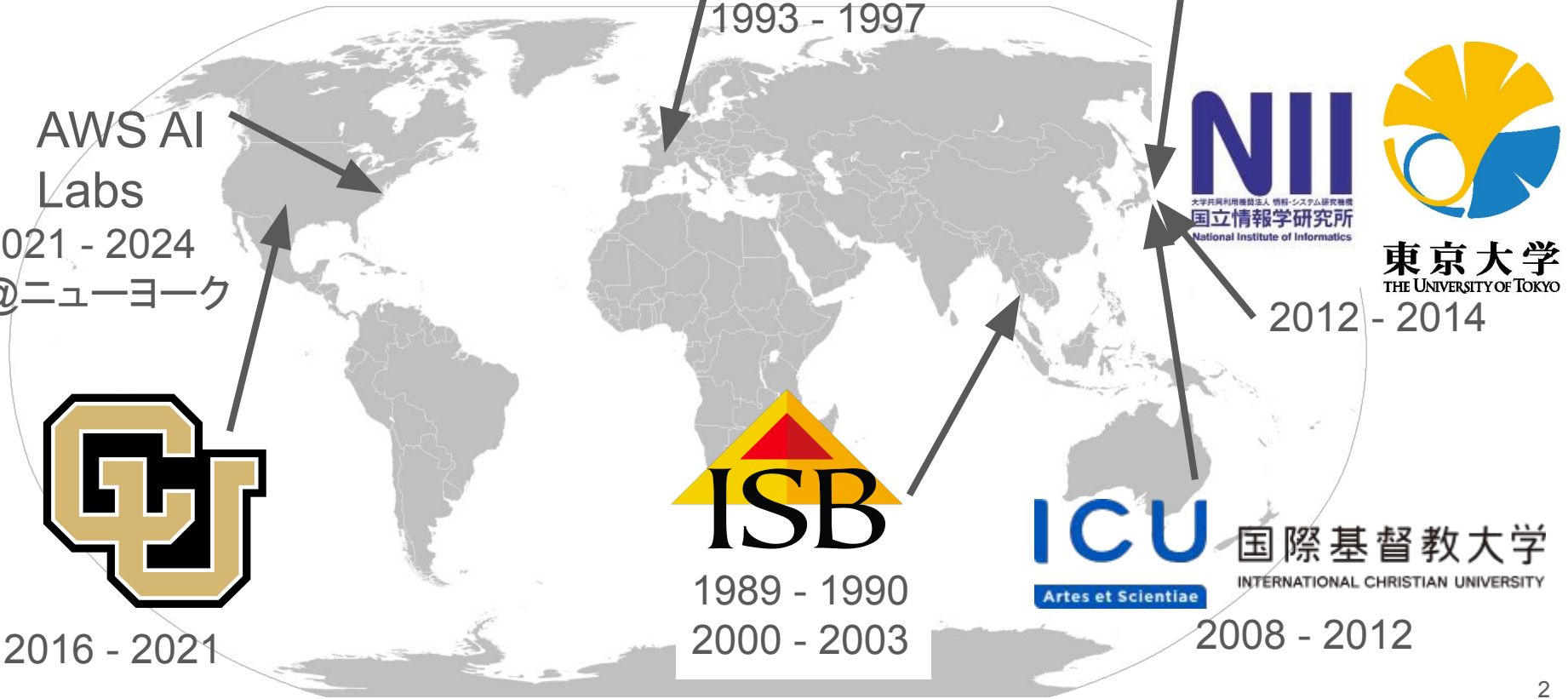
# Visually-Rich Documentを軸とした多言語処理の動向

2023年12月12日@名古屋
2023年12月15日@東京
2023年12月21日@奈良
2024年01月12日@仙台
2024年7月23日＠東京

藤沼祥成

# 自己紹介:略歴



International School of Paris
Educating for complexity
1993 - 1997

Amazon JP
2014 - 2016

AWS AI Labs
2021 - 2024
@ニューヨーク

NII 国立情報学研究所 National Institute of Informatics
東京大学 THE UNIVERSITY OF TOKYO
2012 - 2014

2016 - 2021

ISB
1989 - 1990
2000 - 2003

ICU Artes et Scientiae
国際基督教大学 INTERNATIONAL CHRISTIAN UNIVERSITY
2008 - 2012

# 自己紹介：最近やっていること

- タイの出生証明書→→→→→→→→→→
- 2023年上半期はPDF/文書処理に関する仕事を主に従事
  - 多言語データセット構築の話がEMNLP Findings 2023に

去年の仕事だが今回話さないこと

- Dialogue State Trackingの効率化 [Lesci+ 2023]
- 多言語LMのバイアス解析 [Levy+ 2023]


タイの出生証明書例。在インドタイ大使館のサイトより引用

# 目次

- イントロ：多言語NLP
- Visually-Rich Documentの多言語NLP
- Visually-Rich Documentから見たVision-Languageモデル

# イントロ：多言語 NLP

# 世界で話されている言語は7000言語以上

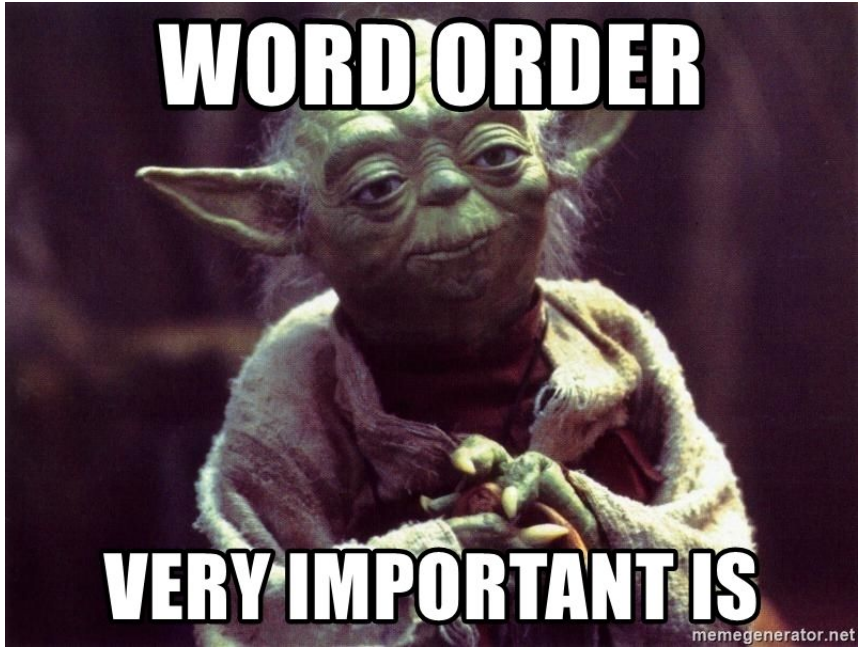

Ref: https://www.eventplanner.net/news/8886_how-to-run-an-efficient-multilingual-conference.html

# 多言語NLPにおける課題

- 例えばあるモデルは英語で学習して評価済み
- 他の言語では？
  - スペイン語は動くかも
  - アムハラ語では？

課題:

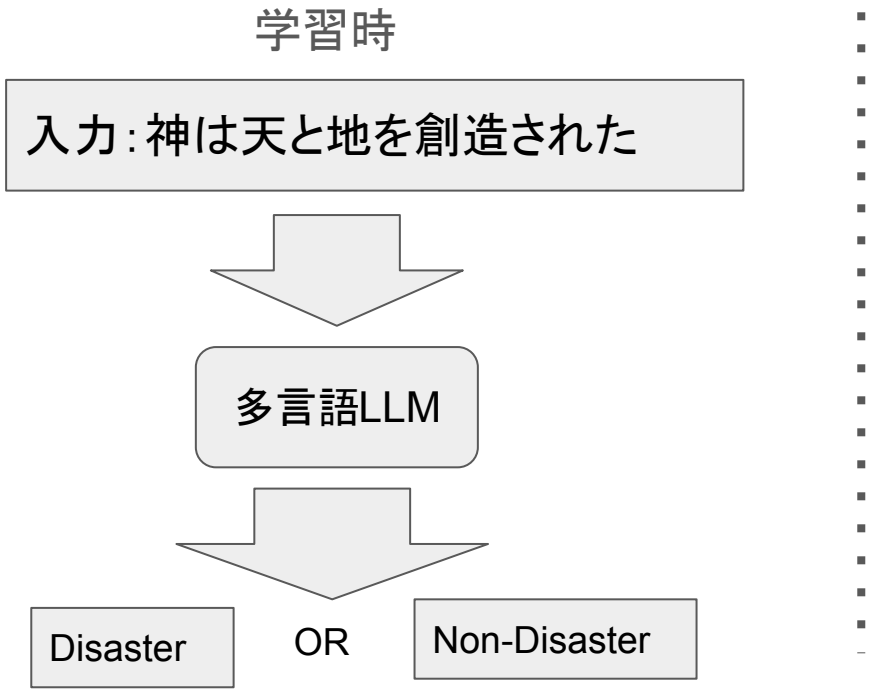
- 言語学観点から
  - 言語の多様性
- 機械学習的観点から
  - 対象言語のデータがない




Ref: https://memegenerator.net/instance/57127615/yoda-word-order-very-important-is

# 言語間転移学習 (Cross-Lingual Transfer Learning)

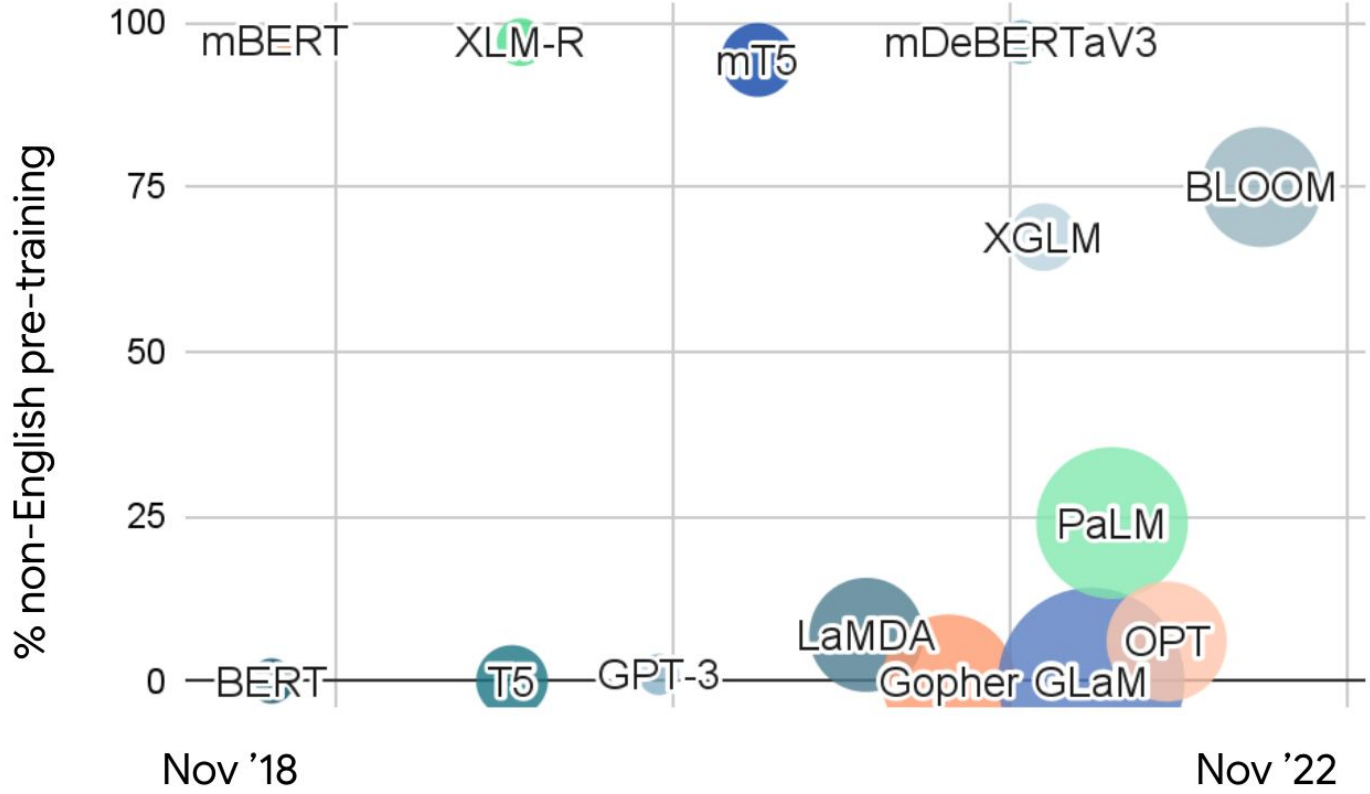
- 言語1で学習し、言語2で推論する



# 言語間転移学習 (Cross-Lingual Transfer Learning)

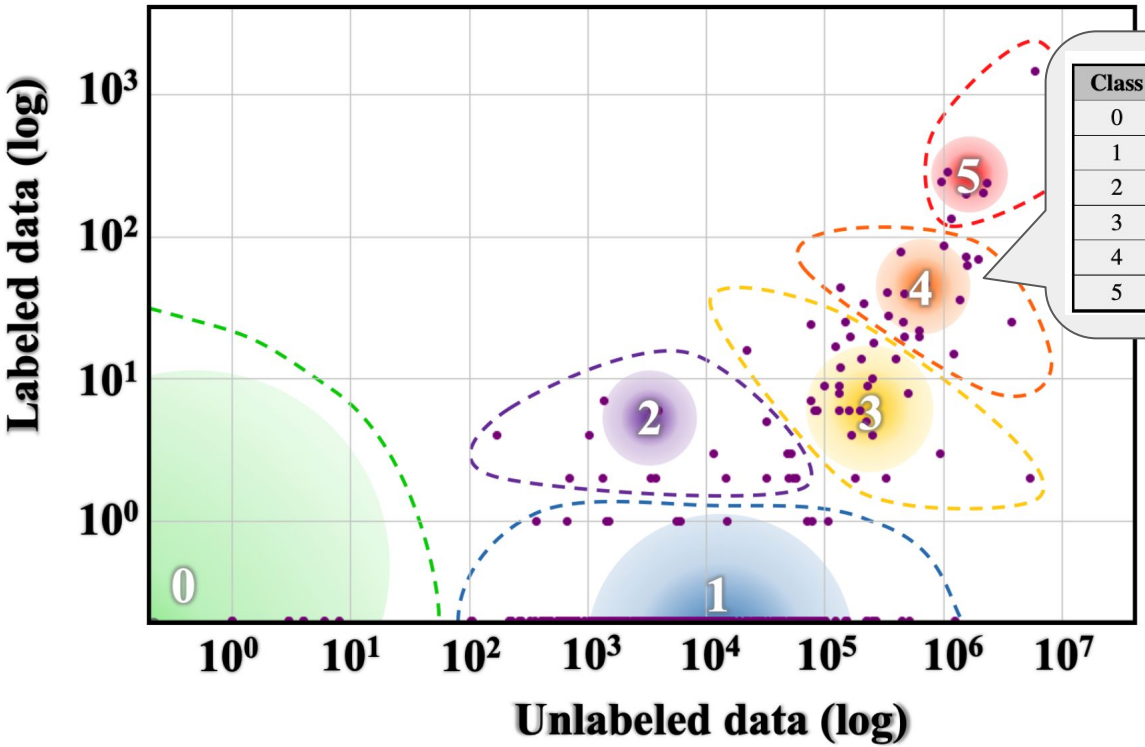
- 言語1で学習し、言語2で推論する

# 多言語事前学習済みモデルは多くある




Ref: https://www.ruder.io/state-of-multilingual-ai/

# 各言語のデータのボトルネックをどうするかが今後の課題




| Class | 5 Example Languages | #Langs |
|---|---|---|
| 0 | Dahalo, Warlpiri, Popoloca, Wallisian, Bora | 2191 |
| 1 | Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo | 222 |
| 2 | Zulu, Konkani, Lao, Maltese, Irish | 19 |
| 3 | Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew | 28 |
| 4 | Russian, Hungarian, Vietnamese, Dutch, Korean | 18 |
| 5 | English, Spanish, German, Japanese, French | 7 |

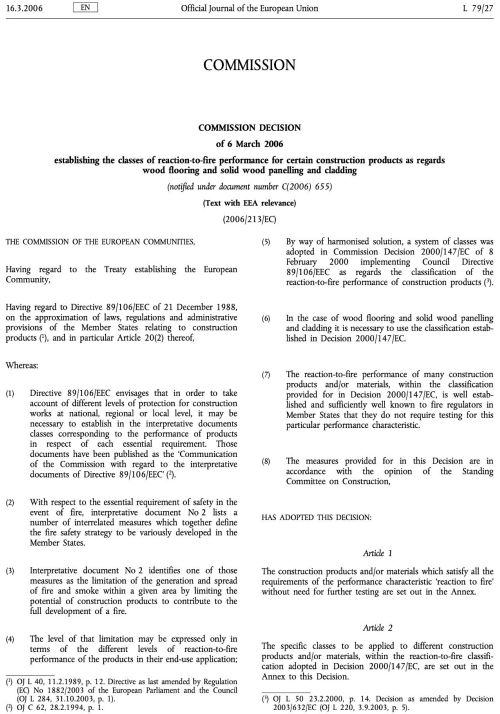
[Joshi+ 2021] The State and Fate of Linguistic Diversity and Inclusion in the NLP World

# Visually-Rich Documentの多言語研究

# Visually-Rich Documentとは何か?

## ウィキペディア


## Eurlex: EUの法令関連文書


16.3.2006 EN Official Journal of the European Union L 79/27

COMMISSION

COMMISSION DECISION

of 6 March 2006

establishing the classes of reaction-to-fire performance for certain construction products as regards wood flooring and solid wood panelling and cladding

(notified under document number C(2006) 655)

(Text with EEA relevance)

(2006/213/EC)

THE COMMISSION OF THE EUROPEAN COMMUNITIES,

Having regard to the Treaty establishing the European Community,

Having regard to Directive 89/106/EEC of 21 December 1988, on the approximation of laws, regulations and administrative provisions of the Member States relating to construction products (1), and in particular Article 20(2) thereof,

Whereas:

(1) Directive 89/106/EEC envisages that in order to take account of different levels of protection for construction works at national, regional or local level, it may be necessary to establish in the interpretative documents classes corresponding to the performance of products in respect of each essential requirement. Those documents have been published as the 'Communication of the Commission with regard to the interpretative documents of Directive 89/106/EEC' (2).

(2) With respect to the essential requirement of safety in the event of fire, interpretative document No 2 lists a number of interrelated measures which together define the fire safety strategy to be variously developed in the Member States.

(3) Interpretative document No 2 identifies one of those measures as the limitation of the generation and spread of fire and smoke within a given area by limiting the potential of construction products to contribute to the full development of a fire.

(4) The level of that limitation may be expressed only in terms of the different levels of reaction-to-fire performance of the products in their end-use application;

(5) By way of harmonised solution, a system of classes was adopted in Commission Decision 2000/147/EC of 8 February 2000 implementing Council Directive 89/106/EEC as regards the classification of the reaction-to-fire performance of construction products (3).

(6) In the case of wood flooring and solid wood panelling and cladding it is necessary to use the classification established in Decision 2000/147/EC.

(7) The reaction-to-fire performance of many construction products and/or materials, within the classification provided for in Decision 2000/147/EC, is well established and sufficiently well known to fire regulators in Member States that they do not require testing for this particular performance characteristic.

(8) The measures provided for in this Decision are in accordance with the opinion of the Standing Committee on Construction,

HAS ADOPTED THIS DECISION:

Article 1

The construction products and/or materials which satisfy all the requirements of the performance characteristic 'reaction to fire' without need for further testing are set out in the Annex.

Article 2

The specific classes to be applied to different construction products and/or materials, within the reaction-to-fire classification adopted in Decision 2000/147/EC, are set out in the Annex to this Decision.

(1) OJ L 40, 11.2.1989, p. 12. Directive as last amended by Regulation (EC) No 1882/2003 of the European Parliament and the Council (OJ L 284, 31.10.2003, p. 1).
(2) OJ C 62, 28.2.1994, p. 1.
(3) OJ L 50 23.2.2000, p. 14. Decision as amended by Decision 2003/632/EC (OJ L 220, 3.9.2003, p. 5).

# Visually-Rich Documentを処理する上での課題

ウィキペディア


OCR

ウィキペディア　フリー

[…]

グループステージ[編集]時間は全て(UTC-4)グループH[編集]グループH[編集]Team Pld W D L GF GA GD Pts キューバ 3 2 i 0 3 0 +3 7

[…]

# Visually-Rich Documentを処理する上での課題

ウィキペディア


OCR

ウィキペディア　フリー

[…]

グループステージ[編集]時間は全て(UTC-4)グループH[編集]グループH[編集]Team Pld W D L GF GA GD Pts キューバ 3 2 i 0 3 0 +3 7

[…]

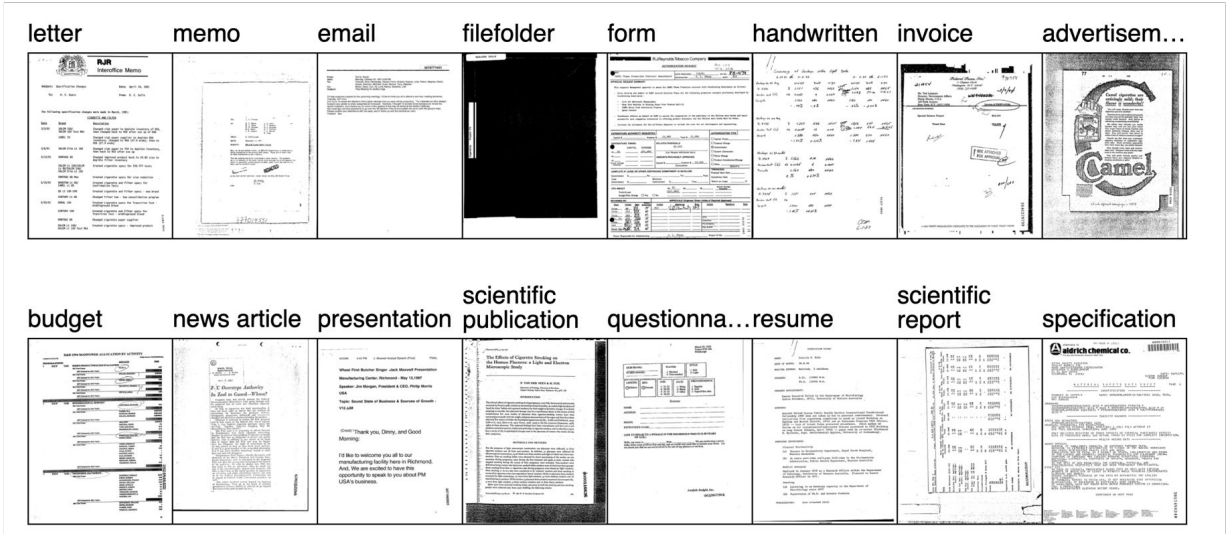
テーブルがわかり辛い

画像はOCRで出力なし

OCRエラー

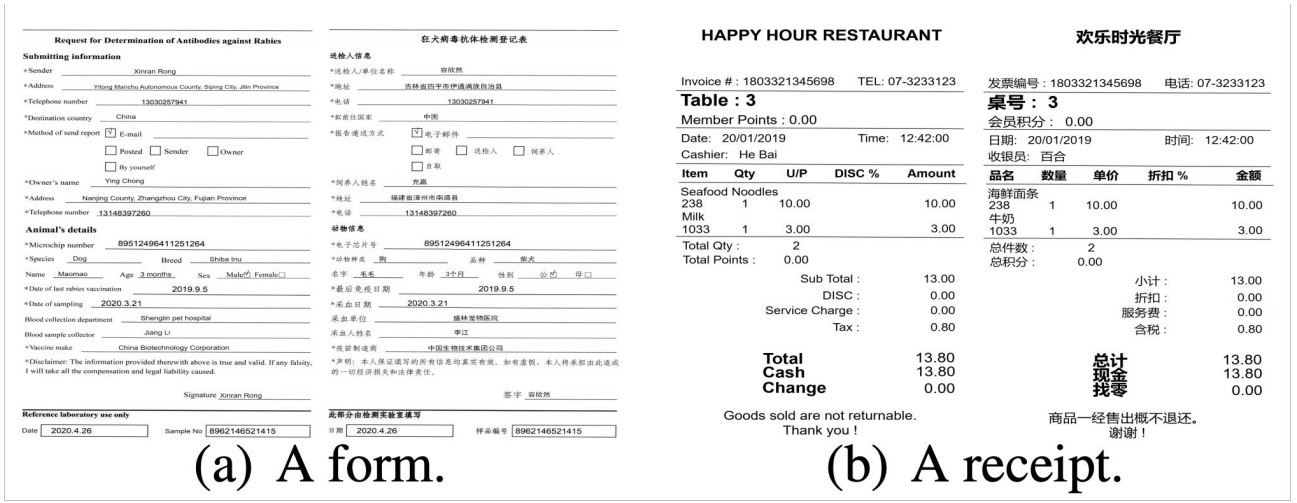
# Visually-Rich Documentの分類データセット[Harley+ 2015]

- 文書画像の16種類の文書タイプ（e.g., メール vs. メモ）に分類
- 問題点
  - 文書の中身を理解しなくても分類可能
  - **英語のみ**



# 英語以外のVisually-Rich Documentは？

- 8言語の文書画像理解データセットは存在する [Xu+ 2021, Wang+ 2022]
- 他の言語、他のタスクにおいてモデルをどう評価するか？



Request for Determination of Antibodies against Rabies

Submitting information

*Sender Xinran Rong
*Address Yitong Manchu Autonomous County, Siping City, Jilin Province
*Telephone number 13030257941
*Destination country China
*Method of send report [V] E-mail
[ ] Posted [ ] Sender [ ] Owner
[ ] By yourself
*Owner's name Ying Chong
*Address Nanjing County, Zhangzhou City, Fujian Province
*Telephone number 13148397260

Animal's details

*Microchip number 89512496411251264
*Species Dog Breed Shiba Inu
Name Maomao Age 3 months Sex Male[V] Female[ ]
*Date of last rabies vaccination 2019.9.5
*Date of sampling 2020.3.21
Blood collection department Shenglin pet hospital
Blood sample collector Jiang Li
*Vaccine make China Biotechnology Corporation
*Disclaimer: The information provided therewith above is true and valid. If any falsity, I will take all the compensation and legal liability caused.

Signature Xinran Rong

Reference laboratory use only

Date 2020.4.26 Sample No 8962146521415

在犬病毒抗体检测登记表

送检人信息

*送检人/单位名称 容欣然
*地址 吉林省四平市伊通满族自治县
*电话 13030257941
*拟前往国家 中国
*报告通讯方式 [V] 电子邮件
[ ] 邮寄 [ ] 送检人 [ ] 饲养人
[ ] 自取
*饲养人姓名 充嘉
*地址 福建省漳州市南靖县
*电话 13148397260

动物信息

*电子芯片号 89512496411251264
*动物种类 狗 品种 柴犬
名字 毛毛 年龄 3个月 性别 公[V] 母[ ]
*最后免疫日期 2019.9.5
*采血日期 2020.3.21
采血单位 盛林宠物医院
采血人姓名 李江
*疫苗制造商 中国生物技术集团公司
*声明：本人保证填写的所有信息均真实有效，如有虚假，本人将承担由此造成的一切经济损失和法律责任。

签字 容欣然

此部分由检测实验室填写

日期 2020.4.26 样品编号 8962146521415

(a) A form.

HAPPY HOUR RESTAURANT

Invoice # : 1803321345698 TEL: 07-3233123
Table : 3
Member Points : 0.00
Date: 20/01/2019 Time: 12:42:00
Cashier: He Bai

| Item | Qty | U/P | DISC % | Amount |
|---|---|---|---|---|
| Seafood Noodles 238 | 1 | 10.00 | | 10.00 |
| Milk 1033 | 1 | 3.00 | | 3.00 |
| Total Qty : | 2 | | | |
| Total Points : | 0.00 | | | |

Sub Total : 13.00
DISC : 0.00
Service Charge : 0.00
Tax : 0.80

Total 13.80
Cash 13.80
Change 0.00

Goods sold are not returnable.
Thank you !

欢乐时光餐厅

发票编号 : 1803321345698 电话: 07-3233123
桌号 : 3
会员积分 : 0.00
日期: 20/01/2019 时间: 12:42:00
收银员: 百合

| 品名 | 数量 | 单价 | 折扣 % | 金额 |
|---|---|---|---|---|
| 海鲜面条 238 | 1 | 10.00 | | 10.00 |
| 牛奶 1033 | 1 | 3.00 | | 3.00 |
| 总件数： | 2 | | | |
| 总积分： | 0.00 | | | |

小计： 13.00
折扣： 0.00
服务费： 0.00
含税： 0.80

总计 13.80
现金 13.80
找零 0.00

商品一经售出概不退还。
谢谢！

(b) A receipt.

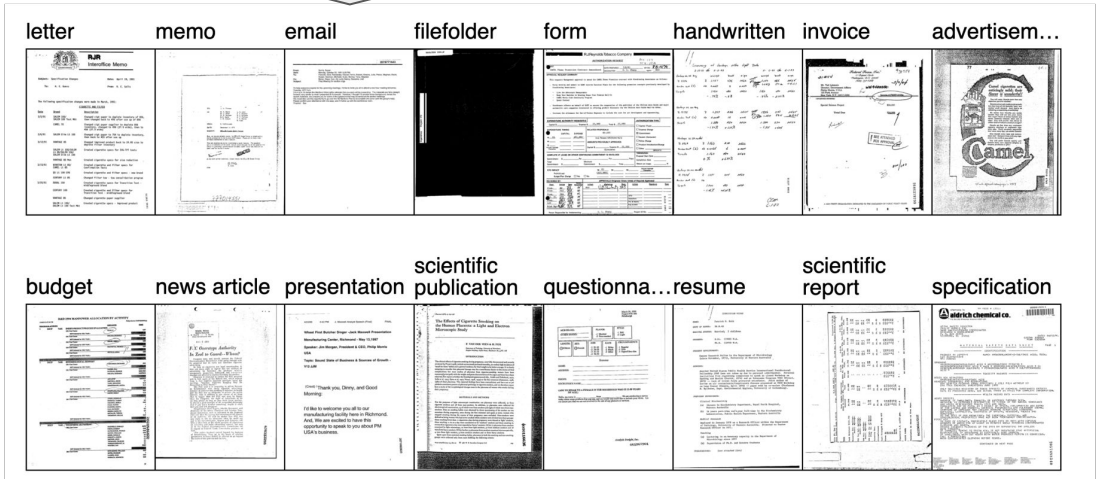
Image from [Wang+ 2022]

# Visually-Rich Docの多言語分類データセット [Fujinuma+ 2023]

新しいデータの良い点:

- トピック分類タスクなので文書内容が重要

リマインド: **RVL-CDIP** [Harley+ 2015]: 16 **文書タイプラベル が付与された** 文書画像分類データセット



# Visually-Rich Docの多言語分類データセット [Fujinuma+ 2023]

新しいデータの良い点:

- トピック分類タスクなので文書内容が重要

# Visually-Rich Docの多言語分類データセット [Fujinuma+ 2023]

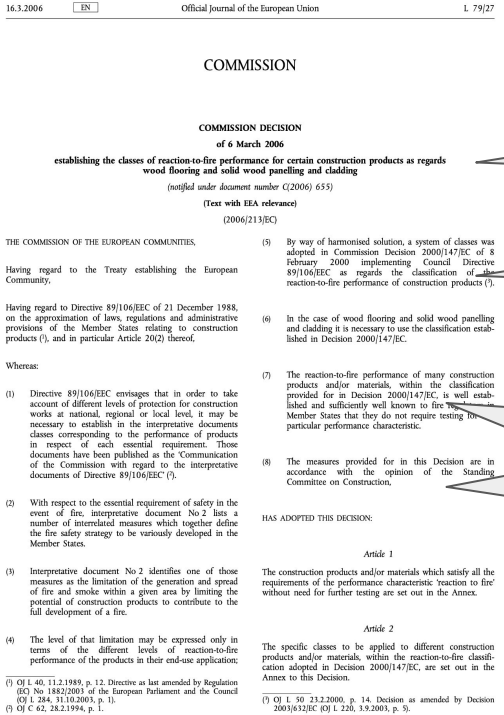
新しいデータの良い点:

- トピック分類タスクなので文書内容が重要

# Visually-Rich Docの多言語分類データセット [Fujinuma+ 2023]

新しいデータの良い点:

- トピック分類タスクなので文書内容が重要
- マルチラベル（1文書に1つ以上のラベル）



16.3.2006 EN Official Journal of the European Union L 79/27

COMMISSION

**COMMISSION DECISION**

**of 6 March 2006**

**establishing the classes of reaction-to-fire performance for certain construction products as regards wood flooring and solid wood panelling and cladding**

*(notified under document number C(2006) 655)*

**(Text with EEA relevance)**

(2006/213/EC)

THE COMMISSION OF THE EUROPEAN COMMUNITIES,

Having regard to the Treaty establishing the European Community,

Having regard to Directive 89/106/EEC of 21 December 1988, on the approximation of laws, regulations and administrative provisions of the Member States relating to construction products [1], and in particular Article 20(2) thereof,

Whereas:

(1) Directive 89/106/EEC envisages that in order to take account of different levels of protection for construction works at national, regional or local level, it may be necessary to establish in the interpretative documents classes corresponding to the performance of products in respect of each essential requirement. Those documents have been published as the 'Communication of the Commission with regard to the interpretative documents of Directive 89/106/EEC' [2].

(2) With respect to the essential requirement of safety in the event of fire, interpretative document No 2 lists a number of interrelated measures which together define the fire safety strategy to be variously developed in the Member States.

(3) Interpretative document No 2 identifies one of those measures as the limitation of the generation and spread of fire and smoke within a given area by limiting the potential of construction products to contribute to the full development of a fire.

(4) The level of that limitation may be expressed only in terms of the different levels of reaction-to-fire performance of the products in their end-use application;

(5) By way of harmonised solution, a system of classes was adopted in Commission Decision 2000/147/EC of 8 February 2000 implementing Council Directive 89/106/EEC as regards the classification of the reaction-to-fire performance of construction products [3].

(6) In the case of wood flooring and solid wood panelling and cladding it is necessary to use the classification established in Decision 2000/147/EC.

(7) The reaction-to-fire performance of many construction products and/or materials, within the classification provided for in Decision 2000/147/EC, is well established and sufficiently well known to fire [illegible] in Member States that they do not require testing for a particular performance characteristic.

(8) The measures provided for in this Decision are in accordance with the opinion of the Standing Committee on Construction,

HAS ADOPTED THIS DECISION:

*Article 1*

The construction products and/or materials which satisfy all the requirements of the performance characteristic 'reaction to fire' without need for further testing are set out in the Annex.

*Article 2*

The specific classes to be applied to different construction products and/or materials, within the reaction-to-fire classification adopted in Decision 2000/147/EC, are set out in the Annex to this Decision.

[1] OJ L 40, 11.2.1989, p. 12. Directive as last amended by Regulation (EC) No 1882/2003 of the European Parliament and the Council (OJ L 284, 31.10.2003, p. 1).
[2] OJ C 62, 28.2.1994, p. 1.
[3] OJ L 50 23.2.2000, p. 14. Decision as amended by Decision 2003/632/EC (OJ L 220, 3.9.2003, p. 9).

Label 1: Industry

Label 2: Environment

Label 3: Production

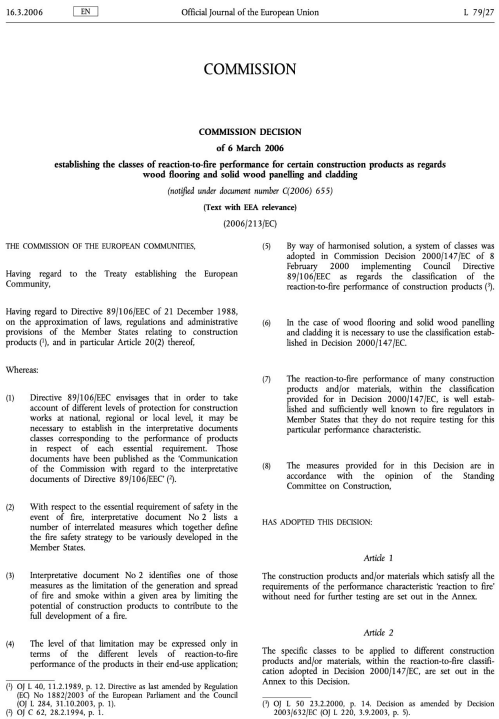
Label 4: Trade

Label 5: Social questions

# Visually-Rich Docの多言語分類データセット [Fujinuma+ 2023]

新しいデータの良い点:

- トピック分類タスクなので文書内容が重要
- マルチラベル（1文書に1つ以上のラベル）
- 多言語
  - EU言語＋アラビア語＋中国語＋日本語



16.3.2006 EN Official Journal of the European Union L 79/27

COMMISSION

**COMMISSION DECISION**
**of 6 March 2006**
**establishing the classes of reaction-to-fire performance for certain construction products as regards wood flooring and solid wood panelling and cladding**
*(notified under document number C(2006) 655)*
**(Text with EEA relevance)**
(2006/213/EC)

THE COMMISSION OF THE EUROPEAN COMMUNITIES,

Having regard to the Treaty establishing the European Community,

Having regard to Directive 89/106/EEC of 21 December 1988, on the approximation of laws, regulations and administrative provisions of the Member States relating to construction products (1), and in particular Article 20(2) thereof,

Whereas:

(1) Directive 89/106/EEC envisages that in order to take account of different levels of protection for construction works at national, regional or local level, it may be necessary to establish in the interpretative documents classes corresponding to the performance of products in respect of each essential requirement. Those documents have been published as the 'Communication of the Commission with regard to the interpretative documents of Directive 89/106/EEC' (2).

(2) With respect to the essential requirement of safety in the event of fire, interpretative document No 2 lists a number of interrelated measures which together define the fire safety strategy to be variously developed in the Member States.

(3) Interpretative document No 2 identifies one of those measures as the limitation of the generation and spread of fire and smoke within a given area by limiting the potential of construction products to contribute to the full development of a fire.

(4) The level of that limitation may be expressed only in terms of the different levels of reaction-to-fire performance of the products in their end-use application;

(5) By way of harmonised solution, a system of classes was adopted in Commission Decision 2000/147/EC of 8 February 2000 implementing Council Directive 89/106/EEC as regards the classification of the reaction-to-fire performance of construction products (3).

(6) In the case of wood flooring and solid wood panelling and cladding it is necessary to use the classification established in Decision 2000/147/EC.

(7) The reaction-to-fire performance of many construction products and/or materials, within the classification provided for in Decision 2000/147/EC, is well established and sufficiently well known to fire regulators in Member States that they do not require testing for this particular performance characteristic.

(8) The measures provided for in this Decision are in accordance with the opinion of the Standing Committee on Construction,

HAS ADOPTED THIS DECISION:

*Article 1*

The construction products and/or materials which satisfy all the requirements of the performance characteristic 'reaction to fire' without need for further testing are set out in the Annex.

*Article 2*

The specific classes to be applied to different construction products and/or materials, within the reaction-to-fire classification adopted in Decision 2000/147/EC, are set out in the Annex to this Decision.

(1) OJ L 40, 11.2.1989, p. 12. Directive as last amended by Regulation (EC) No 1882/2003 of the European Parliament and the Council (OJ L 284, 31.10.2003, p. 1).
(2) OJ C 62, 28.2.1994, p. 1.
(3) OJ L 50 23.2.2000, p. 14. Decision as amended by Decision 2003/632/EC (OJ L 220, 3.9.2003, p. 9).

# Visually-Rich Docの多言語分類データセット [Fujinuma+ 2023]

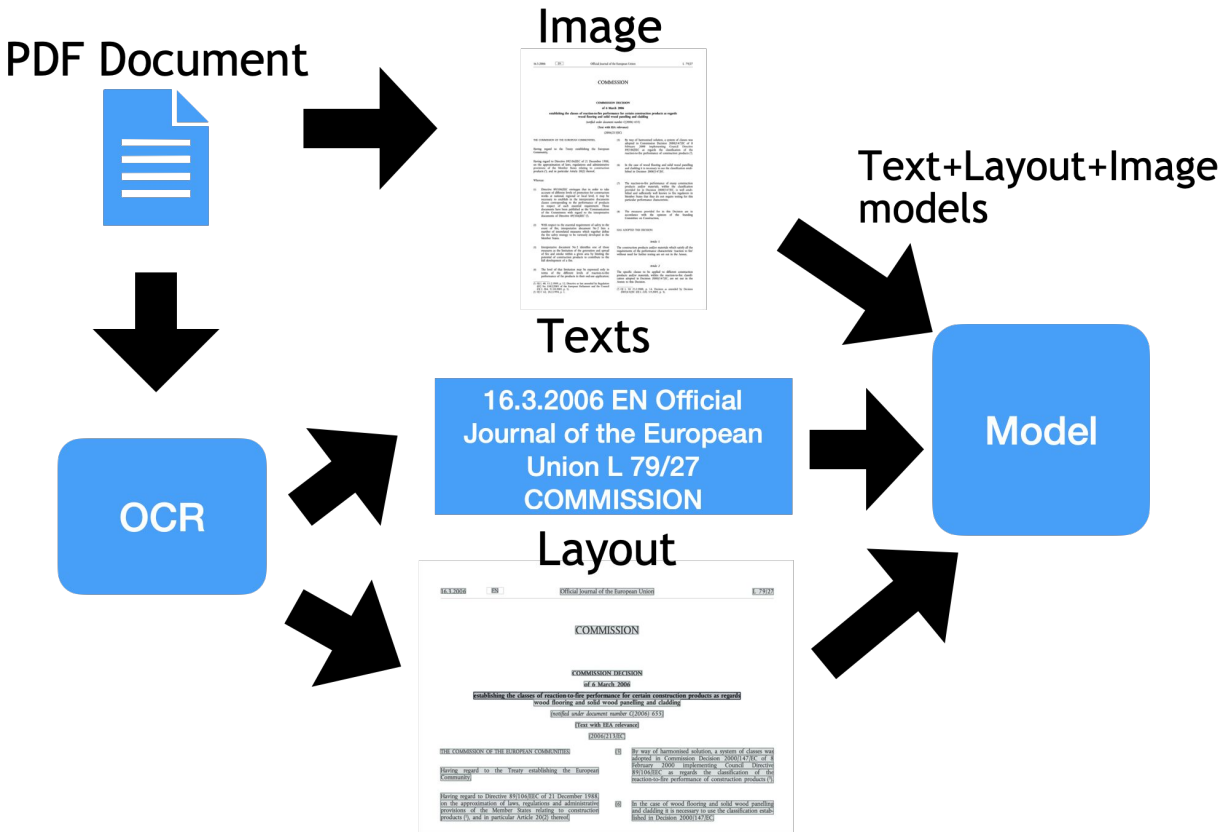
新しいデータでカバーできていないこと:

- 文書レイアウトの多様性
  - レイアウトが学習データの分布と大きく異なると精度が低下 [Chen+ 2023]

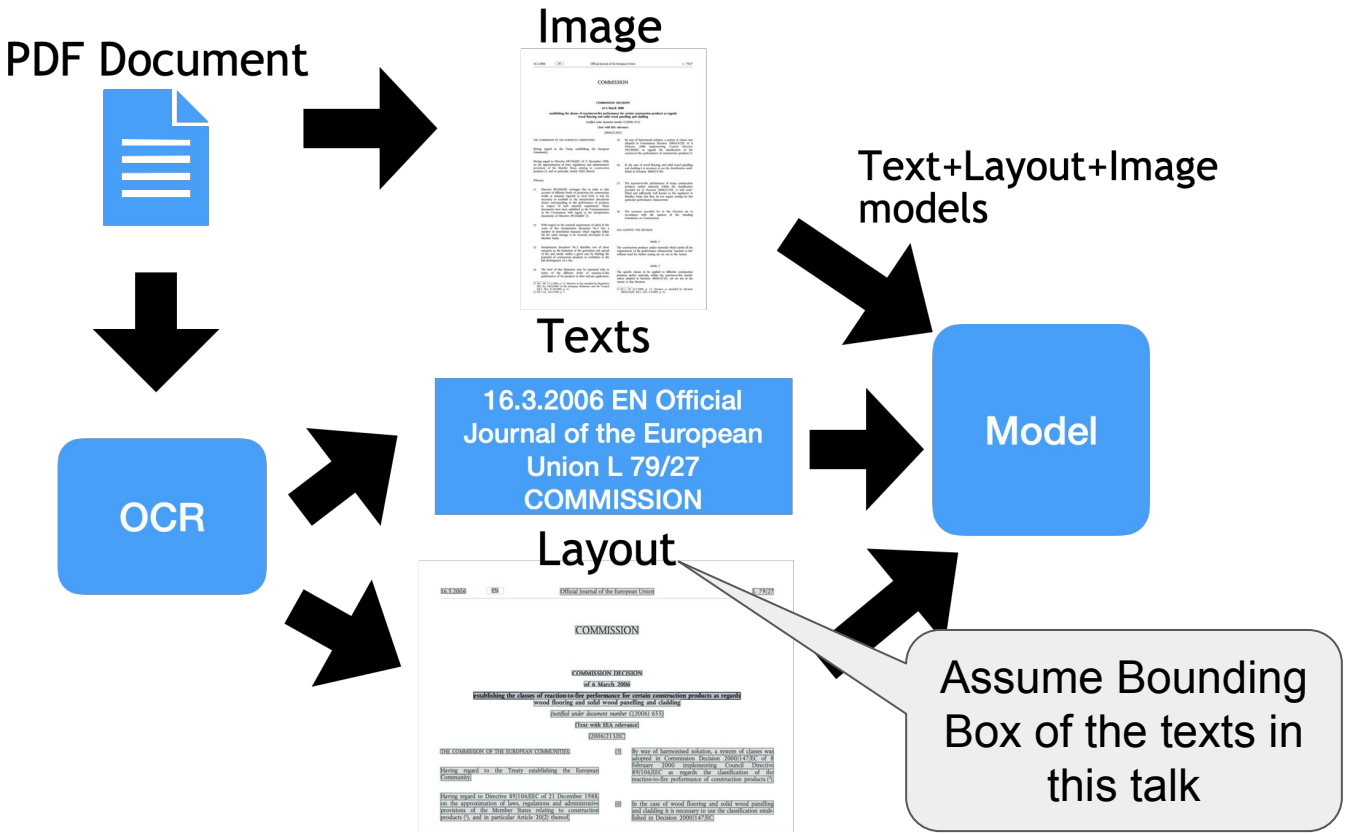
# 背景: Layout-Aware Modelsに関して

- Visually-Rich Docsの処理プロセス
- Visually-Rich Docsに特化したTransformerモデル

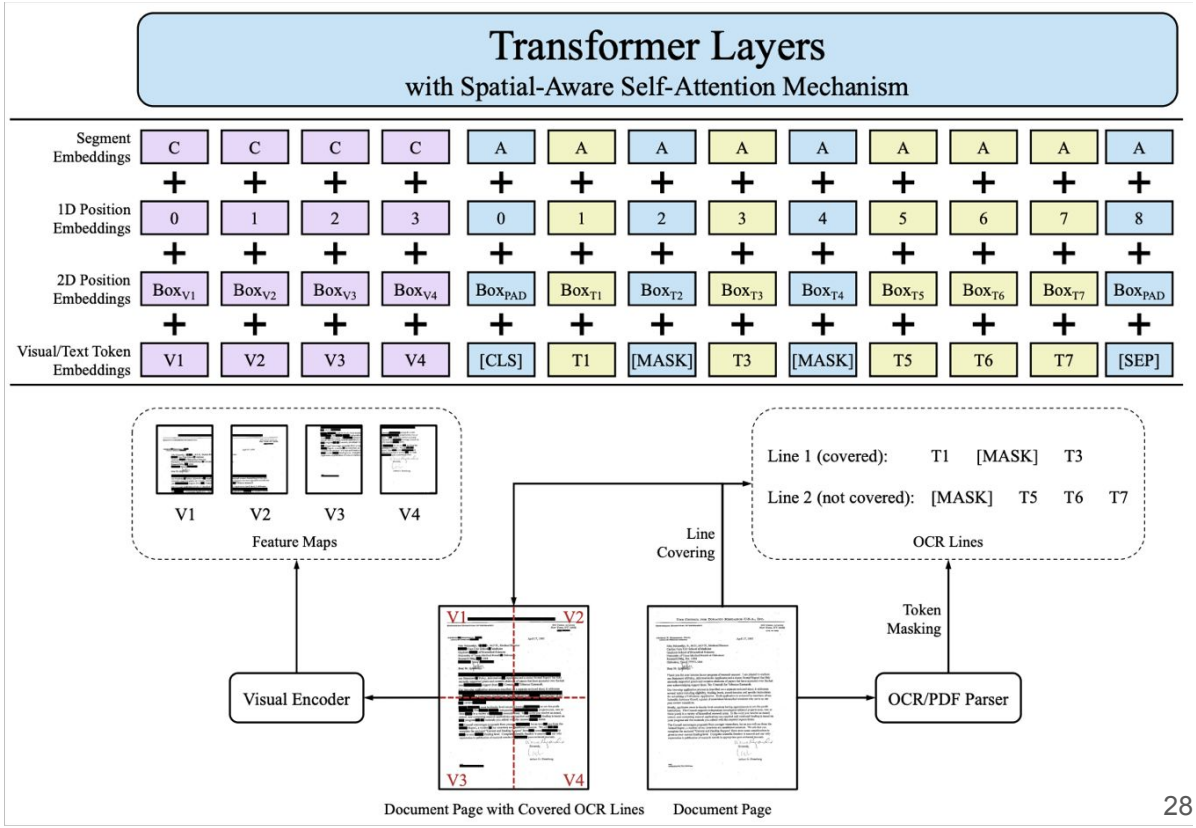
# Visually-Rich Documentの処理プロセス



# Visually-Rich Documentの処理プロセス

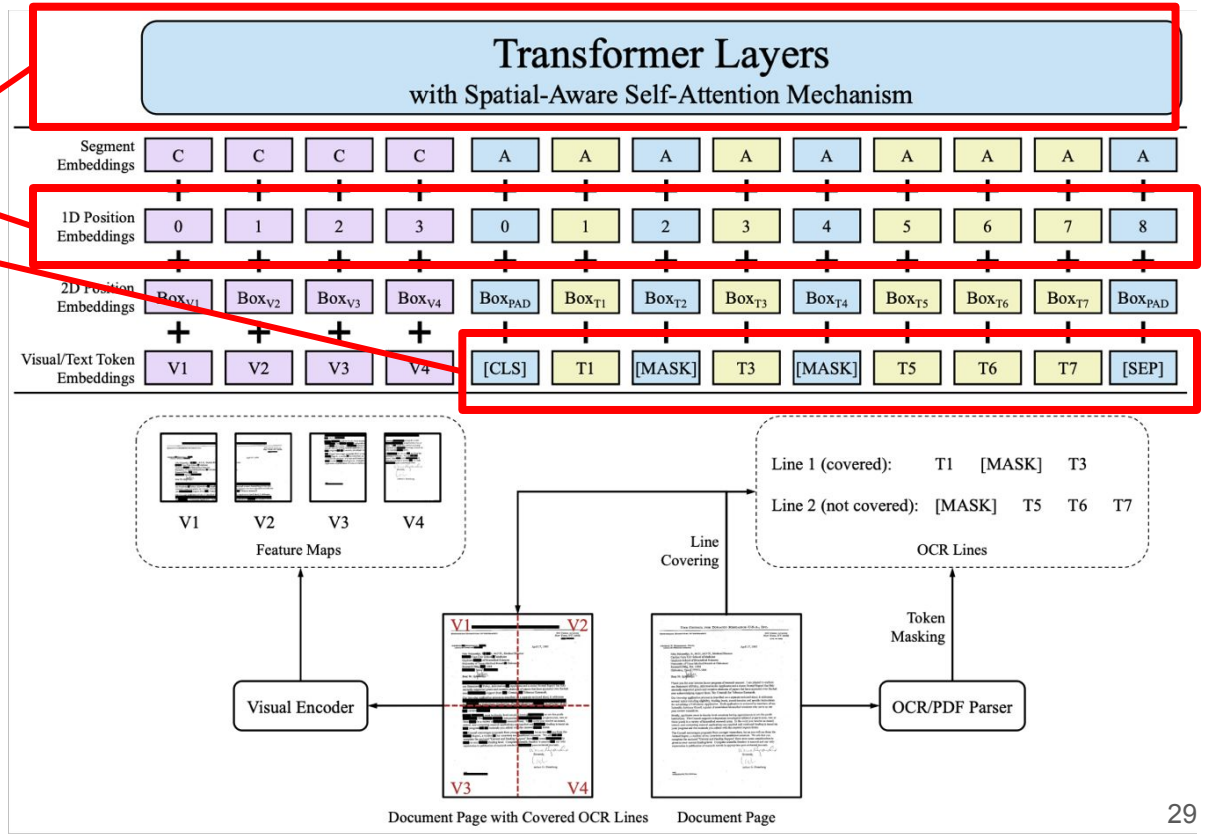


# 数少ない多言語Visually-Rich Doc Model: LayoutXLM [Xu+ 2021]



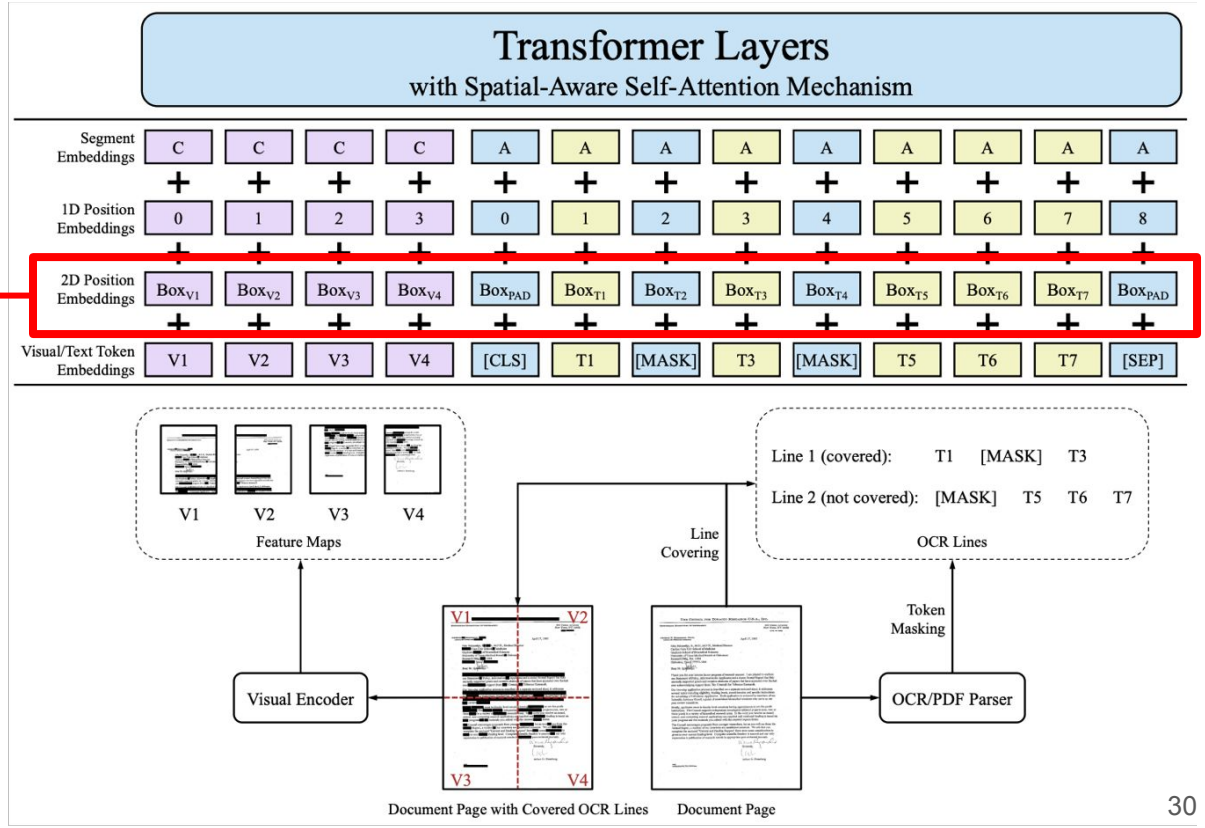
# 数少ない多言語Visually-Rich Doc Model: LayoutXLM [Xu+ 2021]

ここは通常のTransformer
と同じく
トークン＋位置埋め込み



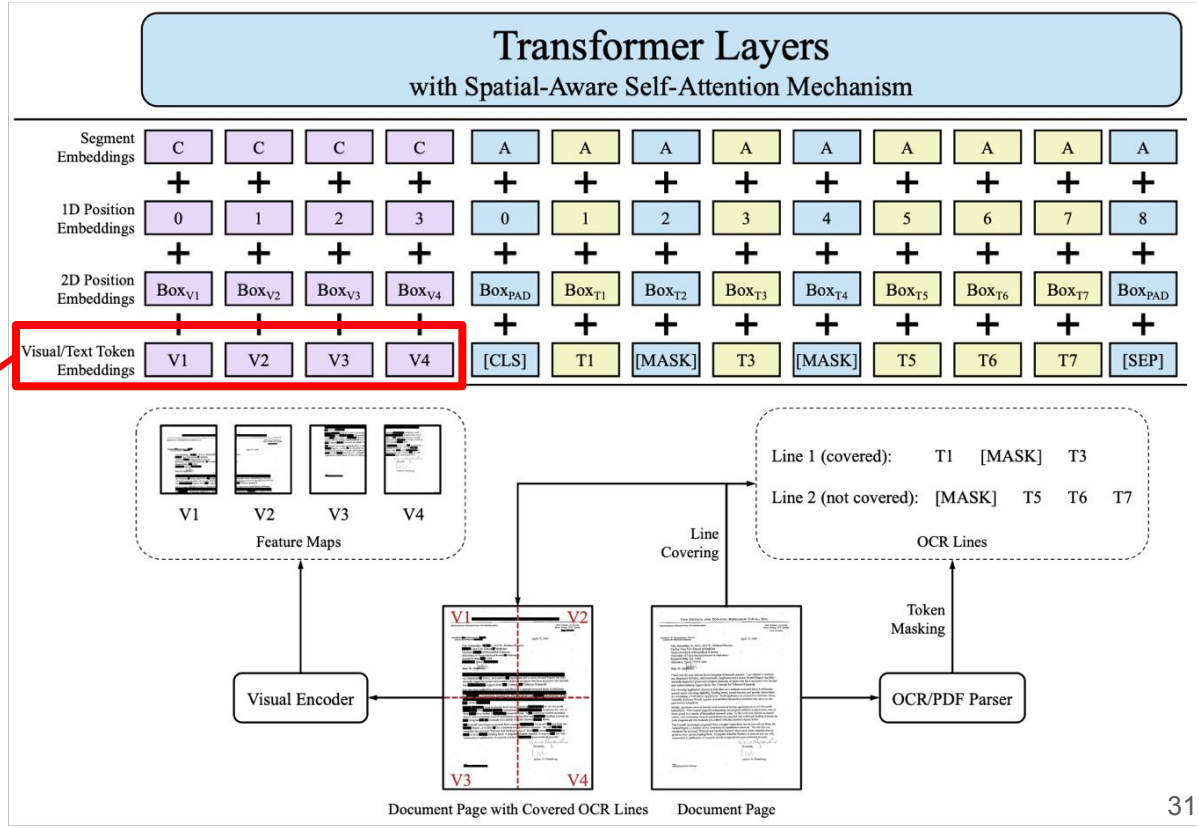
# 数少ない多言語Visually-Rich Doc Model: LayoutXLM [Xu+ 2021]

ここは通常と異なり
2次元位置埋め込み（文書レイアウト）

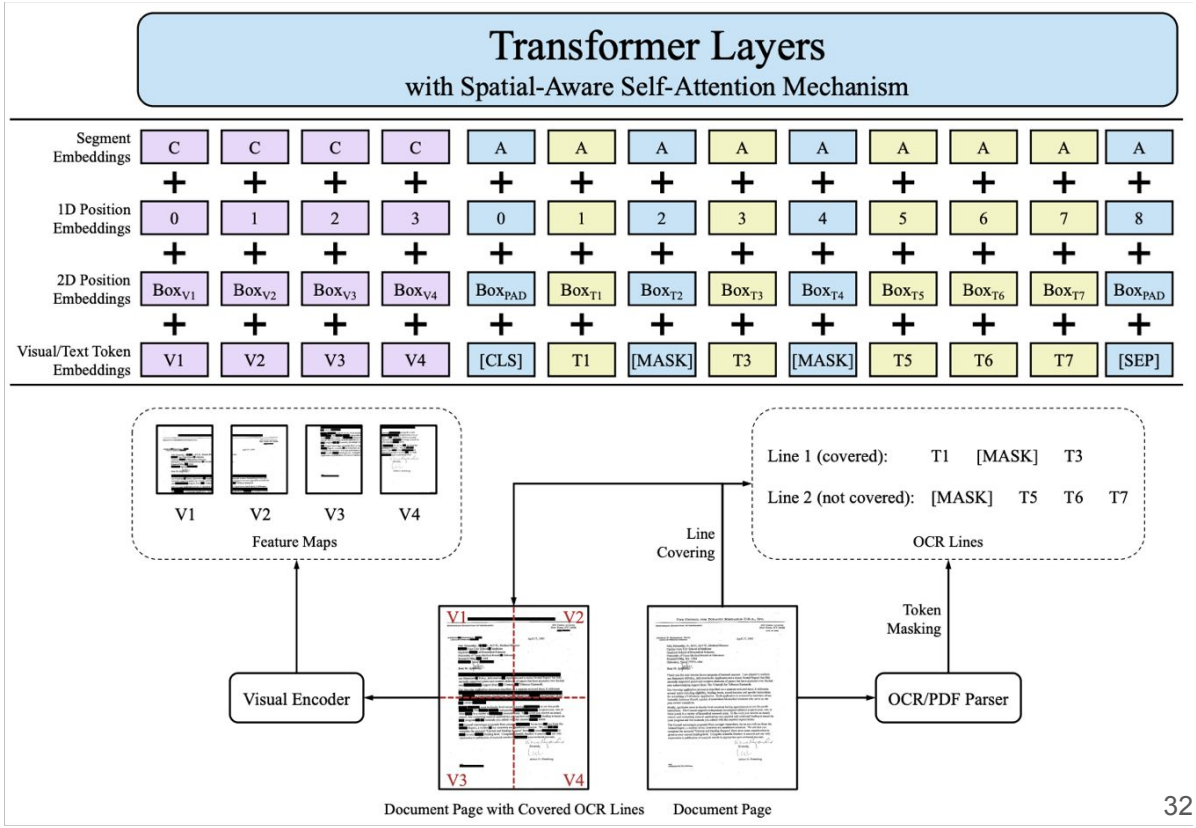


# 数少ない多言語Visually-Rich Doc Model: LayoutXLM [Xu+ 2021]

ここは通常と異なり
画像の埋め込み



# 数少ない多言語Visually-Rich Doc Model: LayoutXLM [Xu+ 2021]




事前学習データは
Common Crawlから抽出した63言語のPDF

# LayoutXLMは多言語には強い

- 学習と推論は同じ言語
- LayoutXLM（マルチモーダル）は精度が言語を問わず安定
  - 多言語PDFの事前学習と入力モダリティが豊富

| Models | en | da | de | nl | sv | ro | es | fr | it |
|---|---|---|---|---|---|---|---|---|---|
| InfoXLM | $64.98_{1.7}$ | $63.16_{1.2}$ | $63.89_{0.9}$ | $62.82_{3.6}$ | $64.08_{1.1}$ | $28.31_{24.7}$ | $63.2_{1.7}$ | $65.12_{0.5}$ | $64.74_{1.4}$ |
| LiLT | $61.56_{2.6}$ | $42.57_{28.9}$ | $61.48_{1.3}$ | $59.14_{2.9}$ | $63.78_{0.5}$ | $1.01_{0.3}$ | $63.1_{1.7}$ | $42.04_{30.6}$ | $62.84_{0.7}$ |
| LayoutXLM | $65.67_{0.5}$ | $65.17_{0.7}$ | $65.09_{0.5}$ | $65.07_{0.2}$ | $64.76_{1.0}$ | $64.15_{1.1}$ | $65.25_{0.3}$ | $65.36_{0.7}$ | $65.22_{0.3}$ |
| Donut | 29.29 | 31.69 | 26.15 | 25.66 | 21.94 | 19.28 | 24.01 | 30.92 | 33.97 |

| Models | pt | pl | bg | cs | hu | fi | el | et | Avg |
|---|---|---|---|---|---|---|---|---|---|
| InfoXLM | $64.01_{1.5}$ | $61.12_{0.8}$ | $14.23_{0.1}$ | $40.99_{34.9}$ | $58.84_{0.9}$ | $63.46_{1.0}$ | $63.95_{0.7}$ | $60.37_{0.8}$ | 56.89 |
| LiLT | $58.10_{2.4}$ | $58.85_{0.5}$ | $1.55_{2.1}$ | $37.60_{31.8}$ | $39.27_{33.8}$ | $61.75_{0.6}$ | $60.45_{1.8}$ | $59.26_{0.9}$ | 49.08 |
| LayoutXLM | $64.26_{0.2}$ | $63.26_{0.7}$ | $63.67_{0.6}$ | $63.6_{0.3}$ | $63.87_{0.8}$ | $63.52_{1.1}$ | $62.19_{0.3}$ | $63.43_{0.2}$ | 64.32 |
| Donut | 26.87 | 22.27 | 20.03 | 19.30 | 23.77 | 25.70 | 32.83 | 22.04 | 25.60 |

# LayoutXLMは言語間転移が難しい

- 学習と推論は異なる言語
- InfoXLM（テキストのみ）とLayoutXLMの精度に差あり
  - 特に学習言語である英語から離れているウラル系言語で（フィンランド語等）

| Models | da | de | nl | sv | ro | es | fr | it | pt |
|---|---|---|---|---|---|---|---|---|---|
| InfoXLM | $51.28_{1.3}$ | $53.27_{1.3}$ | $46.47_{1.4}$ | $47.91_{2.5}$ | $48.73_{2.8}$ | $52.63_{2.4}$ | $52.25_{2.6}$ | $47.75_{2.3}$ | $47.98_{1.0}$ |
| LiLT | $43.94_{6.1}$ | $44.30_{5.8}$ | $38.75_{5.4}$ | $42.11_{7.3}$ | $43.32_{5.7}$ | $47.41_{4.6}$ | $43.96_{6.3}$ | $45.43_{3.4}$ | $42.99_{5.5}$ |
| LayoutXLM | $51.29_{1.7}$ | $46.26_{1.5}$ | $46.49_{2.9}$ | $47.75_{1.5}$ | $50.15_{2.1}$ | $52.35_{1.1}$ | $52.50_{0.7}$ | $49.33_{1.6}$ | $48.46_{1.7}$ |
| Donut | 14.03 | 12.19 | 13.49 | 14.86 | 11.25 | 12.17 | 10.63 | 11.00 | 13.01 |

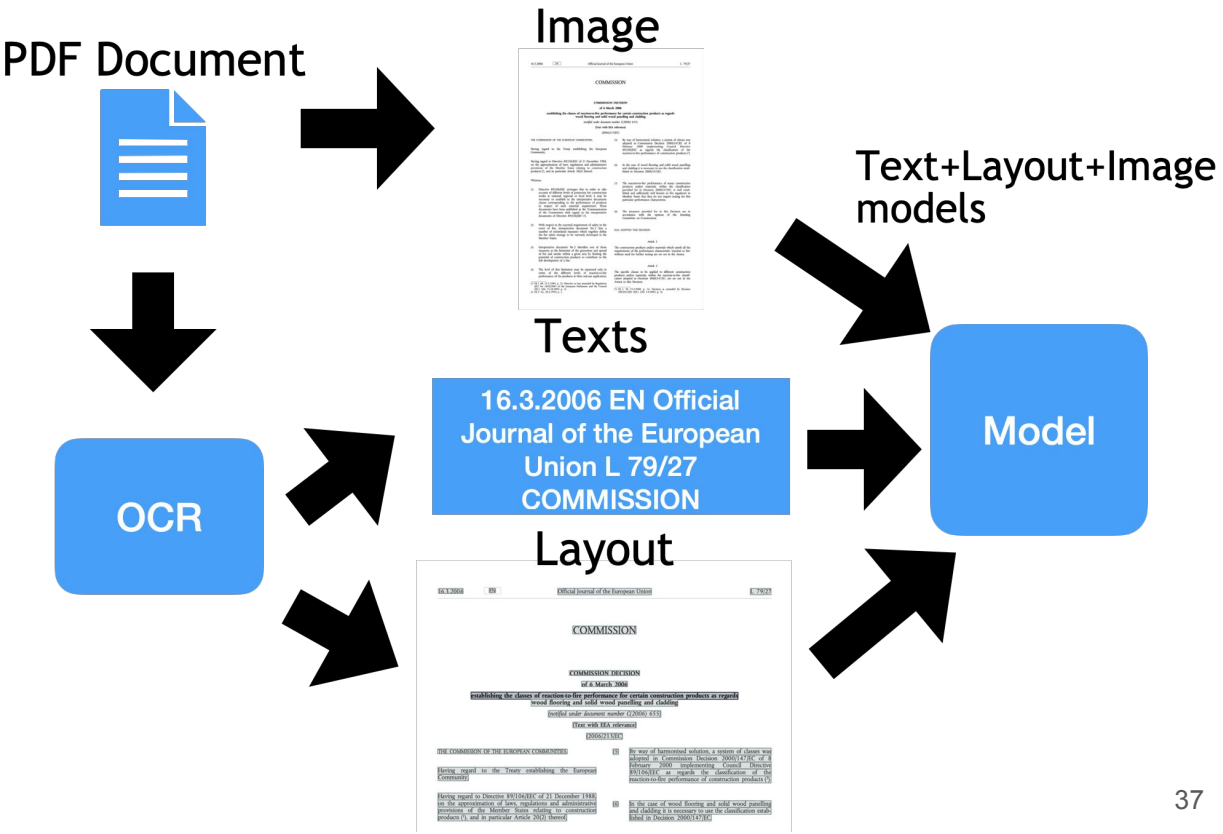
| Models | pl | bg | cs | hu | fi | el | et | Avg |
|---|---|---|---|---|---|---|---|---|
| InfoXLM | $41.62_{0.6}$ | $45.78_{2.3}$ | $46.35_{2.2}$ | $45.74_{3.4}$ | $42.86_{3.4}$ | $34.87_{2.4}$ | $41.78_{3.7}$ | 46.70 |
| LiLT | $35.49_{5.3}$ | $40.77_{6.9}$ | $37.28_{8.0}$ | $39.03_{6.8}$ | $34.02_{7.0}$ | $27.17_{4.1}$ | $34.41_{7.1}$ | 40.02 |
| LayoutXLM | $41.28_{2.7}$ | $47.31_{1.3}$ | $42.32_{2.2}$ | $39.36_{0.9}$ | $31.85_{1.5}$ | $27.15_{1.4}$ | $38.37_{1.8}$ | 44.51 |
| Donut | 12.33 | 9.32 | 13.66 | 8.97 | 12.25 | 9.73 | 16.19 | 12.19 |

# 今後の課題

- 画像を含むマルチモーダルモデルにおいてより良い言語間転移
- 文書分類問題以外での多言語データセットの構築

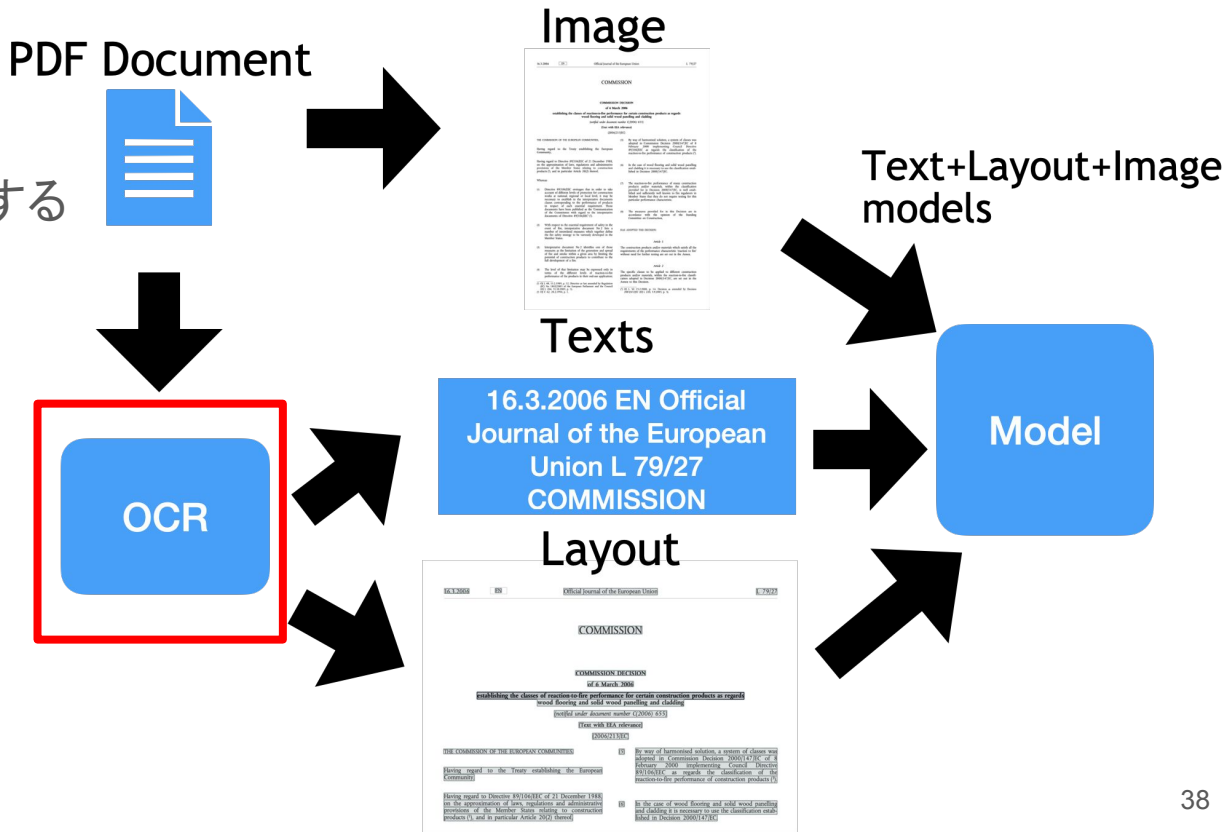
# Visually-Rich Documentを軸として見たVision-Languageモデル（VLM）

# そもそもなぜVLMがVisually-Rich Docに対して重要か？



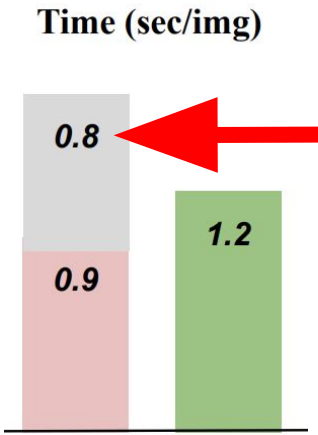
# そもそもなぜVLMがVisually-Rich Docに対して重要か？

- OCR非依存
- OCRは言語依存
  - 多言語にスケールする上でボトルネック

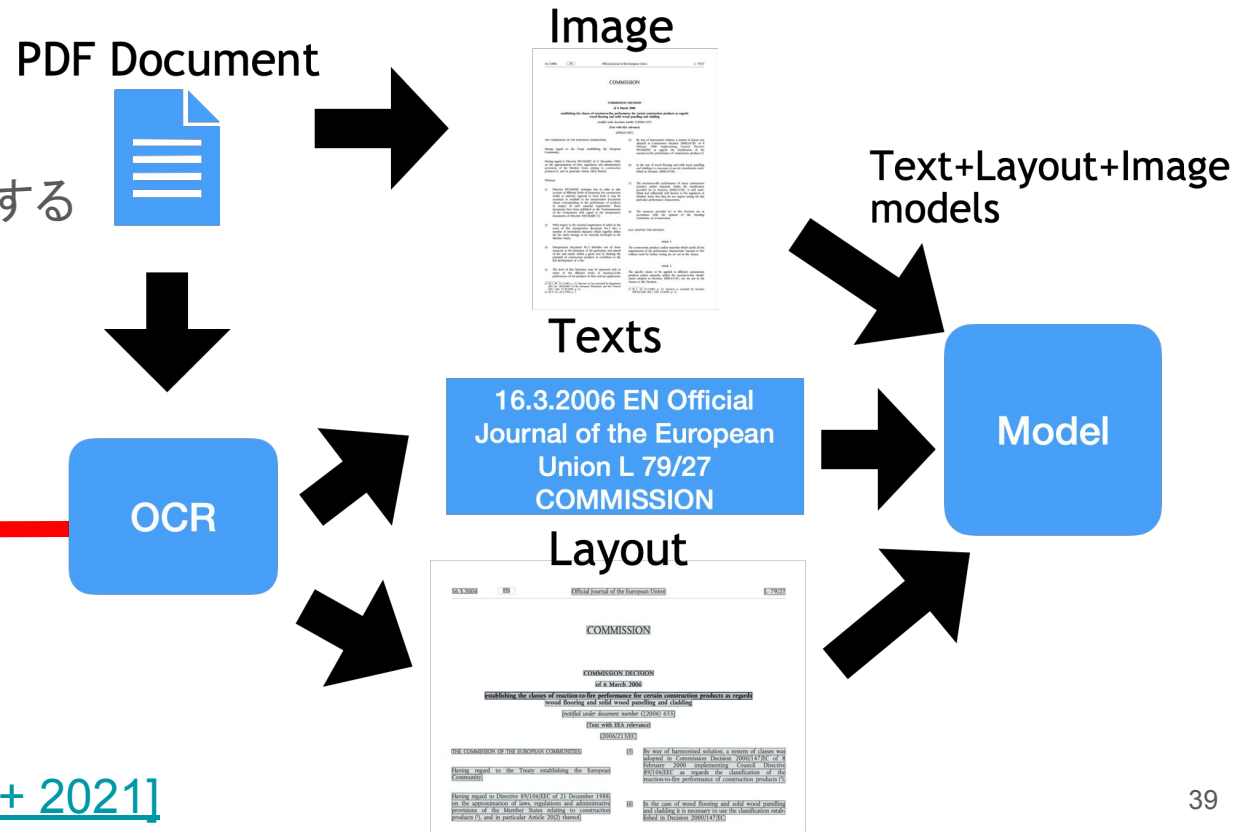


# そもそもなぜVLMがVisually-Rich Docに対して重要か？

- OCR非依存
- OCRは言語依存
  - 多言語にスケールする上でボトルネック
  - 推論時にも遅延

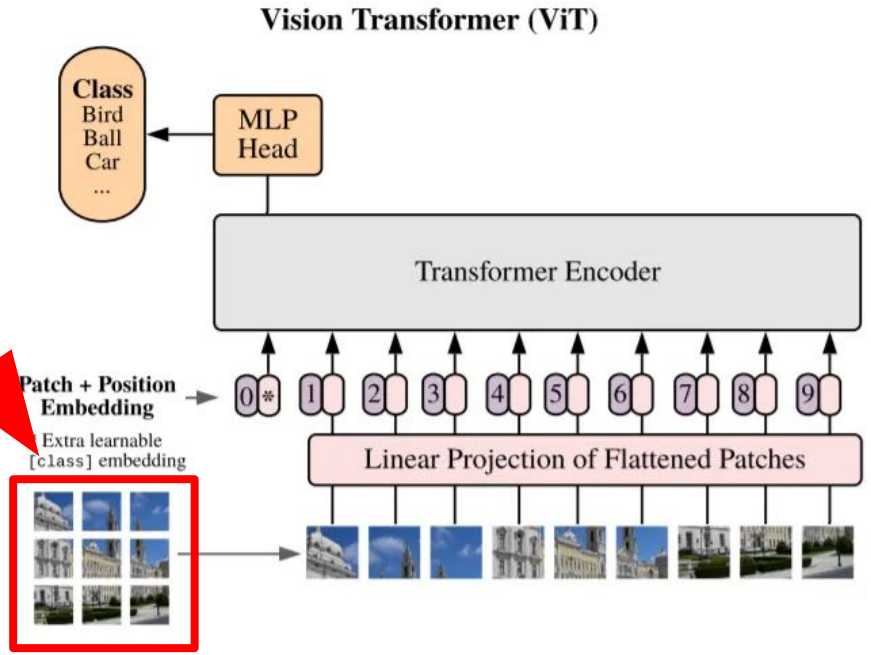



[Kim+ 2021]



# 最近までのVisual-Rich Docから見たVLMの問題点

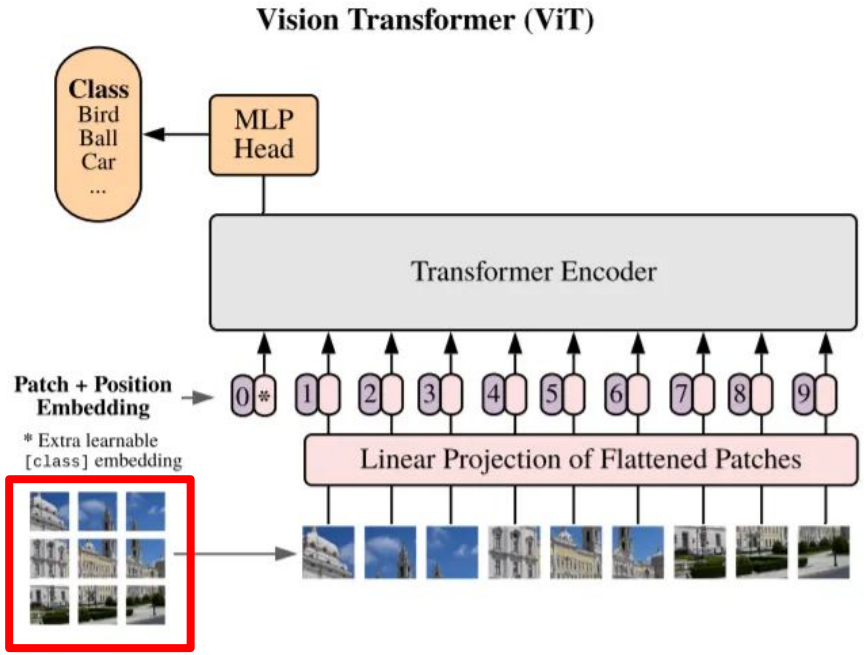
- **入力画像の解像度が低い**
  - 固定長の画像パッチを入力したい
  - 画像サイズ（解像度）Resizeする
  - ViTの224x224が（少し前まで？）主流
  - LayoutXLMも224x224
-




[Dosovitskiy+ 2021]より引用

# 最近までのVisual-Rich Docから見たVLMの問題点

- **入力画像の解像度が低い**
  - 固定長の画像パッチを入力したい
  - 画像サイズ（解像度）Resizeする
  - ViTの224x224が（少し前まで？）主流
  - LayoutXLMも224x224
- 画像はデータ量が多い・密
  - 建物の画像は問題ない
  - 文書の画像になると文字が潰れる




[Dosovitskiy+ 2021]より引用

# 解像度の異なる文書画像

224x224

448x448

# 解像度の異なる文書画像


224x224

448x448

# 解像度の異なる文書画像

224x224


448x448

# 解像度の異なる文書画像



文書画像はタスクや言語によって
必要な解像度が異なる

224x224

**448x448**

# 企業のTech Reportだと解像度が異なるモデルを比較

- AlibabaのQWEN-VL [Bai+ 2023]

Table 6: Results on Text-oriented VQA.

| Model type | Model | TextVQA | DocVQA | ChartQA | AI2D | OCR-VQA |
|---|---|---|---|---|---|---|
| Generalist Models | BLIP-2 (Vicuna-13B) | 42.4 | - | - | - | - |
| | InstructBLIP (Vicuna-13B) | 50.7 | - | - | - | - |
| | mPLUG-DocOwl (LLaMA-7B) | 52.6 | 62.2 | 57.4 | - | - |
| | Pic2Struct-Large (1.3B) | - | **76.6** | 58.6 | 42.1 | 71.3 |
| | **Qwen-VL (Qwen-7B)** | **63.8** | 65.1 | 65.7 | **62.3** | **75.7** |
| | **Qwen-VL-Chat** | 61.5 | 62.6 | **66.3** | 57.7 | 70.5 |
| Specialist SOTAs | PALI-X-55B (Single-task fine-tuning, without OCR Pipeline) | 71.44 | 80.0 | 70.0 | 81.2 | 75.0 |

# 企業のTech Reportだと解像度が異なるモデルを比較

- AlibabaのQWEN-VL [Bai+ 2023]

入力画像サイズ
224x224

Table 6: Results on Text-oriented VQA.

| | Model | TextVQA | DocVQA | ChartQA | AI2D | OCR-VQA |
|---|---|---|---|---|---|---|
| Generalist Models | BLIP-2 (Vicuna-13B) | 42.4 | - | - | - | - |
| | InstructBLIP (Vicuna-13B) | 50.7 | - | - | - | - |
| | mPLUG-DocOwl (LLaMA-7B) | 52.6 | 62.2 | 57.4 | - | - |
| | Pic2Struct-Large (1.3B) | - | **76.6** | 58.6 | 42.1 | 71.3 |
| | **Qwen-VL (Qwen-7B)** | **63.8** | 65.1 | 65.7 | **62.3** | **75.7** |
| | **Qwen-VL-Chat** | 61.5 | 62.6 | **66.3** | 57.7 | 70.5 |
| Specialist SOTAs | PALI-X-55B (Single-task fine-tuning, without OCR Pipeline) | 71.44 | 80.0 | 70.0 | 81.2 | 75.0 |

# 企業のTech Reportだと解像度が異なるモデルを比較

- AlibabaのQWEN-VL [Bai+ 2023]

Table 6: Results on Text-oriented VQA.

448x448

| | Model | TextVQA | DocVQA | ChartQA | AI2D | OCR-VQA |
|---|---|---|---|---|---|---|
| Generalist Models | LIP-2 (Vicuna-13B) | 42.4 | - | - | - | - |
| | InstructBLIP (Vicuna-13B) | 50.7 | - | - | - | - |
| | mPLUG-DocOwl (LLaMA-7B) | 52.6 | 62.2 | 57.4 | - | - |
| | Pic2Struct-Large (1.3B) | - | **76.6** | 58.6 | 42.1 | 71.3 |
| | **Qwen-VL (Qwen-7B)** | **63.8** | 65.1 | 65.7 | **62.3** | **75.7** |
| | **Qwen-VL-Chat** | 61.5 | 62.6 | **66.3** | 57.7 | 70.5 |
| Specialist SOTAs | PALI-X-55B (Single-task fine-tuning, without OCR Pipeline) | 71.44 | 80.0 | 70.0 | 81.2 | 75.0 |

# 企業のTech Reportだと解像度が異なるモデルを比較

- AlibabaのQWEN-VL [Bai+ 2023]

Table 6: Results on Text-oriented VQA.

| | Model | TextVQA | DocVQA | ChartQA | AI2D | OCR-VQA |
|---|---|---|---|---|---|---|
| Generalist Mod | BLIP-2 (Vicuna-13B) | 42.4 | - | - | - | - |
| | InstructBLIP (Vicuna-13B) | 50.7 | - | - | - | - |
| | mPLUG-DocOwl (LLaMA-7B) | 52.6 | 62.2 | 57.4 | - | - |
| | Pic2Struct-Large (1.3B) | - | **76.6** | 58.6 | 42.1 | 71.3 |
| | **Qwen-VL (Qwen-7B)** | **63.8** | 65.1 | 65.7 | **62.3** | **75.7** |
| | **Qwen-VL-Chat** | 61.5 | 62.6 | **66.3** | 57.7 | 70.5 |
| Specialist SOTAs | PALI-X-55B (Single-task fine-tuning, without OCR Pipeline) | 71.44 | 80.0 | 70.0 | 81.2 | 75.0 |

最大
1024x1024

# 企業のTech Reportだと解像度が異なるモデルを比較

- AlibabaのQWEN-VL [Bai+ 2023]

Table 6: Results on Text-oriented VQA.

| Model type | Model | TextVQA | DocVQA | ChartQA | AI2D | OCR-VQA |
|---|---|---|---|---|---|---|
| 448x448 | BLIP-2 (Vicuna-13B) | 42.4 | - | - | - | - |
| | InstructBLIP (Vicuna-13B) | 50.7 | - | - | - | - |
| Generalist Mo | mPLUG-DocOwl (LLaMA-7B) | 52.6 | 62.2 | 57.4 | - | - |
| | Pic2Struct-Large (1.3B) | - | **76.6** | 58.6 | 42.1 | 71.3 |
| | **Qwen-VL (Qwen-7B)** | **63.8** | 65.1 | 65.7 | **62.3** | **75.7** |
| | **Qwen-VL-Chat** | 61.5 | 62.6 | **66.3** | 57.7 | 70.5 |
| Specialist SOTAs | PALI-X-55B (Single-task fine-tuning, without OCR Pipeline) | 71.44 | 80.0 | 70.0 | 81.2 | 75.0 |

# 企業のTech Reportだと解像度が異なるモデルを比較

- AlibabaのQWEN-VL [Bai+ 2023]

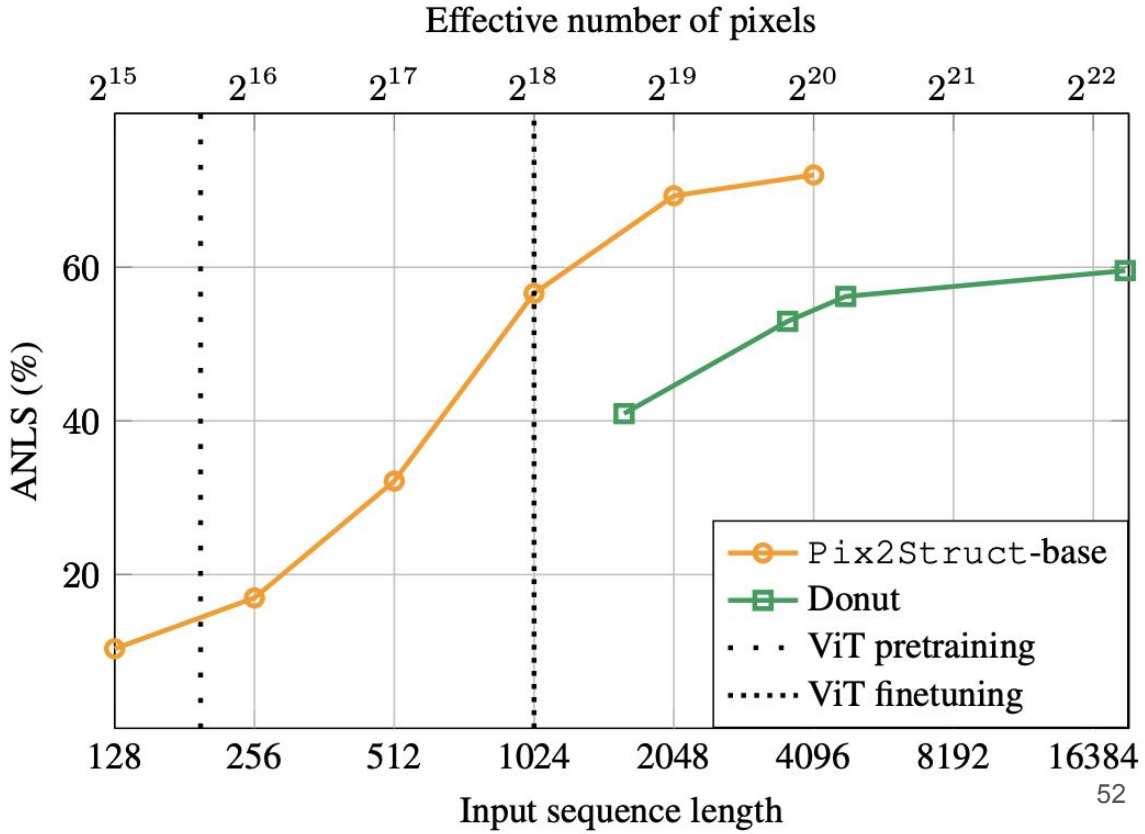
文書画像のQAで解像度が異なるモデルを比較するのは公平かつ有用な知見を得られるのか？

nted VQA.

| 解像度 | Model | TextVQA | DocVQA | ChartQA | AI2D | OCR-VQA |
|---|---|---|---|---|---|---|
| | BLIP-2 (Vicuna-13B) | 42.4 | - | - | - | - |
| | InstructBLIP (Vicuna-13B) | 50.7 | - | - | - | - |
| 448x448 | mPLUG-DocOwl (LLaMA-7B) | 52.6 | 62.2 | 57.4 | - | - |
| **最大1024x1024** | Pic2Struct-Large (1.3B) | - | **76.6** | 58.6 | 42.1 | 71.3 |
| 448x448 | **Qwen-VL (Qwen-7B)** | **63.8** | 65.1 | 65.7 | **62.3** | **75.7** |
| 448x448 | **Qwen-VL-Chat** | 61.5 | 62.6 | **66.3** | 57.7 | 70.5 |
| Specialist SOTAs | PALI-X-55B (Single-task fine-tuning, without OCR Pipeline) | 71.44 | 80.0 | 70.0 | 81.2 | 75.0 |

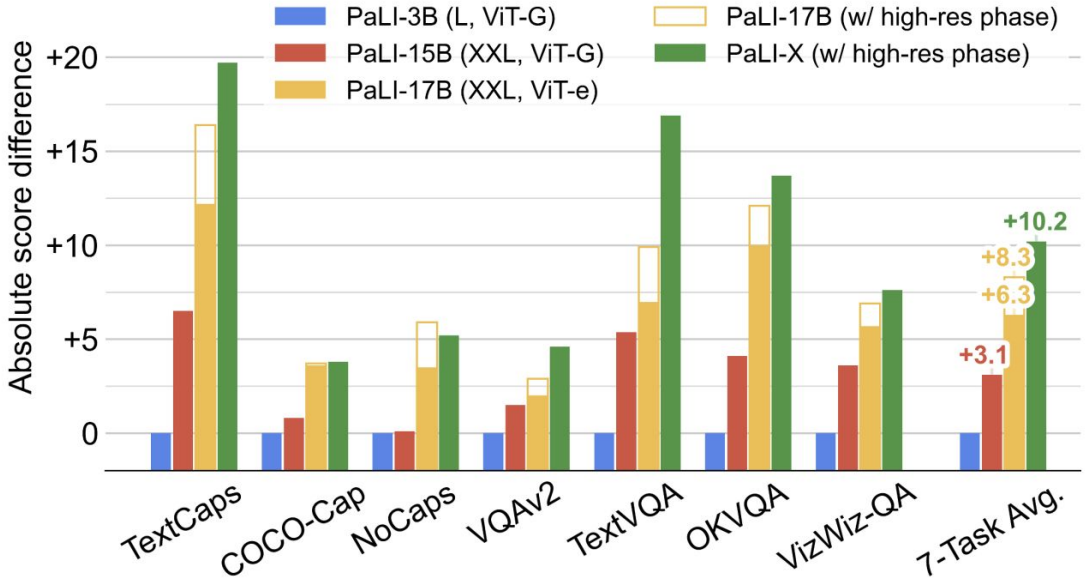
# Pix2structの貢献 [Appendix of Lee+ 2022]

- 横軸が入力解像度
- 縦軸が精度
- DocVQAにおいて、入力画像サイズの重要性が一目瞭然



# PaLI-Xの貢献 [Chen+ 2023]

- Text Transformer(33B)とVision Transformer(22B)両方スケール
- 事前学習で徐々に画像の解像度を上げていっている
  - 224×224→448×448→672×672→756×756



# 研究部分の総括と今後の課題

- 文書画像はタスクや言語によって必要な解像度が異なる
- 企業のTech Reportは半分宣伝資料？
- マルチモーダルモデルでの言語間転移を改善するには？
- 多言語データセットはまだまだ不足