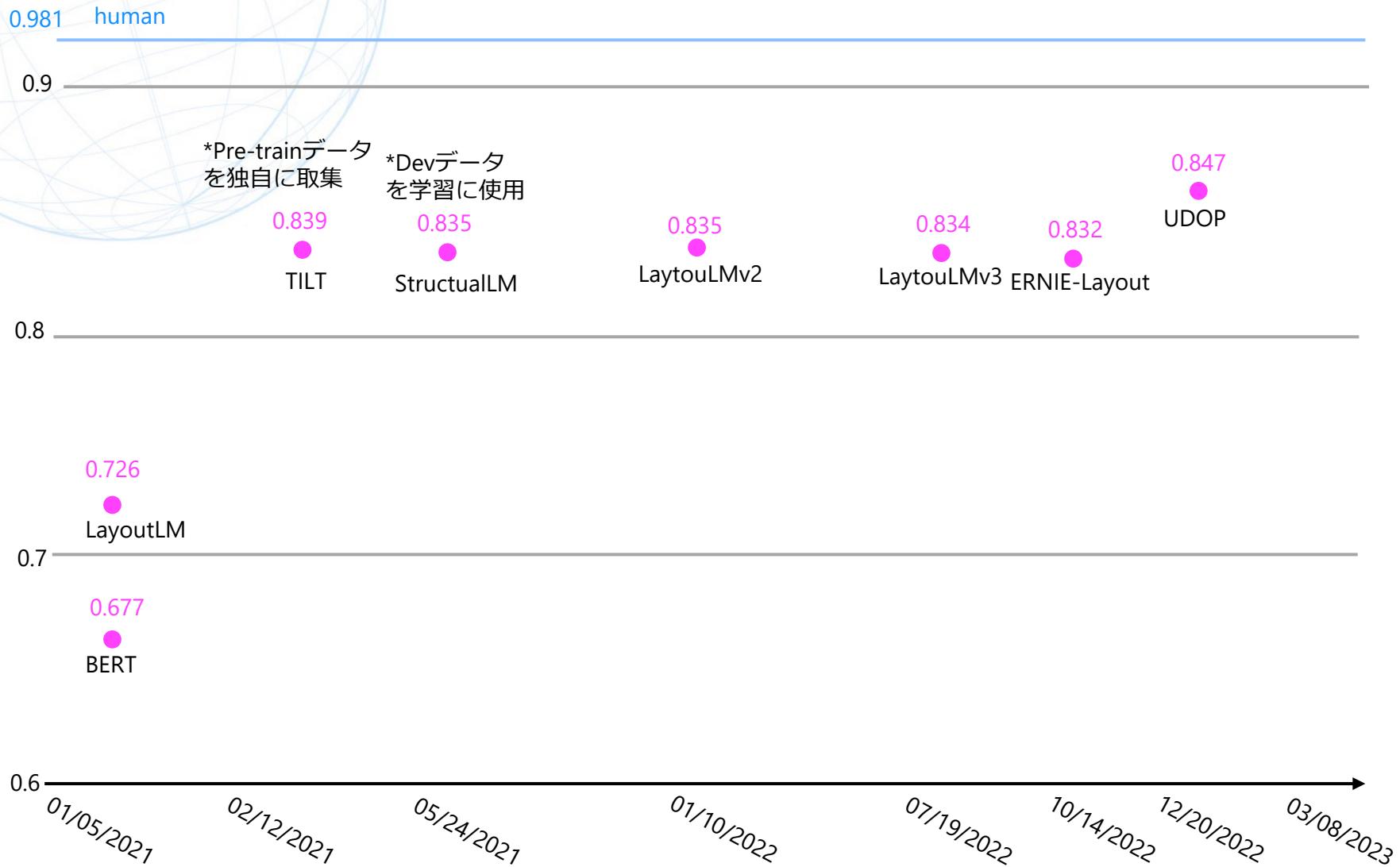
Meta-GUI: モバイルGUIに対するタスク指向対話 [Sun(X-LANCE)+, EMNLP'22]

<https://aclanthology.org/2022.emnlp-main.449.pdf>

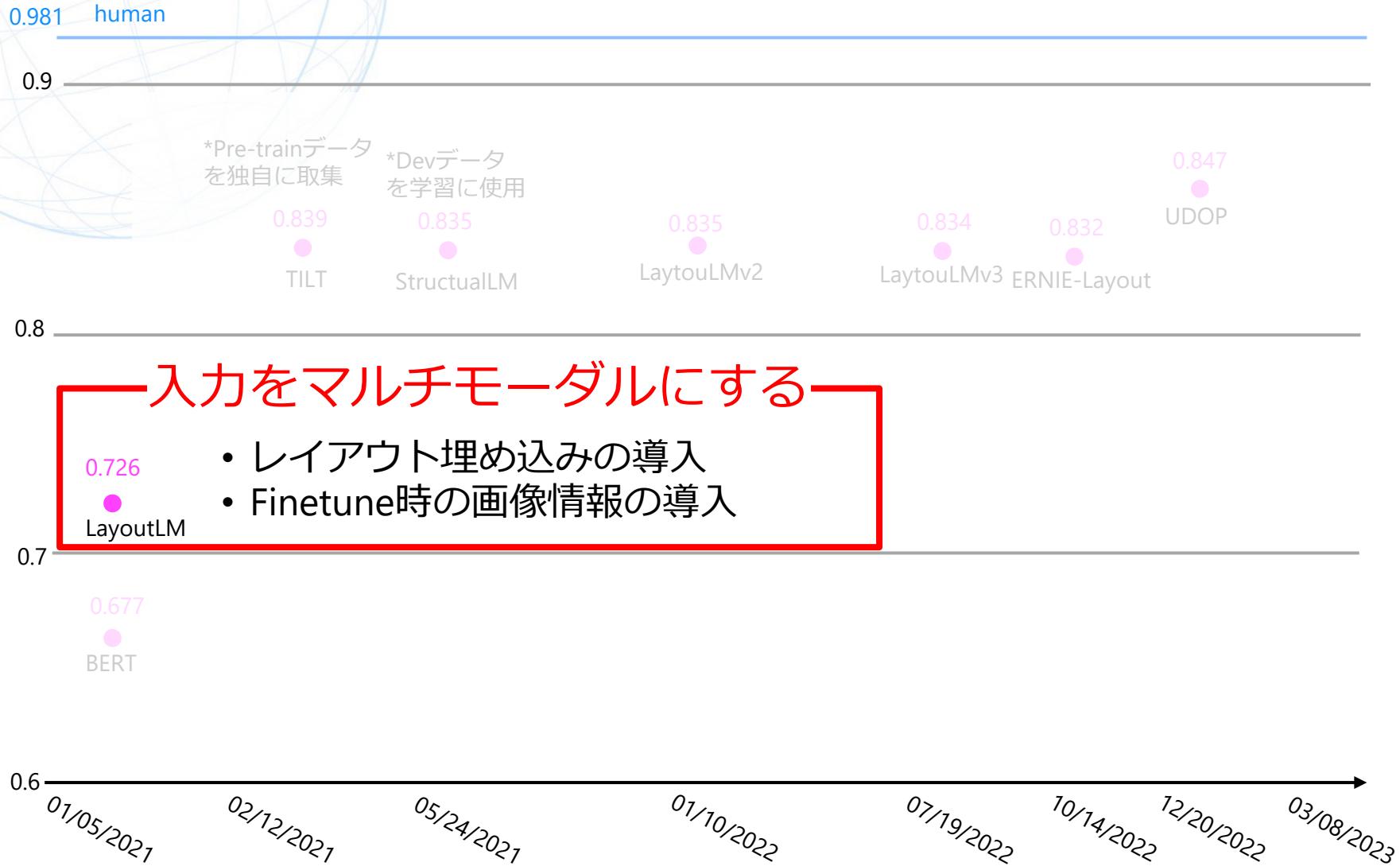
文書画像読解モデル

DocVQAにおける性能の推移

- 文書レイアウト理解を問うDocVQAにおいて、人間のスコアに迫るモデルが登場

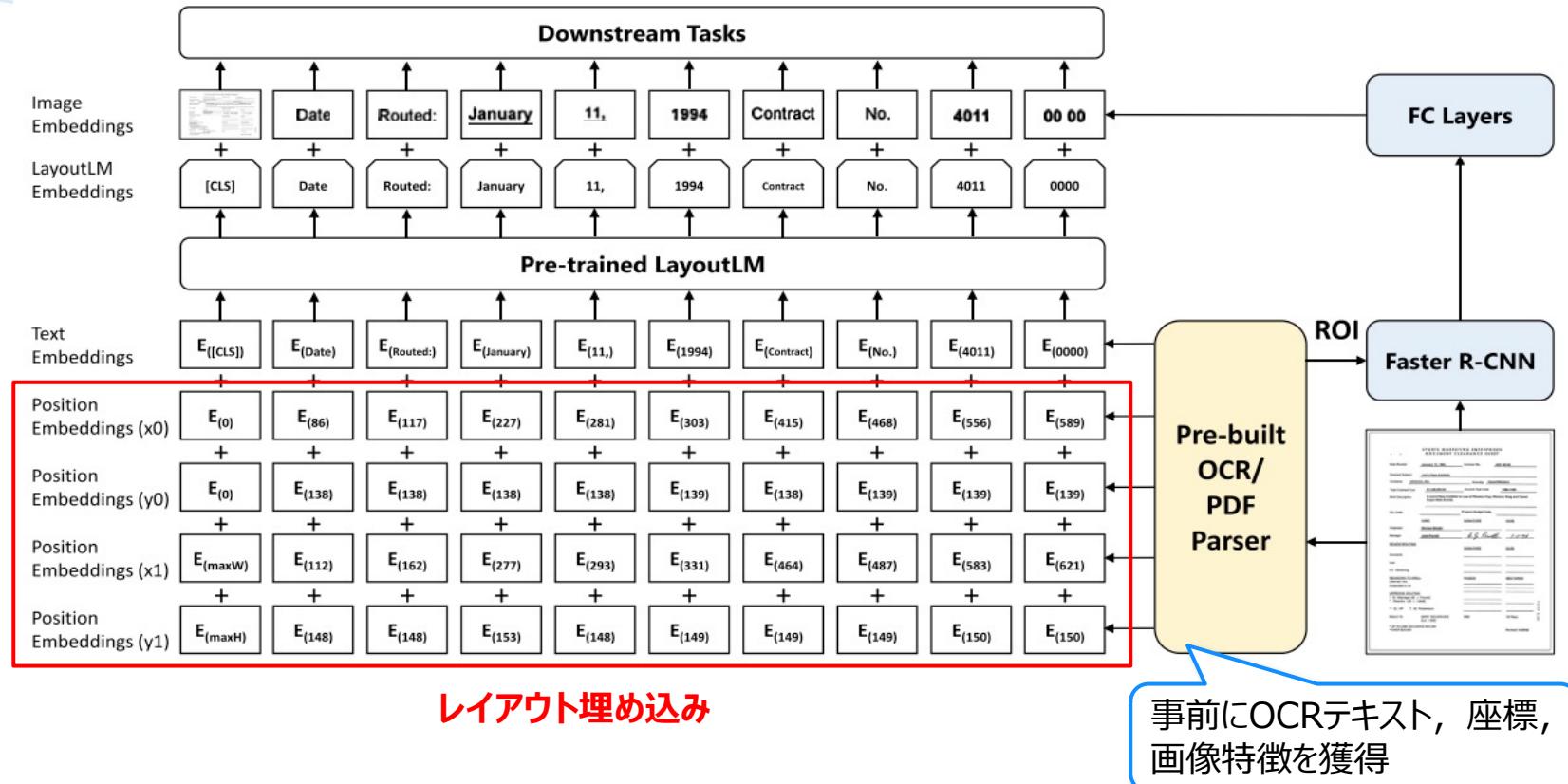


DocVQAにおける性能の推移

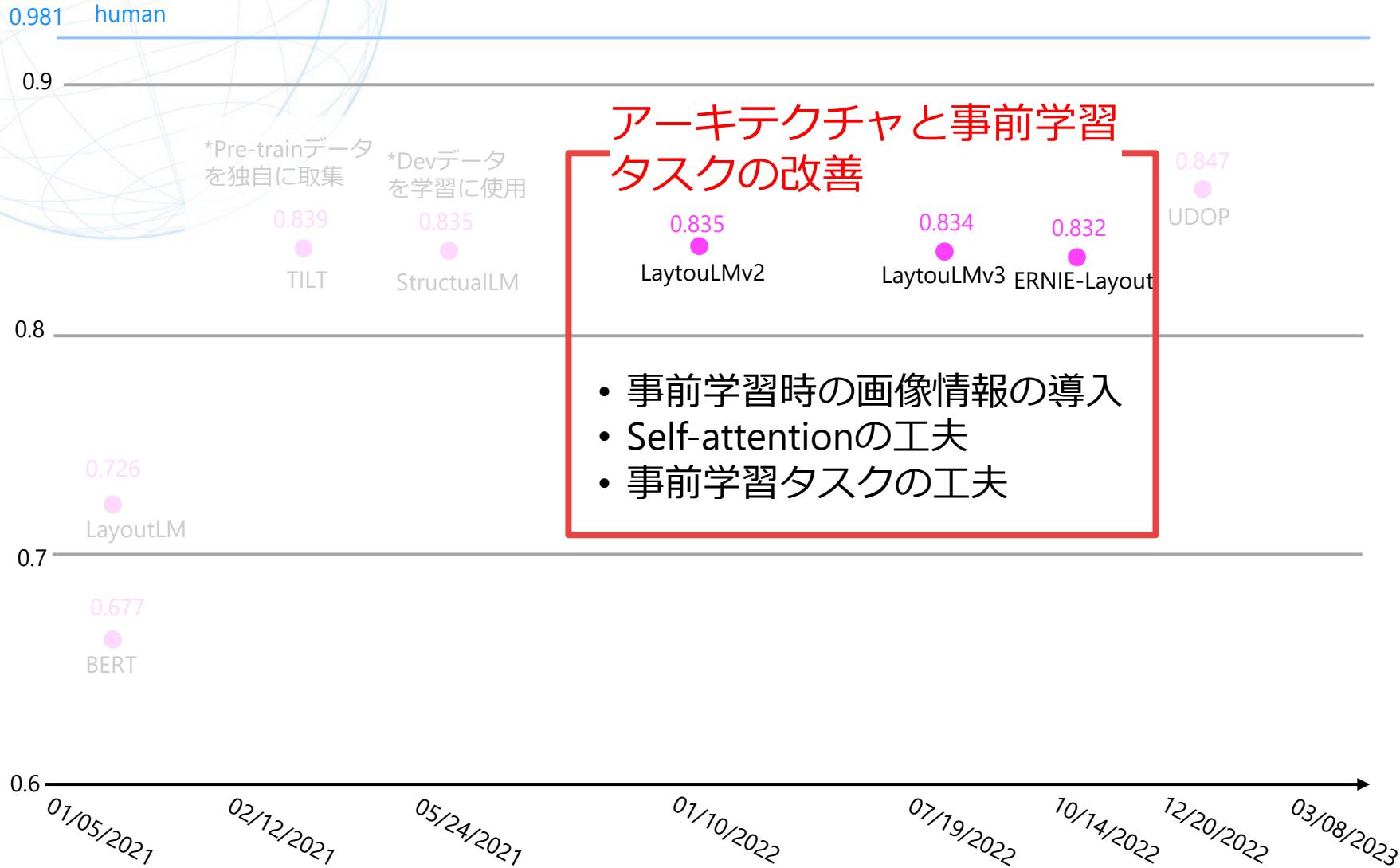


LayoutLM [Xu(HIT)+, KDD'20]

- 入力テキスト埋め込みに対して、テキスト領域の座標を学習可能なパラメータ(レイアウト埋め込み)として追加し、穴埋め事前学習を実施
- Finetune時ののみOCR画像特徴をLayoutLMの出力に足し合わせて、最終出力として使用する

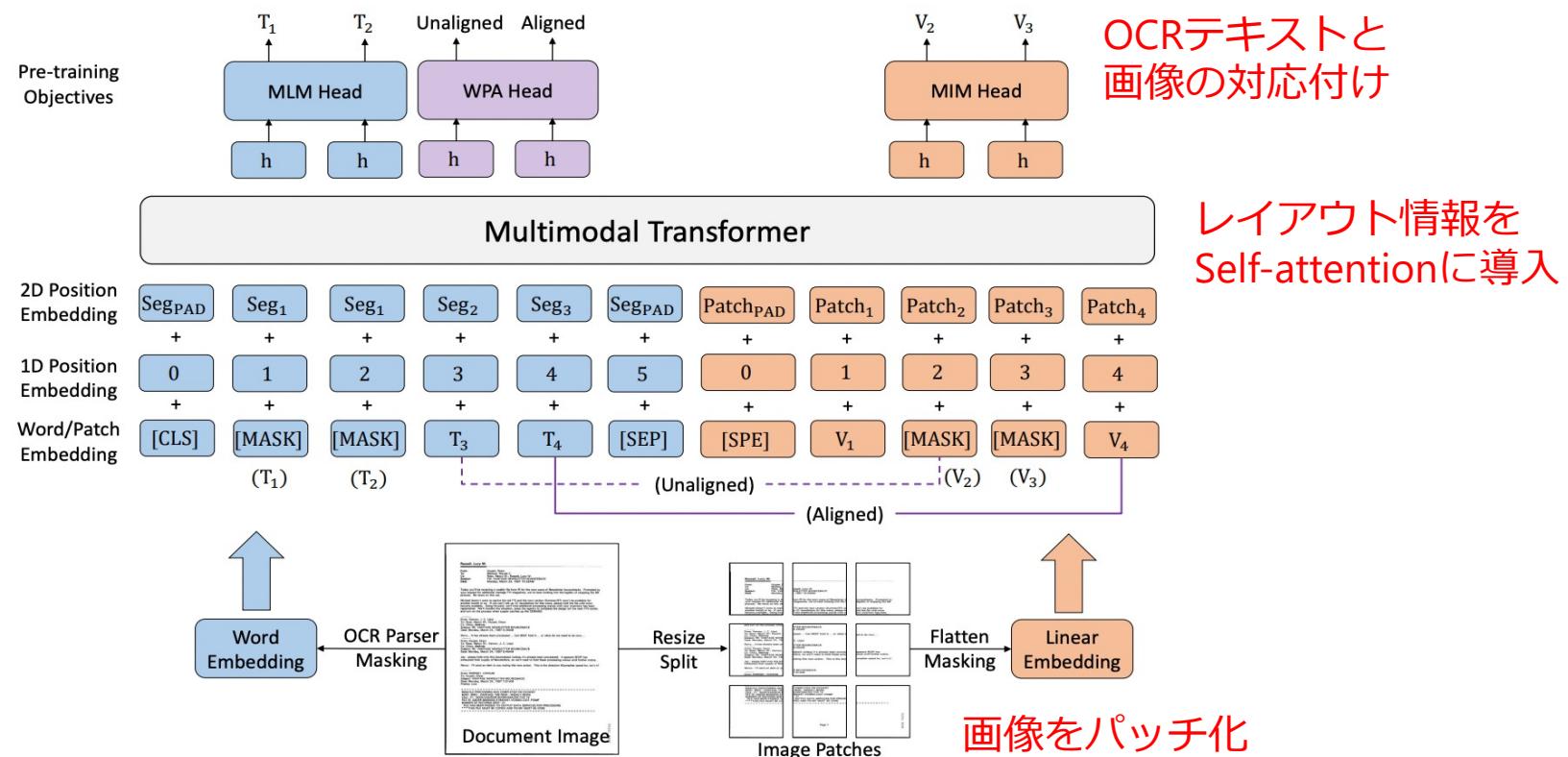


DocVQAにおける性能の推移



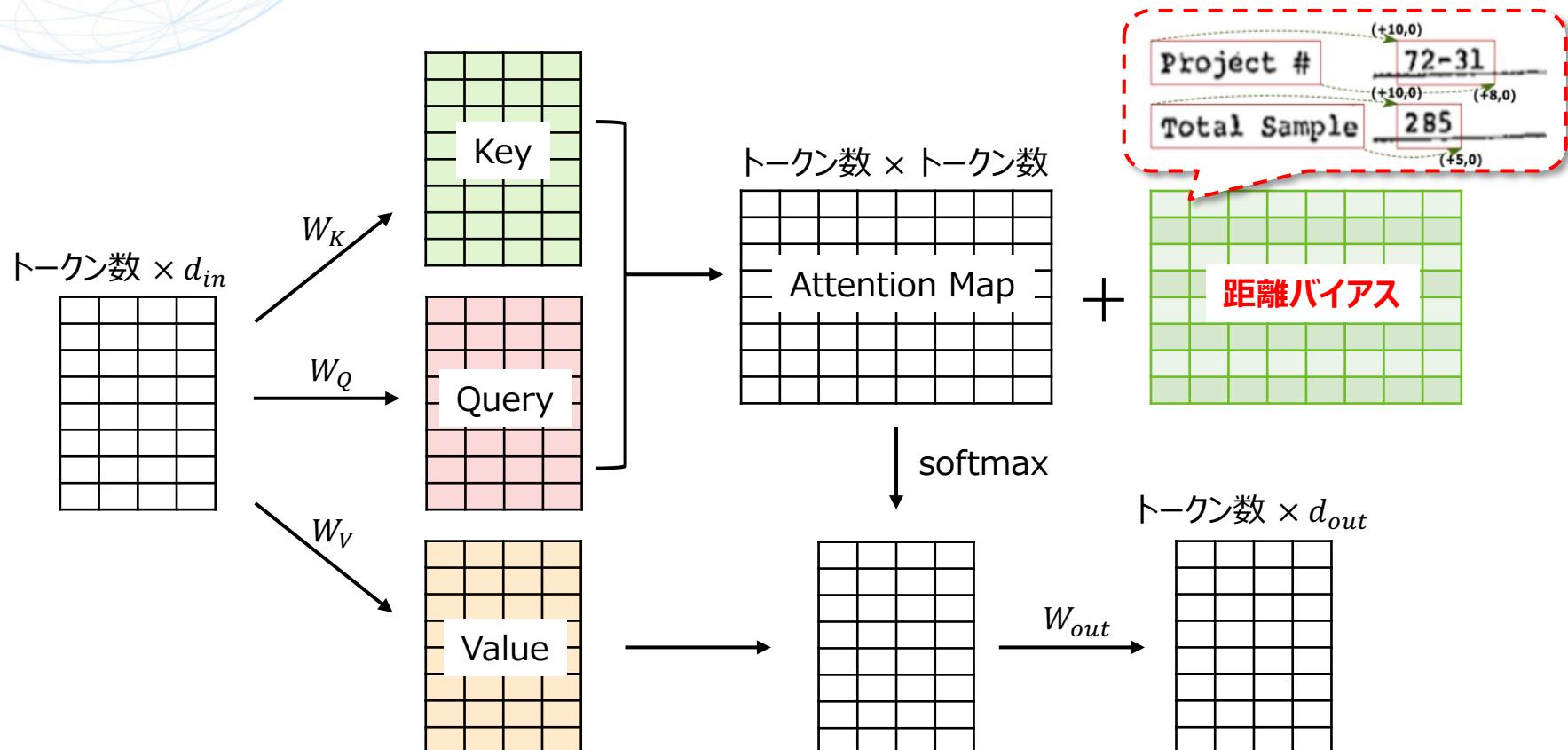
LayoutLMv3 [Huang(Sun Yat-sen Univ.)+, ACMM'22]

- 文書画像をパッチ分割し、ViTを用いてエンコーティング
- 座標間距離を考慮したレイアウト情報をSelf-attentionに導入
- 穴埋め復元と単語-パッチ対応付け学習により、OCRテキストと文書画像の対応を事前学習



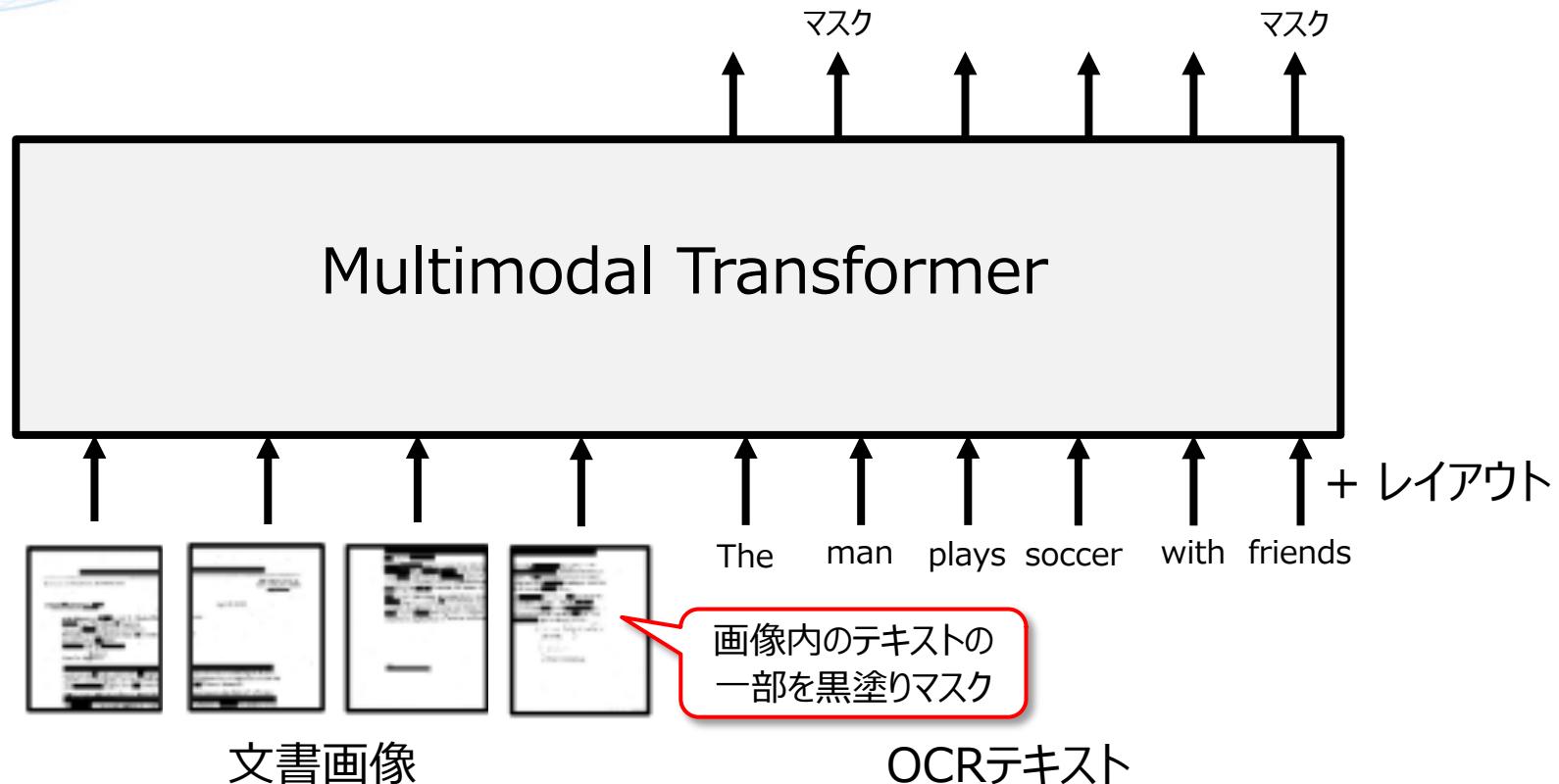
レイアウト情報を考慮したSelf-attention

- テキストの座標間距離をSelf-attentionのバイアス項に追加
 - 座標間の距離を直接与えることで、座標の距離が近いものがKey-Valueの関係性にあるなどの学習が効率的に進む

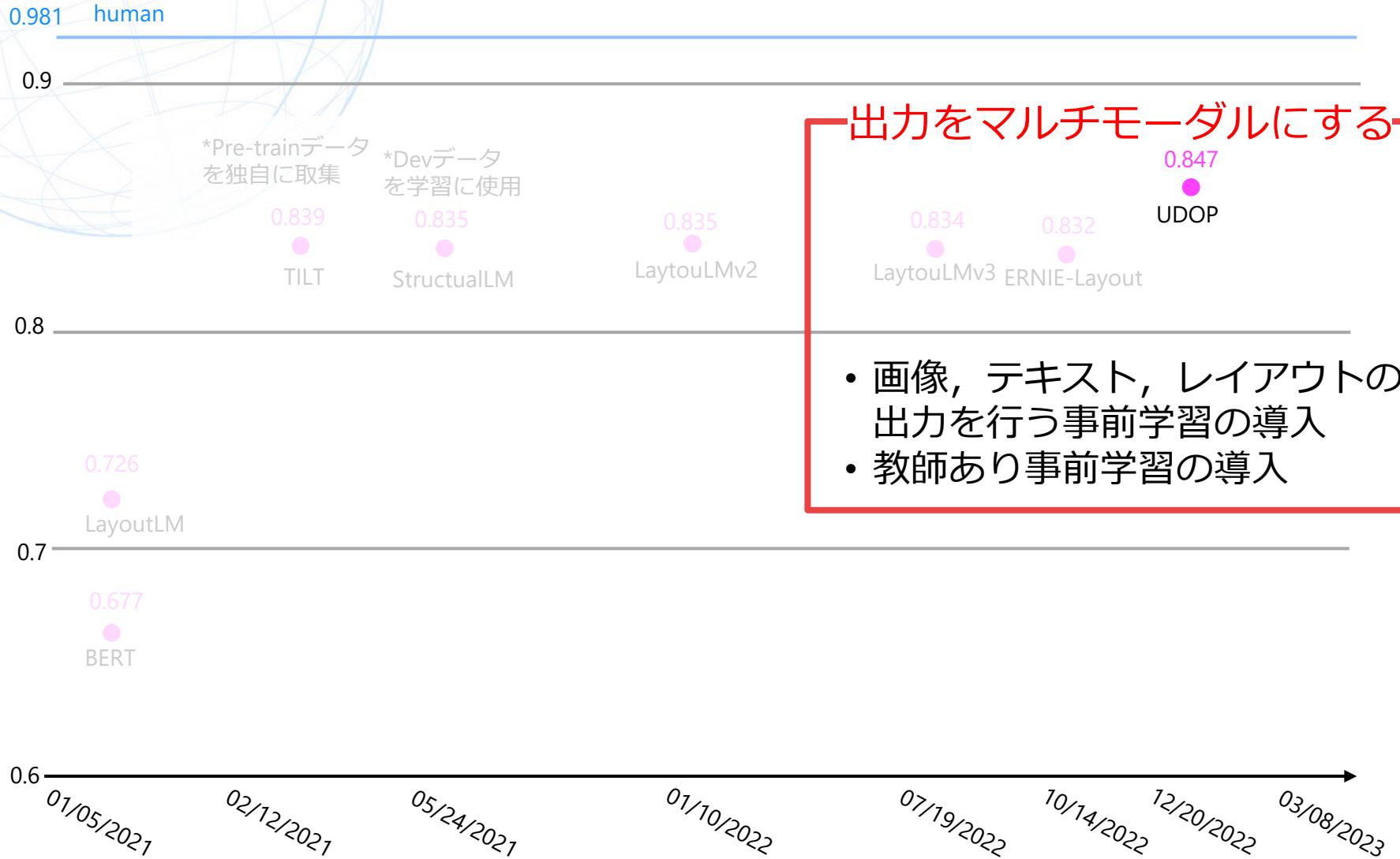


単語-パッチ対応付け学習 (WPA)

- 文書画像のテキスト部分を一部マスクして、OCRテキストのどの部分がマスクされているかを予測することで、文書画像とOCRテキストの対応を学習

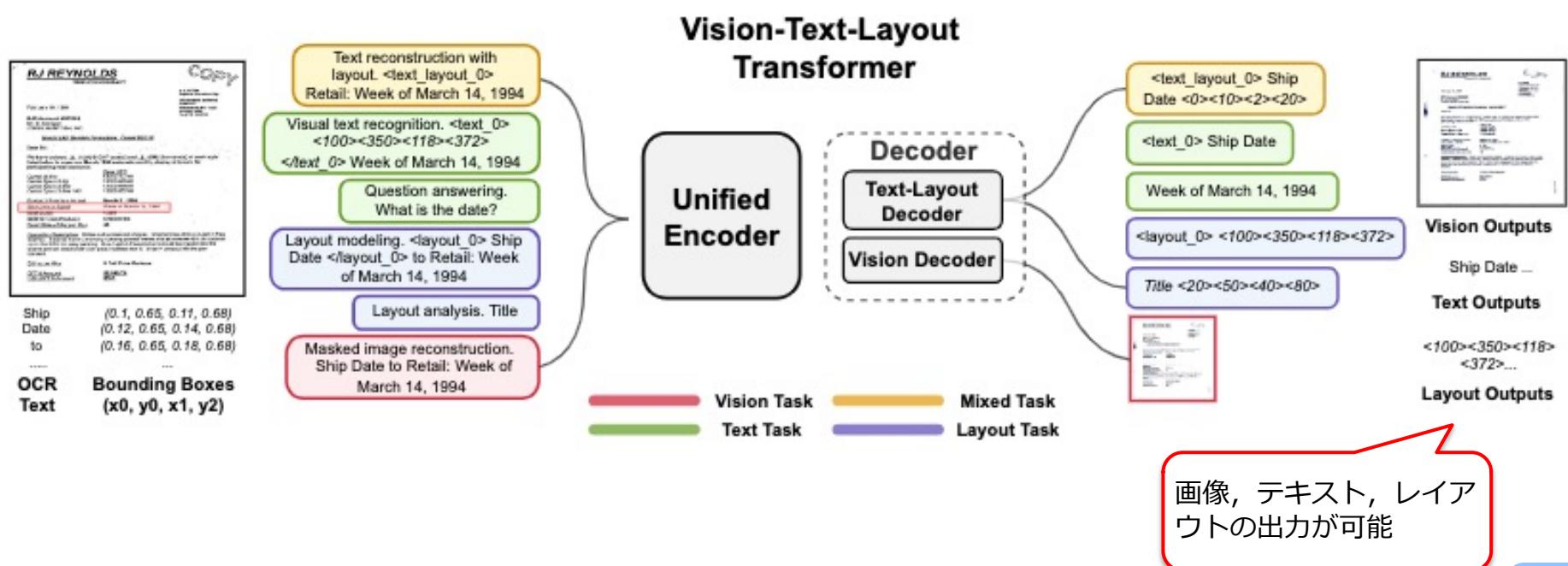


DocVQAにおける性能の推移



UDOP [Tang(UNC)+, arXiv'22]

- Decoderで画像, テキスト, レイアウト(座標)の出力を行う自己教師あり事前学習を実施
- 情報抽出やQAタスクなどの5つのタスク全てを系列変換の枠組みで教師あり事前学習を実施
- 9つの文書画像読解タスクでstate-of-the-artを達成



- **Layout Modeling:** スパン領域の座標を予測
- **Visual Text Recognition:** スパンマスク領域のテキスト予測
- **Joint Text-Layout Recognition:** スパンマスク領域のテキストと座標を予測
- **Masked Image Reconstruction:** マスク領域の画像を予測

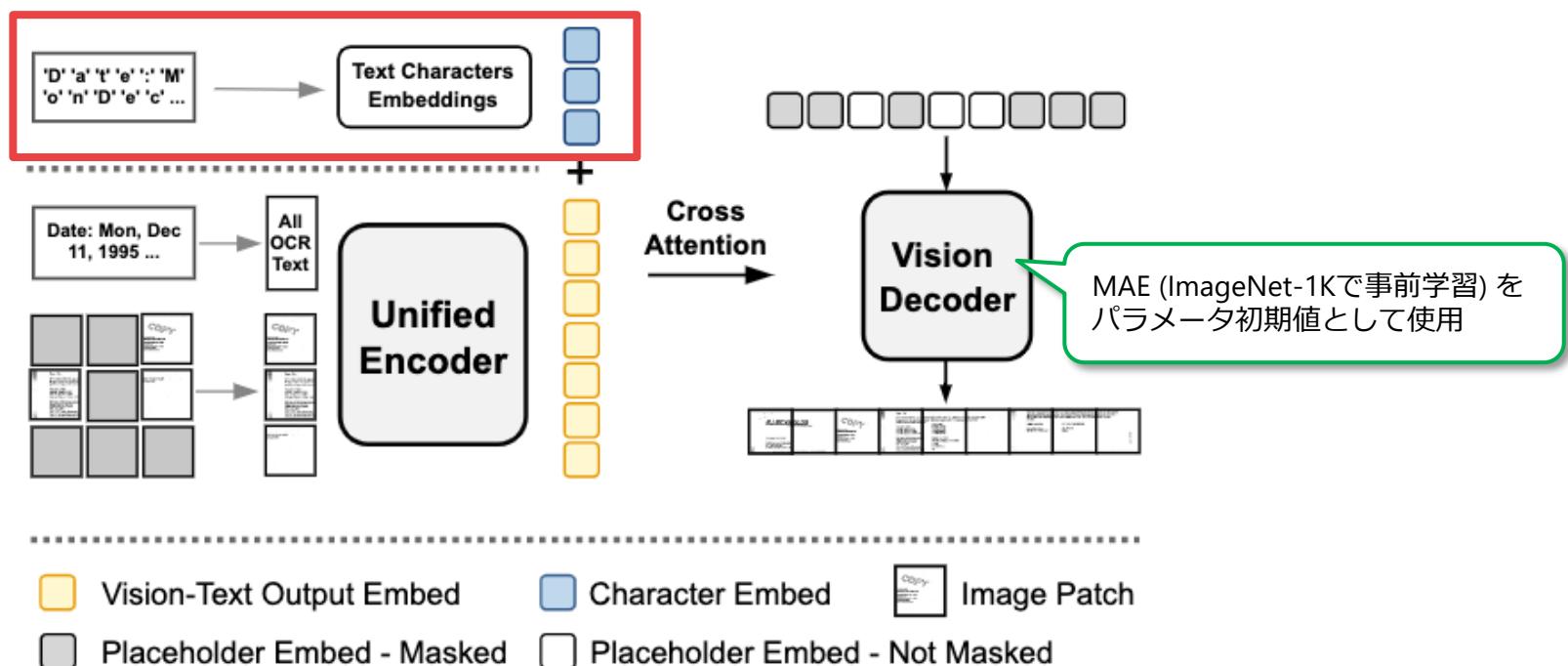
自己教師あり事前学習タスクの例

Self-Supervised Tasks	Task Prompts	Task Targets
Layout Modeling	<i>Layout Modeling.</i> <layout_0> Ship Date to Retail </layout_0> Week of March 14, 1994	<layout_0><100><350><118><372>
Visual Text Recognition	<i>Visual Text Recognition.</i> <text_0> <100><350><118><372> </text_0> to Retail: Week of March 14, 1994	<text_0> Ship Date
Joint Text-Layout Reconstruction	<i>Joint Text-Layout Reconstruction.</i> <text_layout_0> to Retail: Week of March 14, 1994	<text_layout_0> Ship Date <100><350><118><372>
Masked Image Reconstruction	<i>Masked Image Reconstruction.</i> Ship Date to Retail: Week of March 14, 1994	[Pixels of the original image]

文字情報を用いたマスク画像復元

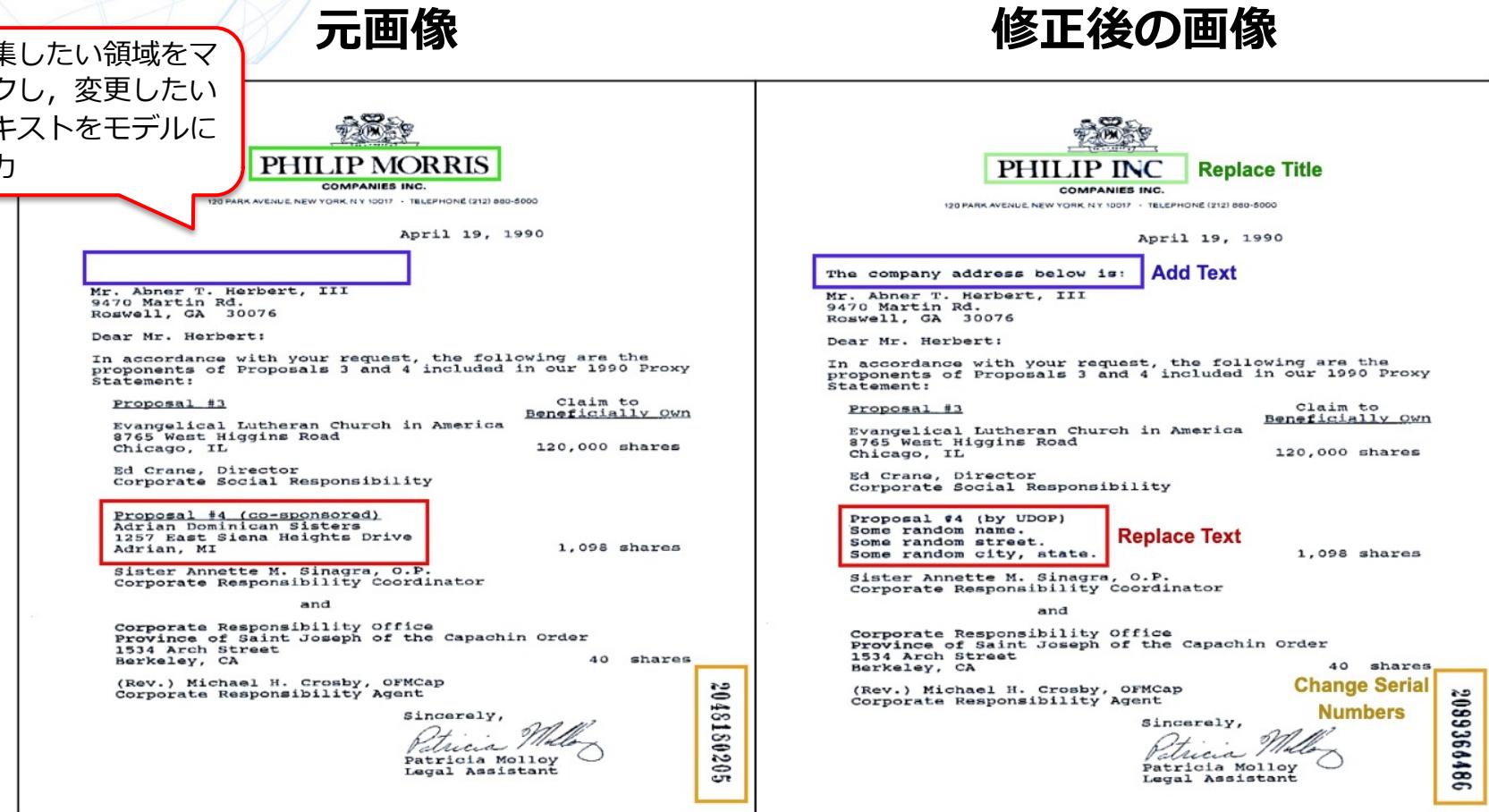
- OCR文字情報を加えることで、文字レベルの詳細なマスク画像復元を実現
- Cross Attentionを用いて、OCR文字情報・OCRテキスト・文書画像情報を参照

OCR文字情報



テキストによる文書画像編集が可能に

- テキストを条件付けたマスク画像復元を事前学習に行なったことで、自然言語指示による文書画像編集が可能に



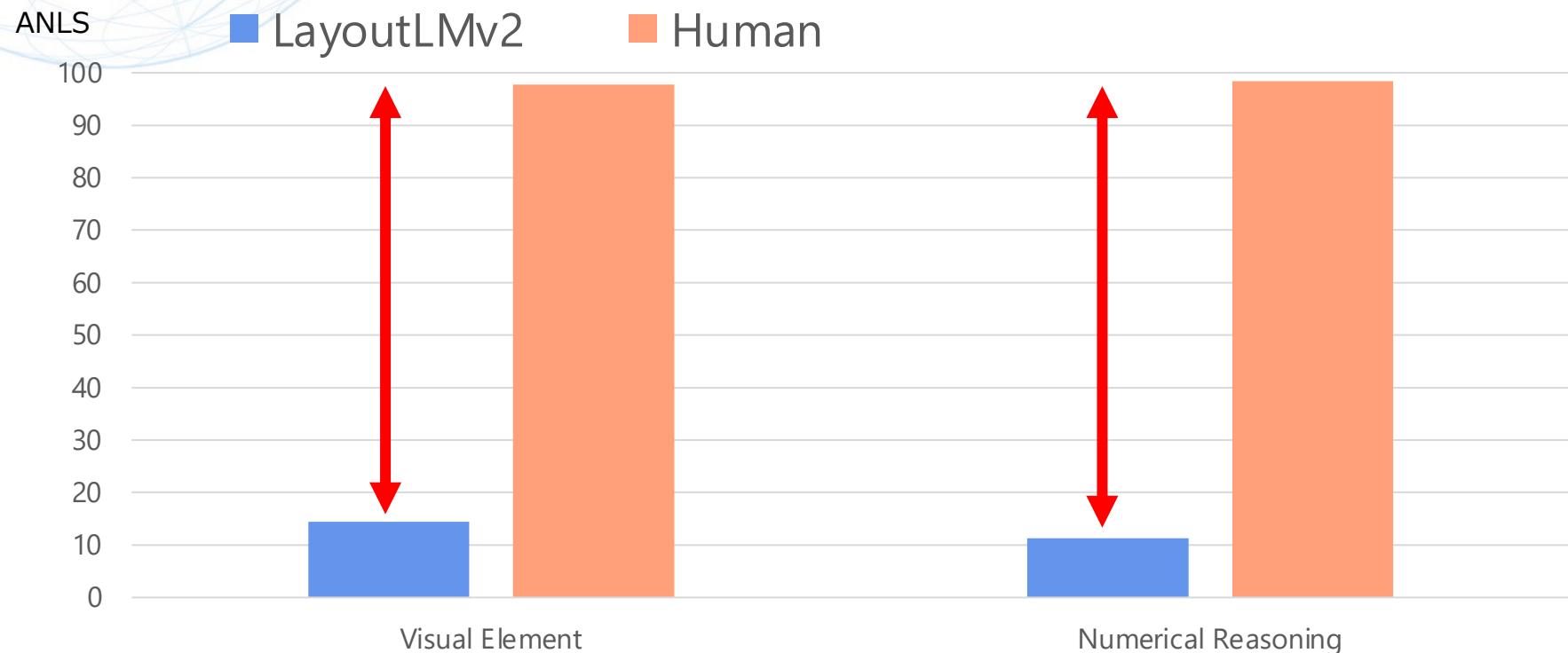
文書画像読解モデル の今後の方向性

今後の方針

- 視覚要素の理解および算術演算の強化
- マルチリンガル化
- End-to-End文書理解
- 大規模言語モデルの活用

視覚物体理解と算術演算

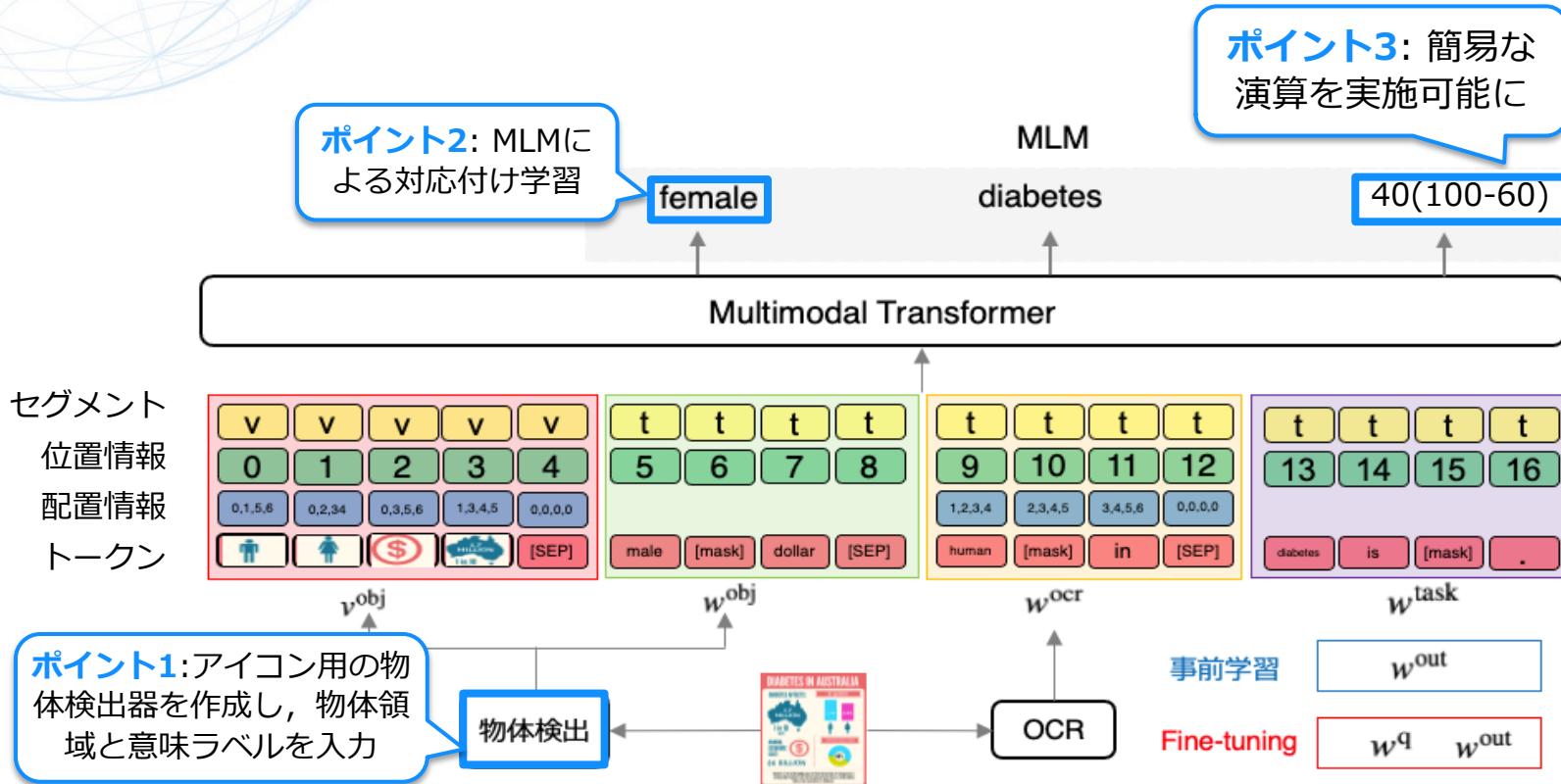
- 文書読解モデルは文書内の視覚物体や算術演算を苦手としている [Tanaka+, 21][Mathew+, 22].



InfographicVQA [Mathew+, WACV'22]における評価結果

IG-BERT [田中+ (NTT), NLP'22 若手奨励賞]

- InfographicVQAコンペティションにおいて、事前学習データ量を従来モデルの1/22に抑えつつ同程度のサイズのモデルの中で最も高い性能を達成し2位に入賞（18チーム337投稿中）



演算過程を考慮したデータ拡張

- 演算の過程を学習可能な新たなデータ拡張手法を提案



1. 文書から数値データの抽出

60% 4.3% 3.6% 1 10 \$6

2. 事前に用意した演算テンプレートに代入

$$60\% + 4.3\% = 64.3\%$$

:

$$100\% - 60\% = 40\%$$

3. 回答正解データと一致するならデータに加える

質問: 糖尿病を防げた事例の割合は?

回答正解データ: 40%

100-60=40の
演算が必要

$$100\% - 60\%$$

この演算過程テキスト
そのものを生成する

マルチリングル化の障壁

- 多言語の文書画像 (請求書, 契約書など) を大規模に収集するのは、大変
 - 言語によって、データ量の偏りが生じてしまいがち
 - 権利上・個人情報保護の観点で、Web上にあまり落ちていない
- 仮説:** 言語が変わっても、文書レイアウトは、普遍的な場合が多い [Wang(SCUT)+, ACL'22]

→ 英語で文書レイアウトを事前学習して、他言語に転移可能か？

英語



This image shows a scanned document form in English. It includes fields for personal information like name, address, and contact details, as well as a signature section at the bottom.

(a) A form.

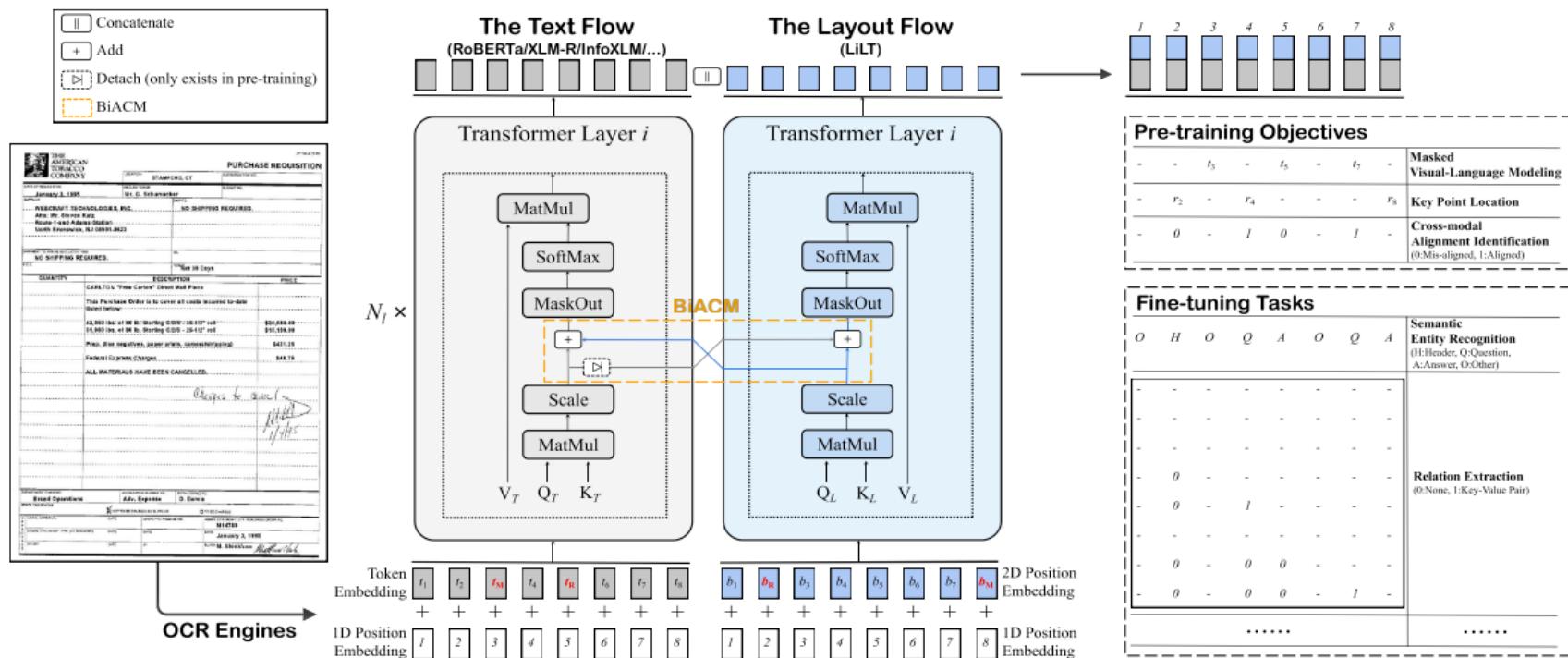
中国語

HAPPY HOUR RESTAURANT					欢乐时光餐厅									
Invoice #: 1803321345698 TEL: 07-3233123					发票编号: 1803321345698 电话: 07-3233123									
账号 : 3					账号 : 3									
Member Points: 0.00					会员积分: 0.00									
Date: 20/01/2019 Time: 12:42:00					日期: 20/01/2019 时间: 12:42:00									
Casher: He Bei 收银员: 何贝														
Item Qty U.P. DISC % Amount					品名 数量 单价 折扣 % 总额									
Beef Noodles 2 10.00 10.00					牛肉面 2 10.00 10.00									
Milk 1 3.00 3.00					牛奶 1 3.00 3.00									
Total Qty: 2 Total Points: 0.00					总件数: 2 总积分: 0.00									
Sub Total: 13.00					小计: 13.00									
Dish: 0.00					折扣: 0.00									
Service Charge: 0.00					服务费: 0.00									
Tax: 0.00					税: 0.00									
Total: 13.00					总计: 13.00									
Cash: 13.00					现金: 13.00									
Change: 0.00					找零: 0.00									
Goods sold are not returnable. Thank you!														
离店—销售出票不退还。 谢！														

(b) A receipt.

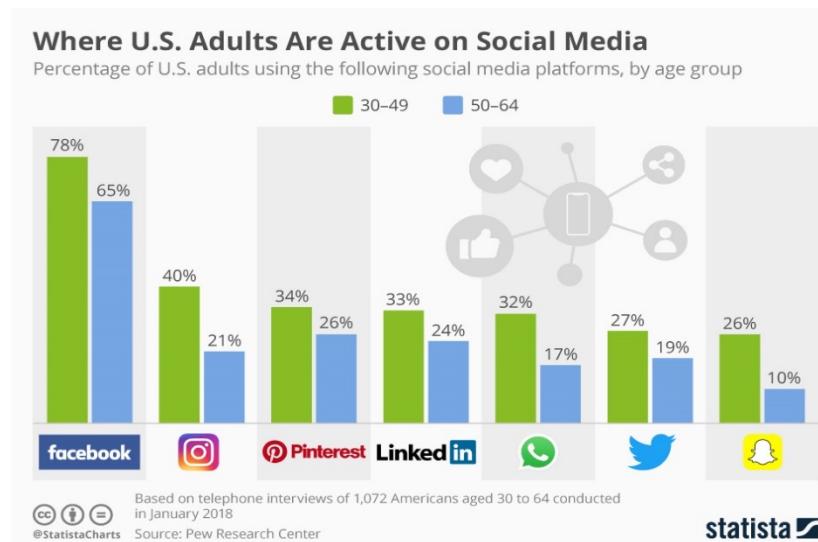
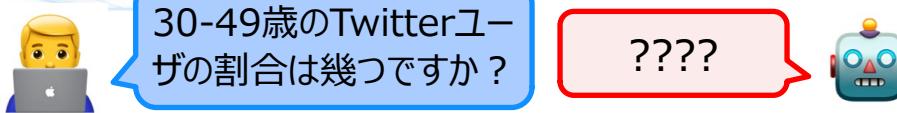
LiLT [Wang(SCUT)+, ACL'22]

- Text FlowとLayout Flowのモジュールに分割
 - 言語に依存しないテキストとレイアウトの関係性をより効率的に学習
- 多言語タスクにおいて、大量の多言語文書で学習したモデル LayoutXLM [Xu,+ arxiv'21] を英語のみの事前学習データで上回る



前処理の無いEnd-to-End文書理解へ

- 前処理 (OCRや物体検出など) により、汎用的な視覚物体理解には限界
- 前処理が全体の処理時間の約半分を占めている、前処理の性能が後段タスクの性能に大きく影響、などの問題がある



商用APIの物体検出結果: オブジェクト検出なし

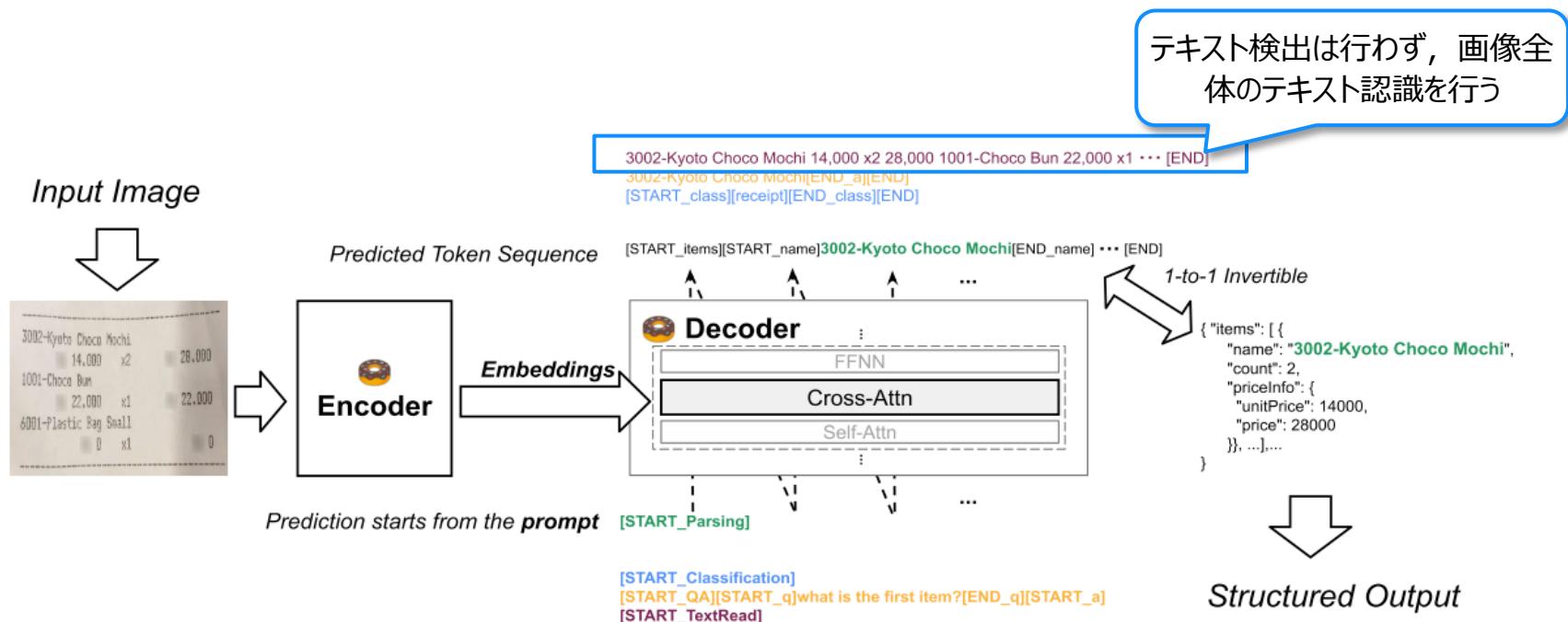
Memory (M) **Time (sec/img)**



文書画像読解のメモリ消費量
と処理時間 [Kim(NAVER+, ECCV'22)]

Donut [Kim(NAVER)+, ECCV'22]

- OCRを事前学習で行うことで、前処理にOCRを行わずに済み、文書画像単体の入力のみでも一定の効果を確認
 - 速度:** OCRありモデルよりも推論速度が2~4倍高速
 - 精度:** OCRありモデルと比べると、QAタスクの精度は劣っているものの、文書画像分類タスクではコンパラの精度。画像全体の雰囲気を掴むことには成功しているものの、文脈理解には改善の余地が見られる



Pix2Struct [Lee(Google)+, arXiv'22]

- Webページのスクリーンショット画像に対してマスクをし, 元のHTMLソースコードを予測
- Altテキストもマスク対象にすることで, 視覚要素(アイコンや写真など)の理解を促進
- 2段階のカリキュラム事前学習を実施. 1段階目の学習ではOCRを行い, 2段階目にマスクHTMLソースコード予測を行う

OCR予測

The elves, it seemed, were possessed
of some mysterious power over the
arts; without eve → The elves, it seemed, were
possessed of some mysterious
power over the arts; without eve

マスクHTMLソースコード予測

Programming Survey

Your favorite programming language:

- Python
- C++
- Java

Submit

Programming Survey

Your favorite programming language:

- Python
- C++
- Java

Submit

画像内のテキストに
対してマスク
を行う

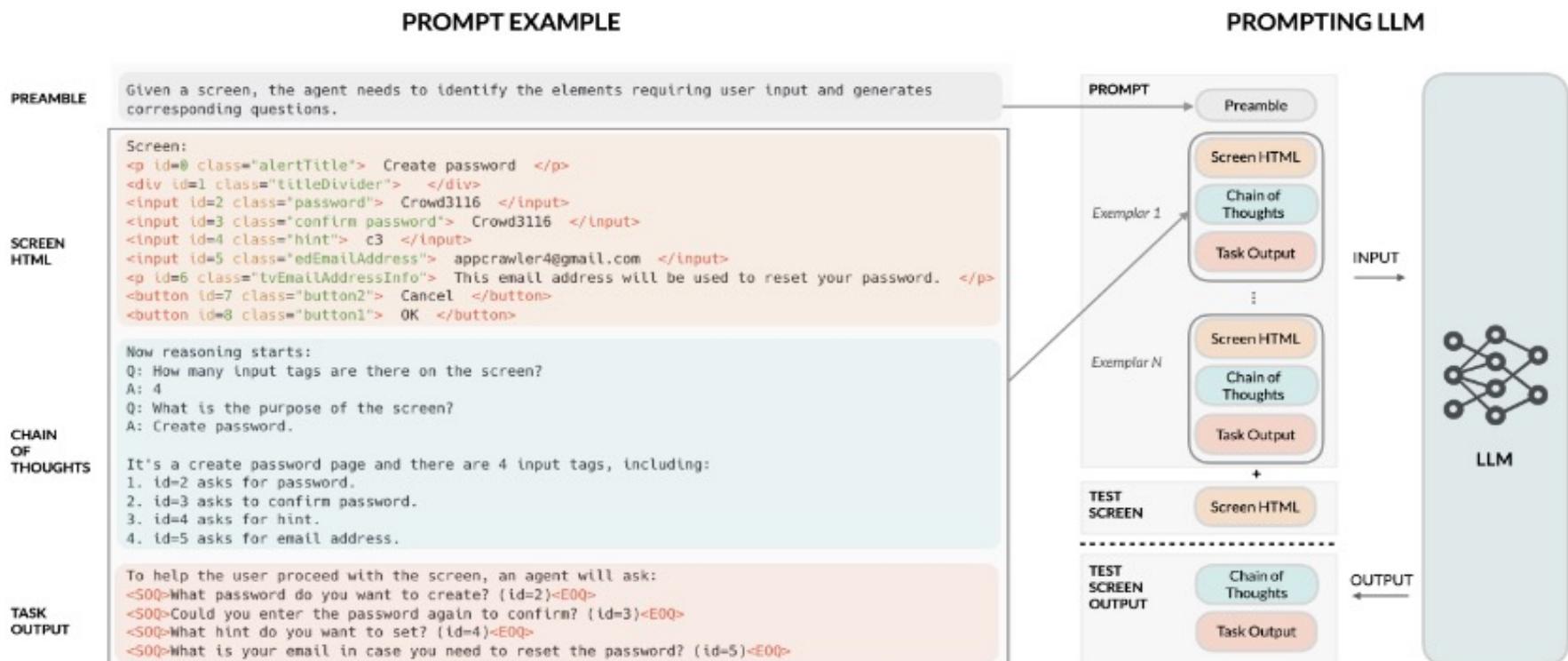
```
<<<Python>


```

大規模言語モデルの活用 [Wang(Utronto)+, CHI'22]

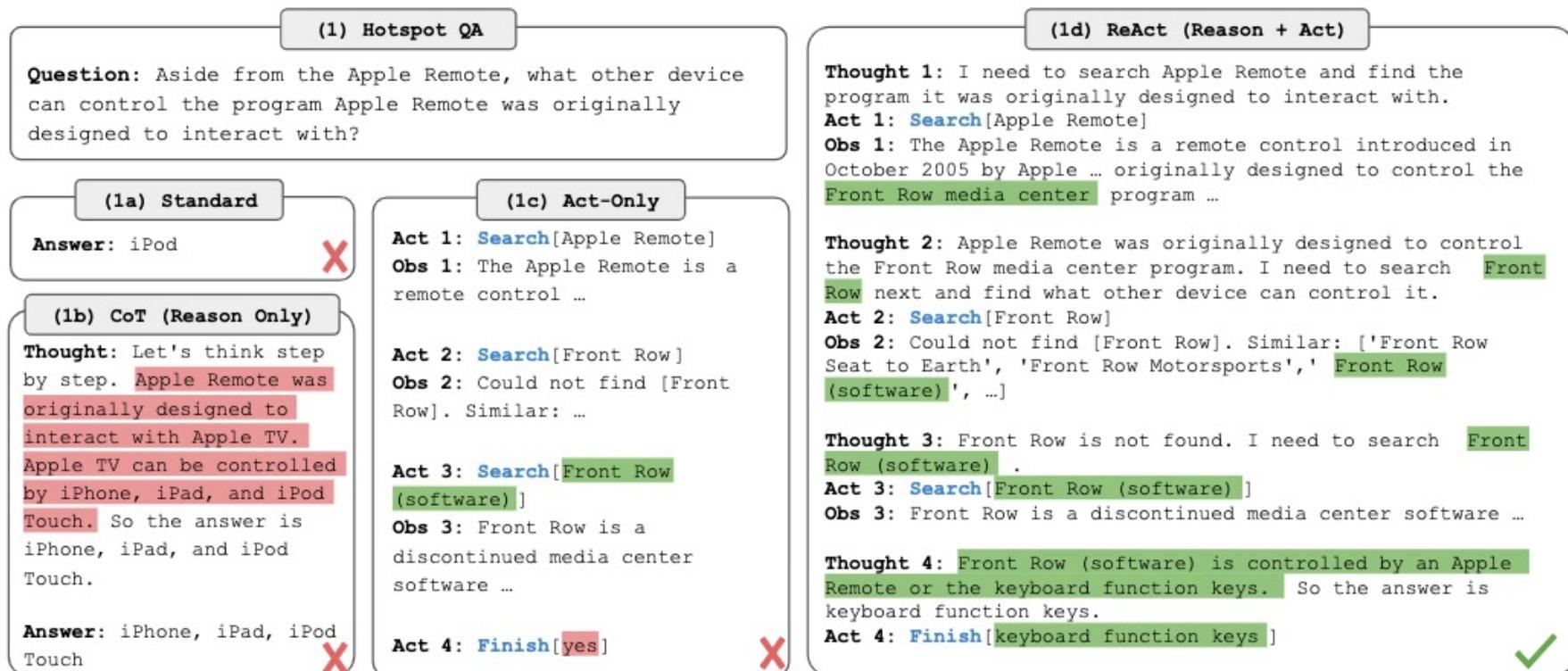
NTT

- 大規模言語モデル (LLM) に含まれる一般常識、事実知識などを活用
- 指示文と例示 (HTML, 推論の過程, 出力, の組) をPromptとして巨大言語モデルに入力する
- LLMの限られた入力長の中で、長いWebページは全て入力できない。LLMに効率よく目的のタスクを認識させることが課題



ReAct [Yao(Princeton Univ.)+, ICLR'23]

- LLMのコンテキスト内で推論と行動決定を融合する手法の提案
- 外部環境（Webページなど）との接続を行い、検索、クリック、スクロールなどを含めた行動決定を行うことも可能
- 画像情報を含む推論や行動決定は、今後の課題



本パートのまとめ

- 汎用的な文書を理解するためには、テキストベースでは限界があり、文書「画像」理解のアプローチが有望である
- 文書画像のレイアウト理解において、人間の性能に近いモデルが登場している
- 今後の方向性
 - 視覚要素の理解および算術演算の強化
 - Multilingual化
 - End-to-End文書理解
 - モデルの大規模化

本講演のまとめ

近年の自然言語処理 および Vision-and-Language 分野について、基礎的な内容・最新動向・今後の展望について紹介しました。

1. モデルは汎用志向、タスクも理解／生成に大別される

2. セルフアテンションを用いる Transformer が主流に

- 系列データに強く、マルチモーダルの統一的なモデリングも可能

3. 自己教師有り学習で大規模事前学習することが主流に

- BERT : フайнチューニングにより各タスクに適応
- GPT-3 : プロンプトによる少量の例示で新タスクに適応

4. ChatGPTによりAIは「新時代」へ

- 自然言語処理におけるBERT以来のパラダイム・シフト
- 人の意図や感覚に適合した言語モデルへの発展

5. 視覚と結びつけた言語理解の大きな発展

- CLIPやFlamingoの成功は大きなマイルストンとなった
- 文書や画面を人のように視覚的に読み取ることも可能に

1. Aditya Ramesh et al.: Hierarchical Text-Conditional Image Generation with CLIP Latents. CoRR abs/2204.06125 (2022)
2. Jean-Baptiste Alayrac et al.: Flamingo: a Visual Language Model for Few-Shot Learning. CoRR abs/2204.14198 (2022)
3. Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS 2015: 91-99
4. Liunian Harold Li et al.: VisualBERT: A Simple and Performant Baseline for Vision and Language. CoRR abs/1908.03557 (2019)
5. Pengchuan Zhang et al: VinVL: Revisiting Visual Representations in Vision-Language Models. CVPR 2021: 5579-5588
6. Alexey Dosovitskiy et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021
7. Alec Radford et al.: Learning Transferable Visual Models From Natural Language Supervision. ICML 2021: 8748-8763
8. Vladimir Karpukhin et al.: Dense Passage Retrieval for Open-Domain Question Answering. EMNLP (1) 2020: 6769-6781
9. Or Patashnik et al.: StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. ICCV 2021: 2065-2074
10. Katherine Crowson et al: VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance. CoRR abs/2204.08583 (2022)
11. Jonathan Ho et al.: Denoising Diffusion Probabilistic Models. NeurIPS 2020
12. Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, Juan Pablo Bello: Wav2CLIP: Learning Robust Audio Representations from Clip. ICASSP 2022: 4563-4567
13. Xiuye Gu et al.: Zero-Shot Detection via Vision and Language Knowledge Distillation. ICLR 2022
14. Yael Vinker et al.: CLIPasso: Semantically-Aware Object Sketching. SIGGRAPH 2022.
15. Guy Tevet et al: MotionCLIP: Exposing Human Motion Generation to CLIP Space. CoRR abs/2203.08063 (2022)

16. Oscar Michel et al.: Text2Mesh: Text-Driven Neural Stylization for Meshes. CVPR 2022: 13482-13492
17. Fangzhou Hong et al.: AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. ACM Trans. Graph. 41(4): 161:1-161:19 (2022)
18. Junnan Li et al.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. CoRR abs/2301.12597 (2023)
19. Shaohan Huang et al.: Language Is Not All You Need: Aligning Perception with Language Models. CoRR abs/2302.14045 (2023)
20. Carlos Soto and Shinjae Yoo: Visual Detection with Context for Document Layout Analysis. EMNLP/IJCNLP 2019
21. Xu Zhong et al.: PubLayoutNet: Largest Dataset Ever for Document Layout Analysis. ICDAR 2019
22. Zilong Wang et al.: LayoutReader: Pre-training of Text and Layout for Reading Order Detection. EMNLP 2021
23. Guillaume Jaume et al.: FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. OST@ICDAR 2019
24. Seunghyun Park et al.: CORD: A Consolidated Receipt Dataset for Post-OCR Parsing, Document Intelligence Workshop @ NeurIPS 2019
25. Adam W. Harley et al.: Evaluation of deep convolutional nets for document image classification and retrieval. ICDAR 2015
26. Minesh Mathew et al.: DocVQA: A Dataset for VQA on Document Images. WACV 2021
27. Ryota Tanaka et al: VisualMRC: Machine Reading Comprehension on Document Images. AAAI 2021
28. Minesh Mathew et al: InfographicVQA. WACV 2022
29. Ryota Tanaka et al.: SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images. AAAI 2023
30. Peter C. Humphreys et al.: A data-driven approach for learning to control computers. ICML 2022: 9466-9482

31. Burns A et al.: A Dataset for Interactive Vision-Language Navigation with Unknown Command Feasibility. ECCV2022
32. Sun L et al.: Towards Multi-modal Conversational Agents on Mobile GUI. EMNLP2022
33. Lee SW et al.: Can Current Task-oriented Dialogue Models Automate Real-world Scenarios in the Wild?. arXiv:2212.10504.
34. Xu Y et al.: Layoutlm: Pre-training of text and layout for document image understanding. KDD2022
35. Xu Y et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. ACL2021
36. Huang Y et al.: Layoutlmv3: Pre-training for document ai with unified text and image masking. ACMM 2022
37. Li C et al.: Structurallm: Structural pre-training for form understanding. ACL21
38. Tang Z et al.: Unifying Vision, Text, and Layout for Universal Document Processing. arXiv:2212.02623
39. Peng Q et al. ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding. Findings of EMNLP2022
40. 田中涼太 et al. : テキストと視覚的に表現された情報の融合理解に基づくインフォグラフィク質問応答. NLP2022
41. Kim G et al.: Ocr-free document understanding transformer. ECCV2022
42. Lee K et al.: Toutanova K.: Pix2Struct: Screenshot parsing as pretraining for visual language understanding. arXiv:2210.03347.
43. Wang J et al.: Lilt: A simple yet effective language-independent layout transformer for structured document understanding. ACL2022
44. Wang B et al.: Enabling Conversational Interaction with Mobile UI using Large Language Models. CHI2023
45. Yao S et al.: React: Synergizing reasoning and acting in language models. ICLR2023