

DOC2PPT: Automatic Slide Deck Generation from Documents

AAAI'22



Tsu-Jui Fu



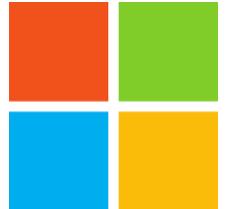
William Wang



Daniel McDuff



Yale Song



DOC2PPT

- Generate a slide from an academic paper



Figure 1: Some samples from the IMAGE-CHAT training set. For each sample we asked humans to engage in a conversation about the given image, where the two speakers, A & B, each have a given provided style.

B, and collect the dialogue using crowd-workers who are asked to help annotate those roles, and to be engaging to the other speaker while doing so. It was emphasized in the data collection instructions that the style trait describes a trait of the speaker, not properties of the content of the image they are discussing. Some examples from the training set are given in Figure 1.

4 Model

Data Quality During data collection crowd-sources were manually monitored, checking to ensure they were following the instructions. Poor

We consider two major types of dialogue model retrieval and generative. Both approaches make use of the same components as building blocks. We

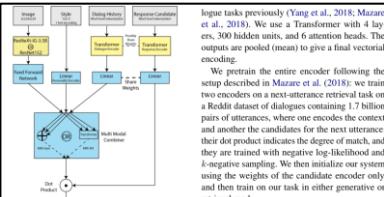


Figure 2: The TransResNETGEN multimodal architecture for grounded dialogue. There are several options: different image encoders (ResNet52 or ResNeXt-IG-3.5B), text encoders (shared or separate Transformers for history and response), and different multimodal combiners (sum or attention-based).

per ResNet152 features. We used the implementation provided in the torchvision project (Marcel and Rodriguez, 2010). The second is a ResNeXt-32x-48 (Xie et al., 2017) trained on 5.4 billion images.

| Model | Combiner | Text Encoders | Image Encoder | Turn 1 | Turn 2 | Turn 3 | All |
|--------------------------------|-----------|---------------|-----------------|--------|--------|--------|------|
| IR Baseline | n/a | n/a | n/a | 2.15 | 5.86 | | |
| TRANSRESNET _{IG-3.5B} | All | Separate | ResNeXt-IG-3.5B | 9.7 | 16.5 | 40.5 | 40.2 |
| TRANSRESNET _{IG-3.5B} | MM-Sum | Separate | ResNet152 | 34.5 | 46.0 | 50.0 | 47.0 |
| TRANSRESNET _{IG-3.5B} | MM-Shared | Shared | ResNeXt-IG-3.5B | 53.6 | 47.0 | 41.3 | 47.5 |
| TRANSRESNET _{IG-3.5B} | MM-Att | Separate | ResNeXt-IG-3.5B | 53.5 | 50.0 | 48.8 | 49.3 |
| TRANSRESNET _{IG-3.5B} | MM-Sum | Separate | ResNeXt-IG-3.5B | 54.0 | 51.9 | 44.8 | 50.3 |

Table 2: Module choices on IMAGE-CHAT. We compare different module variations of TRANSRESNET_{IG-3.5B}.

| Modules | TRANSRESNET _{IG-3.5B} (PROTUD) | | | TRANSRESNET _{IG-3.5B} (POVGEN) | | |
|------------------------------------|---|-------------|-------------|---|-------------|-------------|
| | Turn 1 | Turn 2 | Turn 3 | Turn 1 | Turn 2 | Turn 3 |
| Image Only | 37.6 | 28.1 | 20.7 | 26.7 | 21.1 | 21.9 |
| Style Only | 18.0 | 18.0 | 18.0 | 18.0 | 18.0 | 18.0 |
| Dialogue History Only | 1.0 | 33.7 | 32.3 | 22.3 | 18.9 | 22.7 |
| Style + Dialogue (no condition) | 18.3 | 45.4 | 43.1 | 35.4 | 20.4 | 24.1 |
| Image + Style (no condition) | 37.6 | 41.1 | 35.2 | 43.1 | 26.6 | 25.6 |
| Image + Style (no dialogue) | 54.0 | 41.1 | 35.2 | 43.1 | 23.7 | 23.2 |
| Image + Style + Image (full model) | 54.0 | 51.9 | 44.8 | 50.3 | 25.7 | 24.2 |

Table 3: Ablation on IMAGE-CHAT. We compare variants of our best TRANSRESNET generative and retrieval models (ResNeXt-IG-3.5B image encoder, and MM-Sum+ separate text encoders for retrieval) where we remove modality, image, dialogue history and relevant style (which is the same for both human author and model, so there is no advantage). We ask the evaluators in a blind test to choose the “more engaging”

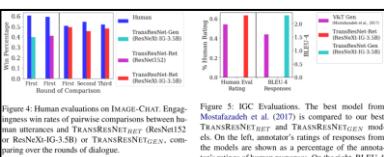


Figure 4: Human evaluations on IMAGE-CHAT. Engagement win rates of pairwise comparisons between human utterances and TRANSRESNET_{IG-3.5B} or ResNet152 or ResNeXt-IG-3.5B or TRANSRESNET_{IG-3.5B}, comparing over the rounds of dialogue.

base a response. This is clearly related to our task, except it focuses on answering questions, which our task does not. Our task is more varied as it was collected in an unconstrained way, unlike in IGC where they were asked to write a question. Nevertheless, an average question contains a ‘? or !’ state word, while ours, where most of our dataset contains 40,076 training utterances that are questions (11.3% of the data) and so it could be possible to produce responses to them. Without any fine-tuning at all, we thus simply took exactly the same best trained models and used them for their question answering task.

Unfortunately, after contacting the authors of Mostafazadeh et al. (2017) they no longer have the predictions of their model available, nor have they made available the code for their human evaluations. This work shows that our best proposed model can

DOC2PPT

Image-Chat

- Speaker b : a & b .
- We apply a set of 215 possible style traits , using an existing set from shuster et al .
- Who will be assigned to a person ?



A. Peaceful - B. Abstemious
A. I'm so thankful for this delicious food.
B: What is it called again?
A: Not sure but fried goodness.

A. Fearful - B. Miserable
A: I just heard something out there and I have no idea what it was.
B: It was probably a Wall coming to us in the middle of your speech.
A: I would never go camping in the woods for this very reason.

A. Evasive - B. Skeptical
A: What is the difference between the forest and the trees? Oh look, dry pine needles!
B: I doubt that's even a forest, it looks like a line of trees.
A: There's probably more lame lame poems on the other side!

Human Evaluations on IMAGE-CHAT

- Ablation study for both retrieval and generative models
- What is the best of both worlds ?
- Resnet 152 , resnet densenets .
- We ask the evaluators to choose the two possible utterances :

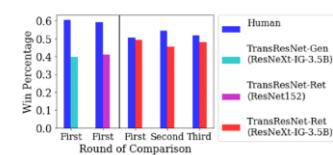


Figure 5: ICG Evaluations. The best model from Mostafazadeh et al. (2017) is compared to our best TRANSRESNET_{IG-3.5B} and TRANSRESNET_{IG-3.5B} (ResNet152).

On the left, the human's ratings of responses from the models are shown as a percentage of the annotator's ratings of human responses. On the right, BLEU-4 scores of the response task are shown.

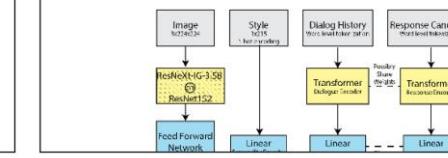
responses, to give a percentage. The results are shown in Figure 5. (left) human vs model gap between human and model performance, yielding a higher percentage of the human score (62.9% vs 54.2%). More detailed results and example predictions of our model can be found in Appendices E and F, including examples of highly rated and poorly rated outputs from our model.

6 Conclusion

This paper presents an approach for improving the way machines can generate grounded conversations that humans find engaging. Focusing on the case of chit-chatting about a given image, a naturally useful application for end-users of social dialogue agents, this work shows that our best proposed model can

Retrieval Models

- Two major types of dialogue model :
- In the retrieval model , the three modalities are fed into a combiner module .
- Resnet 152 , resnet densenets .
- Dialogue decoder : dialogue decoder the encoding from the image
- Style encoder to obtain its representation rs .

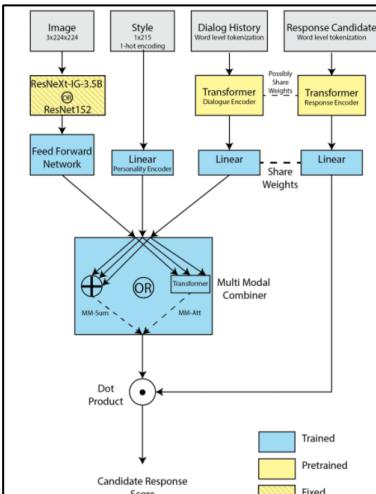


Conclusion

- Can be studied in future work .
- (zhang et al . , 2018)
- Humans can relate to social dialogue agents
- Retrieval models outperformed their generative models .
- A new dataset is made of a new dataset .

DOC2PPT

- Multi-modal summarizer
 - **Text Summarization + Figure Retrieval + Multi-Page**



logue tasks previously (Yang et al., 2018; Mazare et al., 2018). We use a Transformer with 4 layers, 300 hidden units, and 6 attention heads. The outputs are pooled (mean) to give a final vectorial encoding.

We pretrain the entire encoder following the setup described in Mazare et al. (2018): we train two encoders on a next-utterance retrieval task on a Reddit dataset of dialogues containing 1.7 billion pairs of utterances, where one encodes the context and another the candidates for the next utterance; their dot product indicates the degree of match, and they are trained with negative log-likelihood and k -negative sampling. We then initialize our system using the weights of the candidate encoder only, and then train on our task in either generative or retrieval mode.

4.1 Retrieval Models

Multimodal combiner module We consider two possible combiner modules for the inputs:

Multimodal sum combiner (MM-sum): Given an input image, style trait and dialogue (I, S, D), together with a candidate response C , the score of the final combination is computed as $s(I, S, D, C) = (r_I + r_S + r_D) \cdot r_C$.

Multimodal attention combiner (MM-att): A more sophisticated approach is to use an attention mechanism to choose which modalities are most relevant for each example by stacking Transformers. We concatenate the three representation vectors r_I , r_S and r_D and feed them to a second

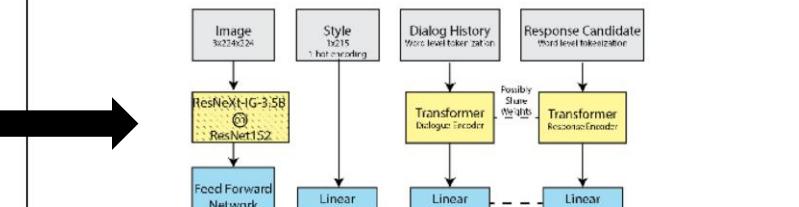
per as *ResNet152* features. We used the implementation provided in the torchvision project (Marcel and Rodriguez, 2010). The second is a *ResNeXt* $32 \times 48d$ (Xie et al., 2017) trained on 3.5 billion In-

Text Summarization

Figure Retrieval

Retrieval Models

- Two major types of dialogue model :
- In the retrieval model , the three modalities are fed into a combiner module .
- Resnet 152 , resnet densenets .
- Dialogue decoder : dialogue decoder the encoding from the image
- Style encoder to obtain its representation rs .



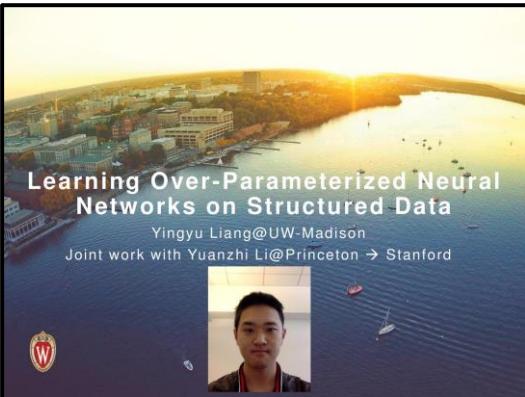
Multi-Page

Dataset Building

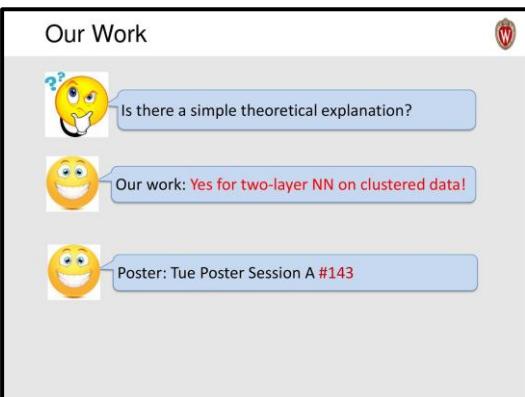
- Crawl **paper-slide pairs** from AI conferences
 - Computer Vision (CVPR, ECCV, ...)
 - Natural Language Processing (ACL, NAACL, ...)
 - Machine Learning (ICLR, ICML, ...)
- **5,873** in total
 - 4,686 / 592 / 595 (train / val / test)
- To prepare the data for training, needs some **preprocessing** in advance

Dataset Building

- Extract **text content** from a slide
 - Azure CV to do **Optical Character Recognition (OCR)**



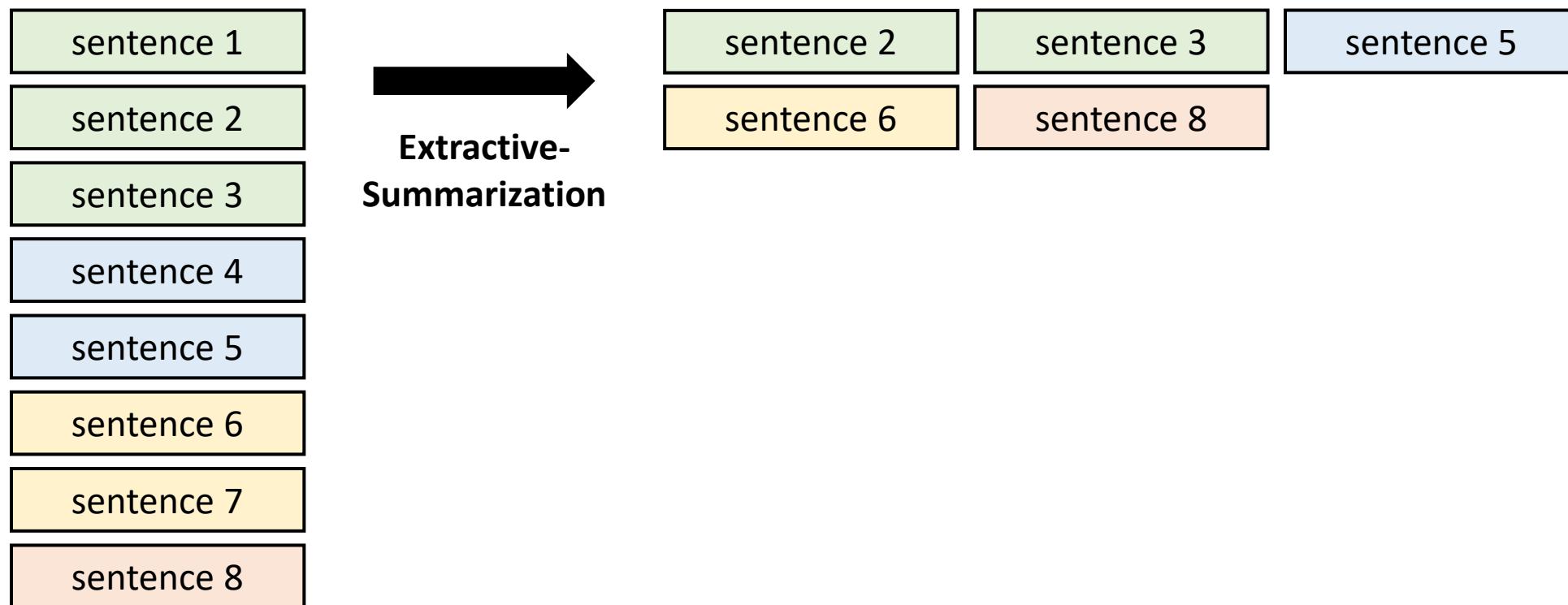
- Learning Over-Parameteiized –Neural
- Networks on Structured Data
- Yingyu Liang@UWLMadison
- Joint work with Yuanzhi Li@Princeton -Y Stanford



- Our Work
- Is there a simple theoretical explanation?
- Our work: Yes for two-layer NN on clustered data!
- Poster: Tue Poster Session A #143

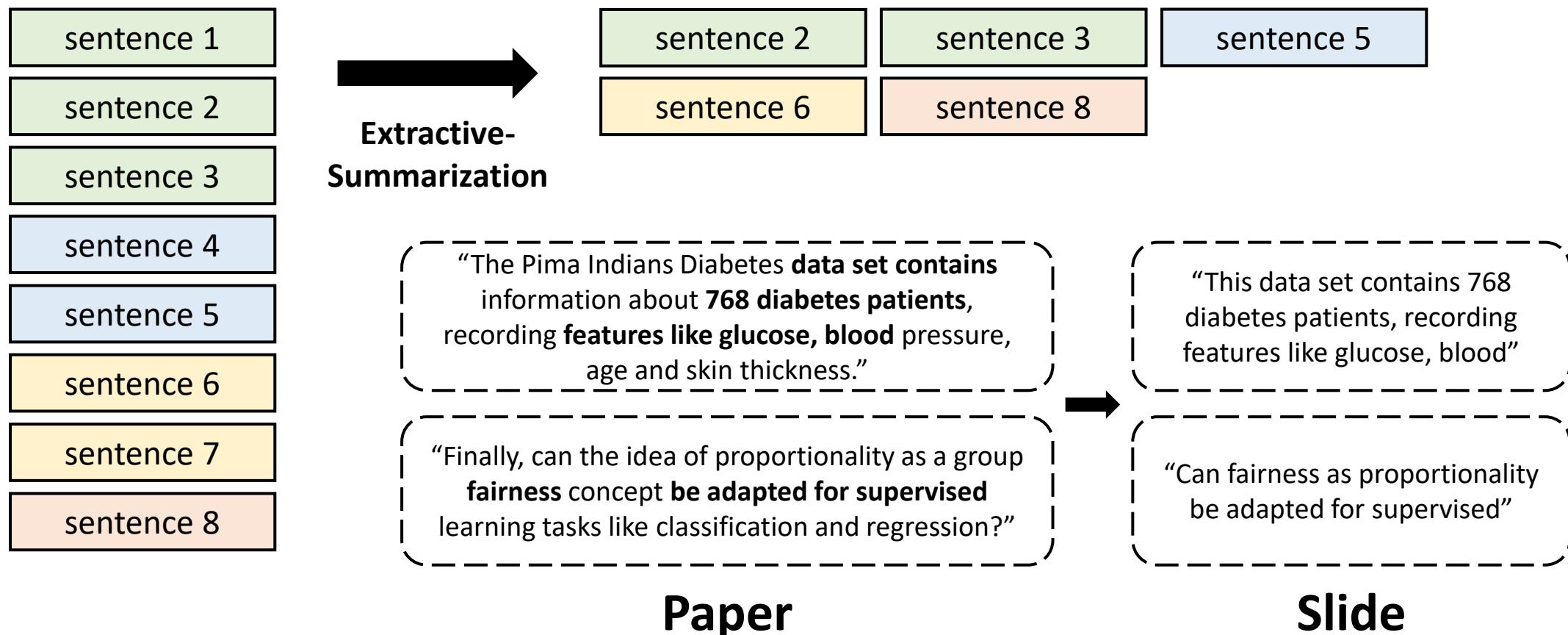
Dataset Building

- Match sentences from slide to paper
 - Extractive-based summarization



Dataset Building

- Match sentences from slide to paper
 - Extractive-based summarization



Dataset Building

- Match figures from slide to paper
 - CNN feature to do similarity matching

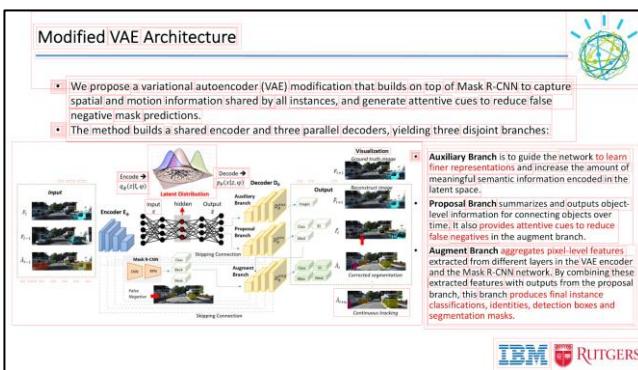
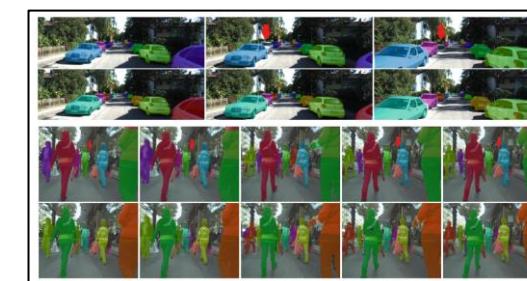
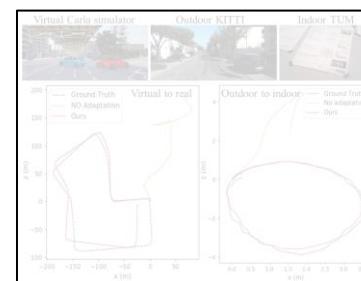


| Method | MOTSOTA | MOTS17 | TP↑ | FP↓ | FN↓ | |
|------------------------------------|---------|--------|------|-------|-----|-----|
| (a) KITTI MOTS Dataset [59] - Cars | 74.9 | 85.8 | 80.1 | 7,109 | 148 | 920 |
| Mask R-CNN [34]+HT [6] | 75.5 | 86.1 | 80.5 | 7,135 | 140 | 894 |
| MaskTrack R-CNN [60] | 75.5 | 86.1 | 80.5 | 7,135 | 140 | 894 |
| TrackR-CNN [61] | 75.5 | 86.1 | 80.5 | 7,135 | 140 | 894 |
| Ours | 77.6 | 88.8 | 87.7 | 7,355 | 130 | 674 |

| Method | MOTSOTA | MOTS17 | TP↑ | FP↓ | FN↓ | |
|---|---------|--------|------|-------|-----|-----|
| (b) KITTI MOTS Dataset [59] - Pedestrians | 44.6 | 63.8 | 74.1 | 2,479 | 295 | 868 |
| Mask R-CNN [34]+HT [6] | 45.9 | 64.6 | 74.9 | 2,497 | 280 | 850 |
| MaskTrack R-CNN [60] | 46.8 | 65.1 | 75.7 | 2,523 | 267 | 822 |
| TrackR-CNN [61] | 49.7 | 67.6 | 77.0 | 2,607 | 251 | 740 |

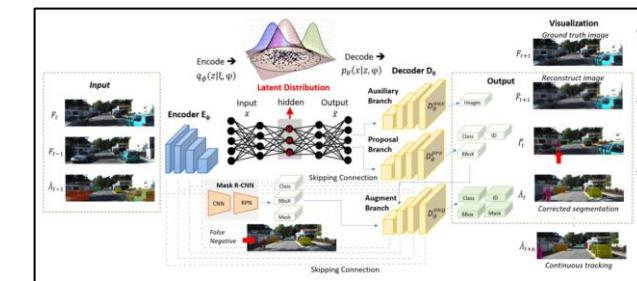
| Method | MOTSOTA | MOTS17 | TP↑ | FP↓ | FN↓ | |
|-------------------------------|---------|--------|------|--------|-------|-------|
| (c) MOTChallenge Dataset [69] | 48.6 | 65.5 | 77.6 | 16,726 | 1,039 | 7,318 |
| Mask R-CNN [34]+HT [6] | 48.6 | 65.5 | 77.6 | 16,862 | 1,082 | 7,012 |
| MaskTrack R-CNN [60] | 52.1 | 67.5 | 79.5 | 20,255 | 1,702 | 6,639 |
| TrackR-CNN [61] | 59.5 | 71.5 | 84.7 | 21,253 | 1,537 | 5,641 |

| Method | MOTSOTA | MOTS17 | TP↑ | FP↓ | FN↓ | |
|------------------------------|---------|--------|------|-------|-----|-----|
| (d) YouTube-VIS Dataset [66] | 33.7 | 46.4 | 78.8 | 2,751 | 790 | 596 |
| Mask R-CNN [34]+HT [6] | 34.1 | 47.2 | 78.7 | 2,767 | 780 | 580 |
| MaskTrack R-CNN [60] | 34.6 | 48.3 | 79.8 | 2,801 | 778 | 546 |
| TrackR-CNN [61] | 35.1 | 50.4 | 80.8 | 2,866 | 783 | 481 |



nts: performance comparison

| Method | Full | DiffCat | Cat | Cat&attr | Cat&cat | WithoutDist |
|--------------------|------|---------|------|----------|---------|-------------|
| Chance | 0.4 | 1.7 | 1.8 | 1.9 | 1.7 | 6.6 |
| GroundreR [35] | 19.1 | 60.2 | 38.5 | 35.7 | 38.9 | 75.7 |
| Deaf-GroundreR | 2.2 | 7.7 | 7.9 | 8.0 | 8.0 | 27.1 |
| Shuffle-GroundreR | 13.1 | 41.8 | 28.6 | 27.2 | 27.6 | 58.5 |
| Obj-Attr-GroundreR | 15.2 | 53.1 | 32.6 | 29.6 | 32.7 | 68.8 |
| MattNet-refCOCO | 8.7 | 22.7 | 17.0 | 16.7 | 18.9 | 42.4 |
| MattNet [44] | 26.3 | 69.1 | 45.2 | 42.5 | 45.8 | 77.9 |
| CM-Att-Erase [27] | 28.0 | 71.3 | 47.1 | 43.4 | 48.4 | 80.4 |
| SCAN [22]+MattNet | 18.8 | - | - | - | - | - |
| MattNet-Mine | 33.8 | 70.5 | 54.4 | 46.8 | 52.0 | 78.4 |

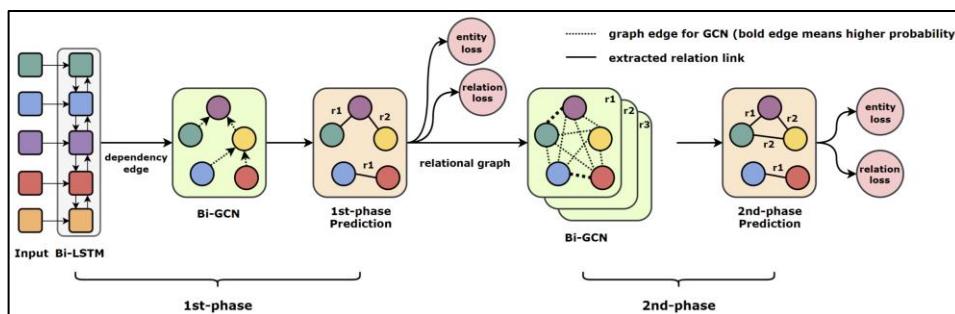


Slide

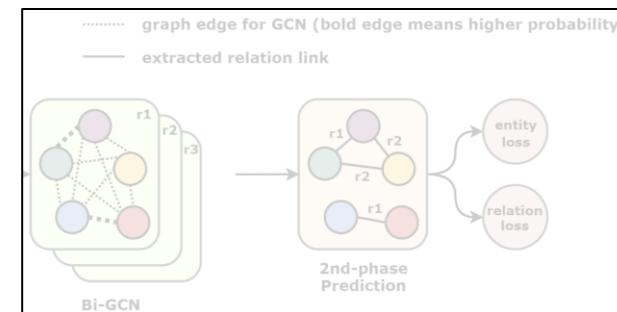
Figure from Paper

Dataset Building

- Match figures from slide to paper
- Not always perfect (currently 50.5% F1)
 - Leave as future work for better label to learn from



Partial Matching



Different Expression

| Method | NYT | | | WebNLG | | |
|------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| NovelTagging | 62.4% | 31.7% | 42.0% | 52.5% | 19.3% | 28.3% |
| OneDecoder | 59.4% | 53.1% | 56.0% | 32.2% | 28.9% | 30.5% |
| MultiDecoder | 61.0% | 56.6% | 58.7% | 37.7% | 36.4% | 37.1% |
| GraphRel _{1p} | 62.9% | 57.3% | 60.0% | 42.3% | 39.2% | 40.7% |
| GraphRel _{2p} | 63.9% | 60.0% | 61.9% | 44.7% | 41.1% | 42.9% |

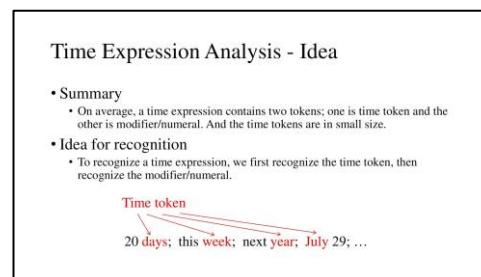
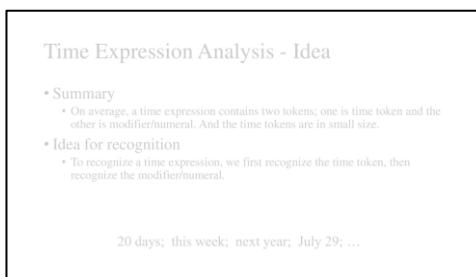
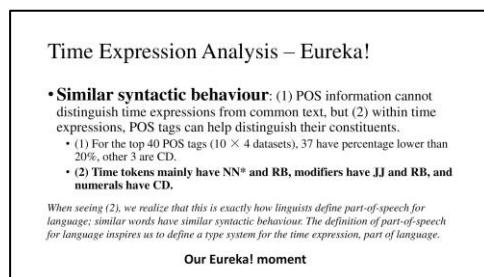
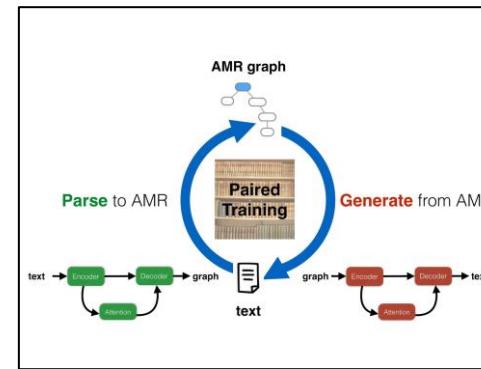
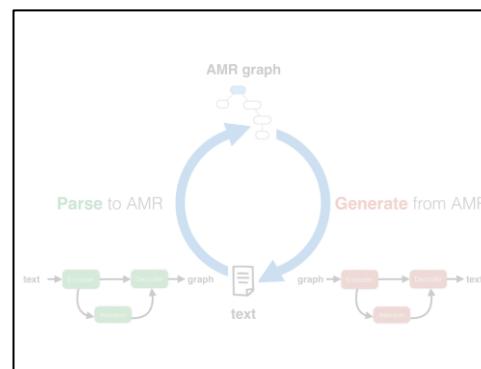
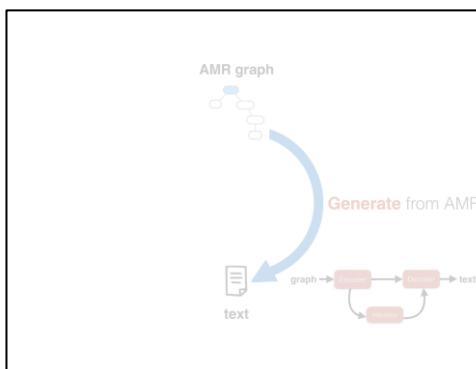
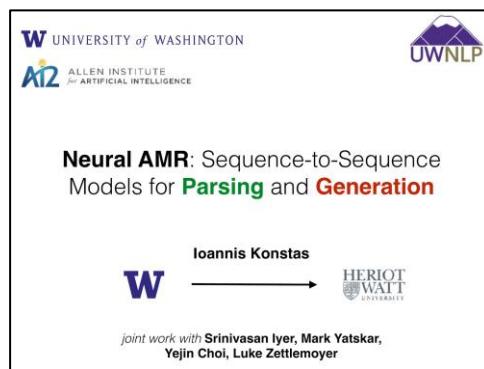
| Method | P | R | F1 | NER |
|------------------------|--------------|--------------|--------------|--------------|
| NovelTag | 62.4% | 31.7% | 42.0% | - |
| CopyRE | 61.0% | 56.6% | 58.7% | - |
| GraphRel _{1p} | 62.9% | 57.3% | 60.0% | 88.8% |
| GraphRel _{2p} | 63.9% | 60.0% | 61.9% | 89.2% |

Dataset Building

- **Match figures** from slide to paper
- **Not** always perfect (currently 50.5% F1)
- Apply **human labeling** for testing set
 - **Golden testing set** for fair evaluation

Dataset Building

- Remove the **progressive** page
 - **OCR cover rate > 80% (Acc ~90%)**
 - Keep the last one



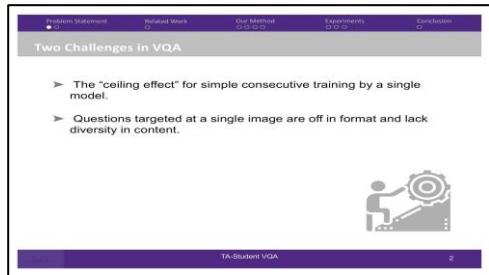
Dataset Building

- Generate pages **for each section** and combine them all
 - BERT to match **text** (page) with **paragraph** (section)
 - Consider **continuity**

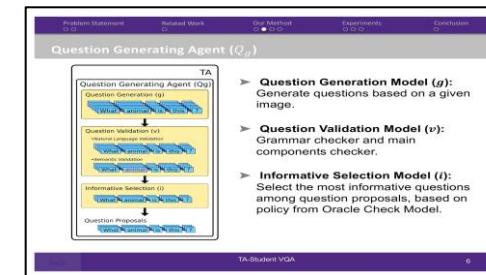
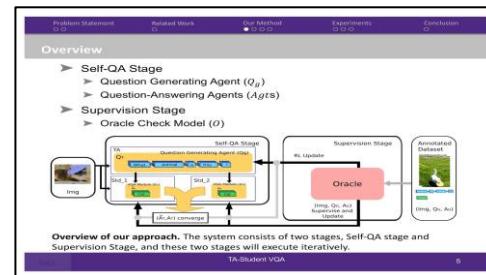
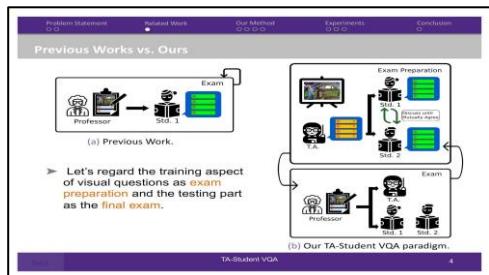


Dataset Building

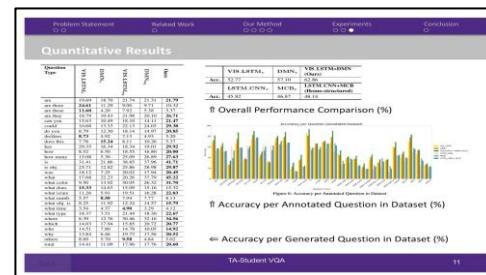
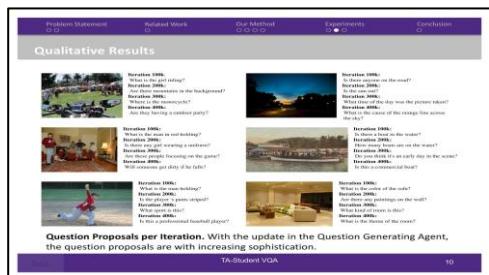
- Generate pages for each section and combine them all
 - BERT to match **text** (page) with **paragraph** (section)



→ 1. Introduction

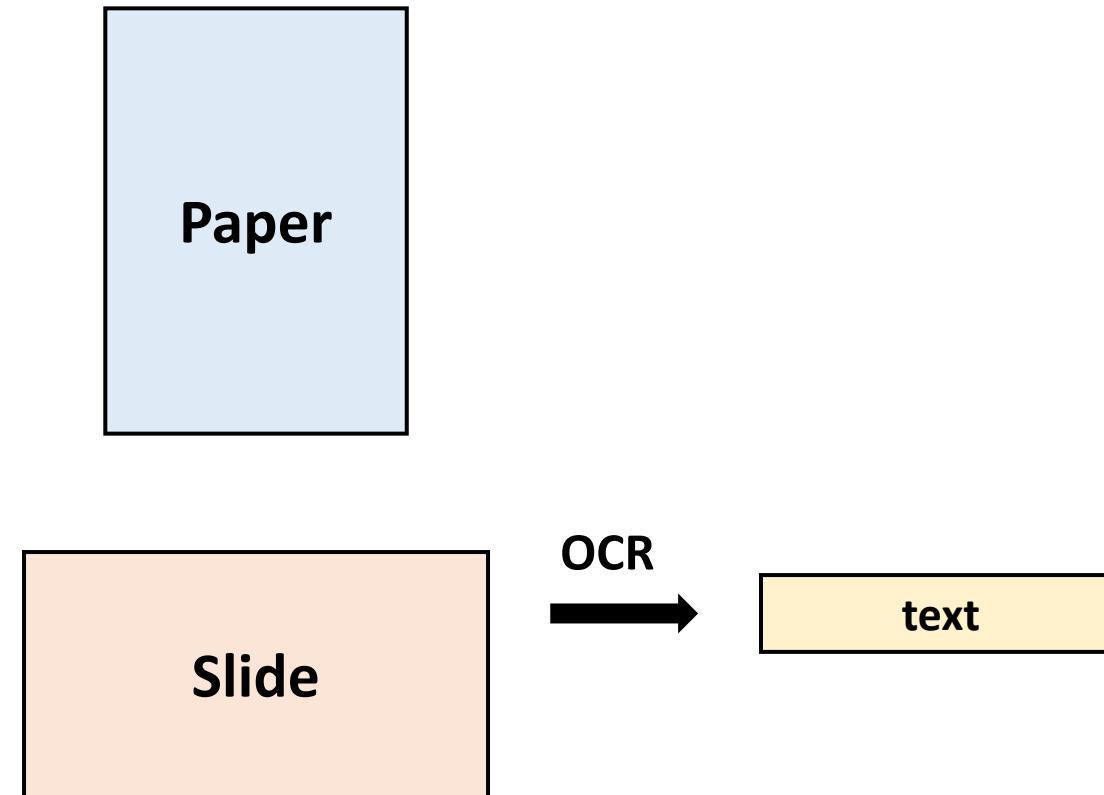


→ 3. Approach

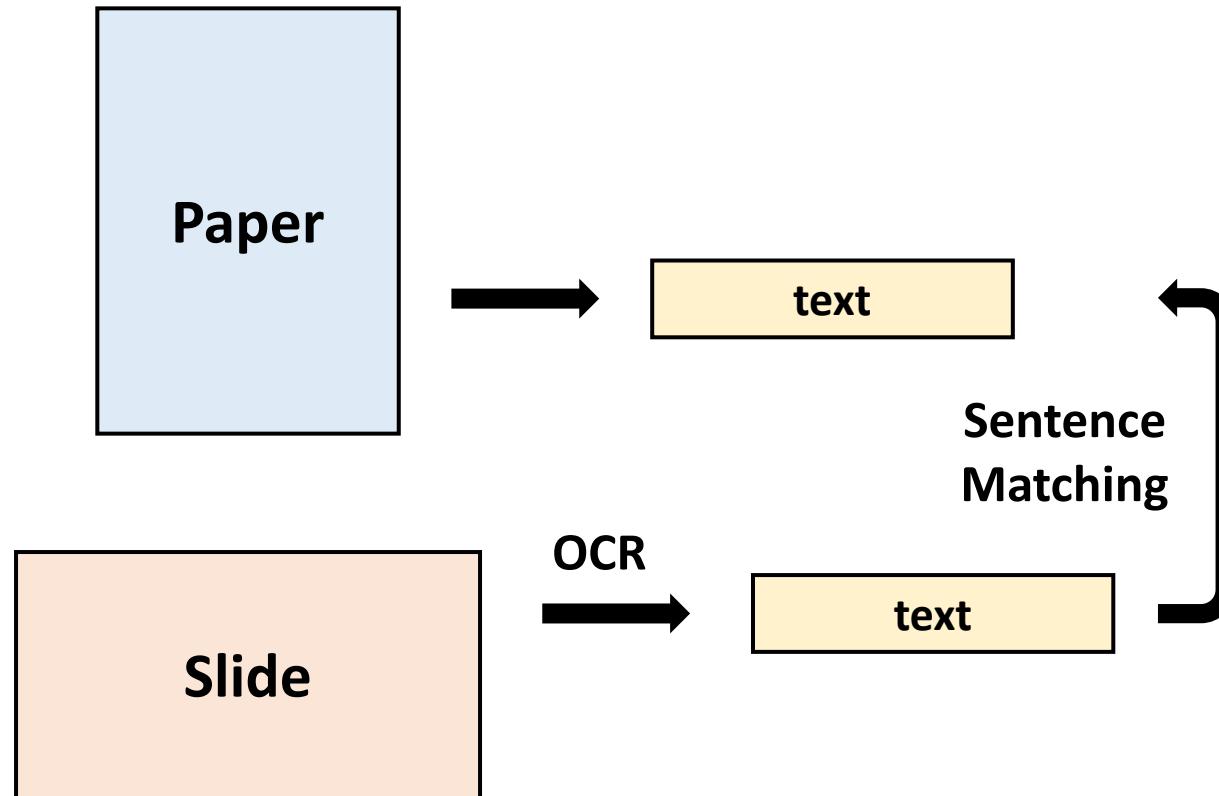


→ 4. Experiments

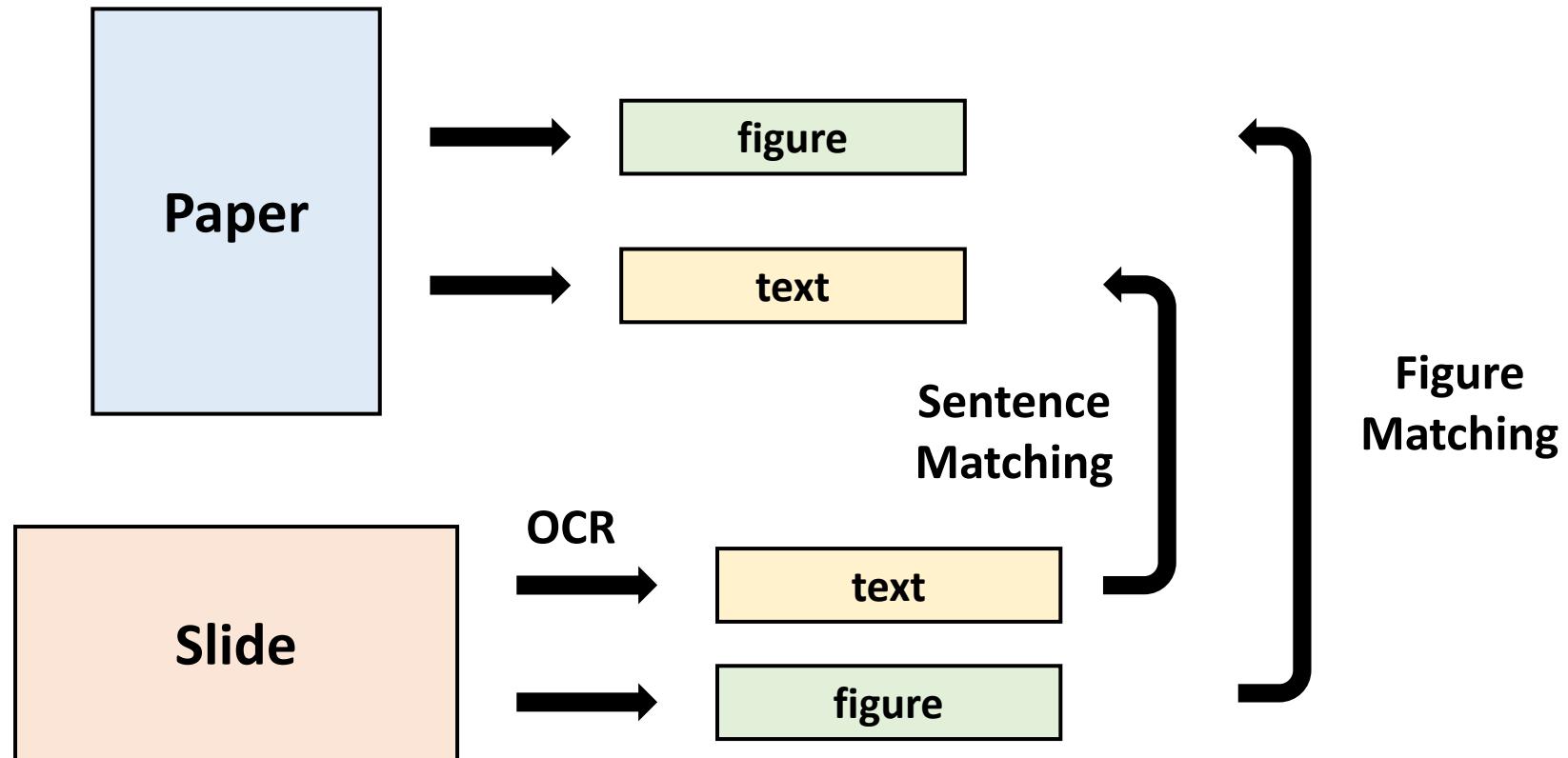
Dataset Building



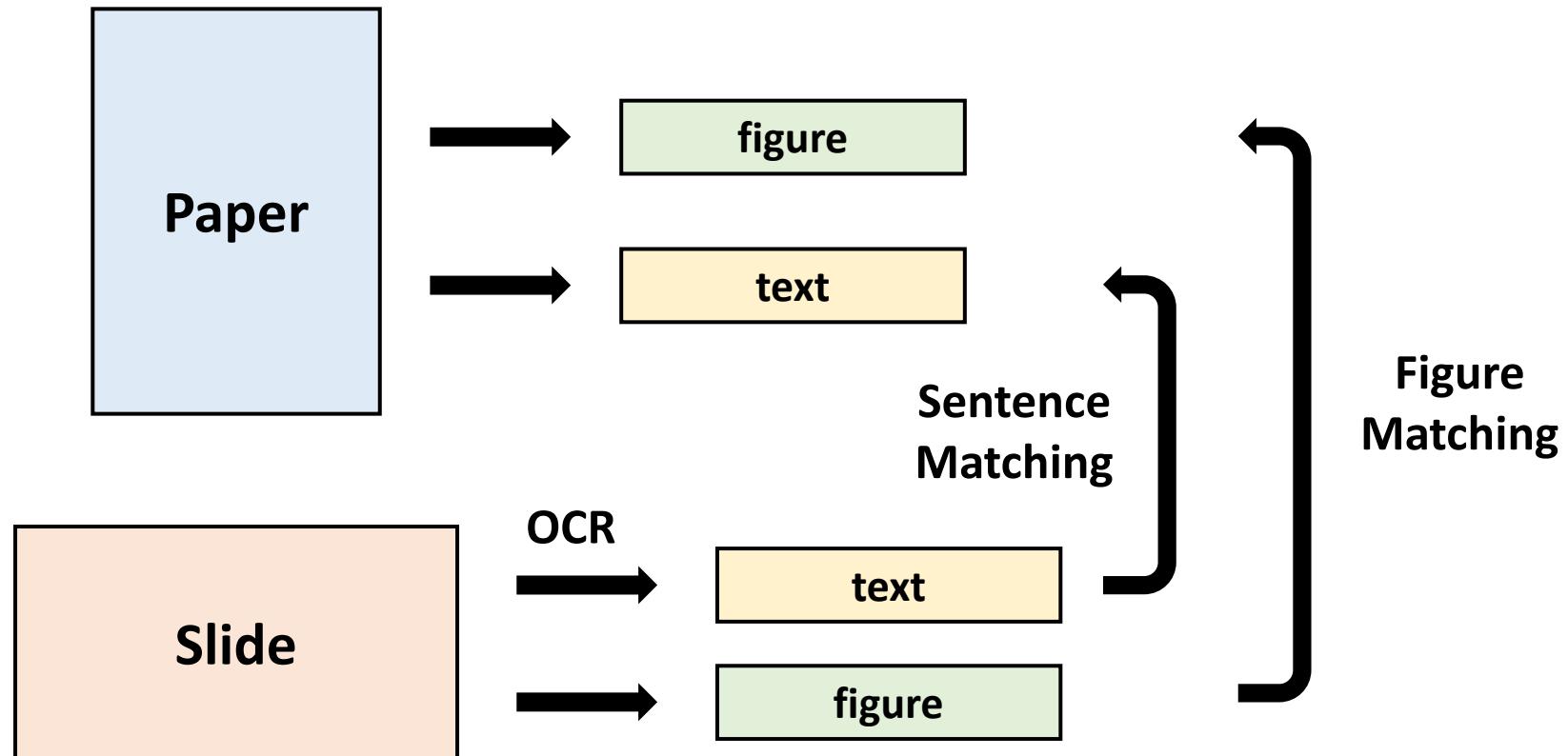
Dataset Building



Dataset Building



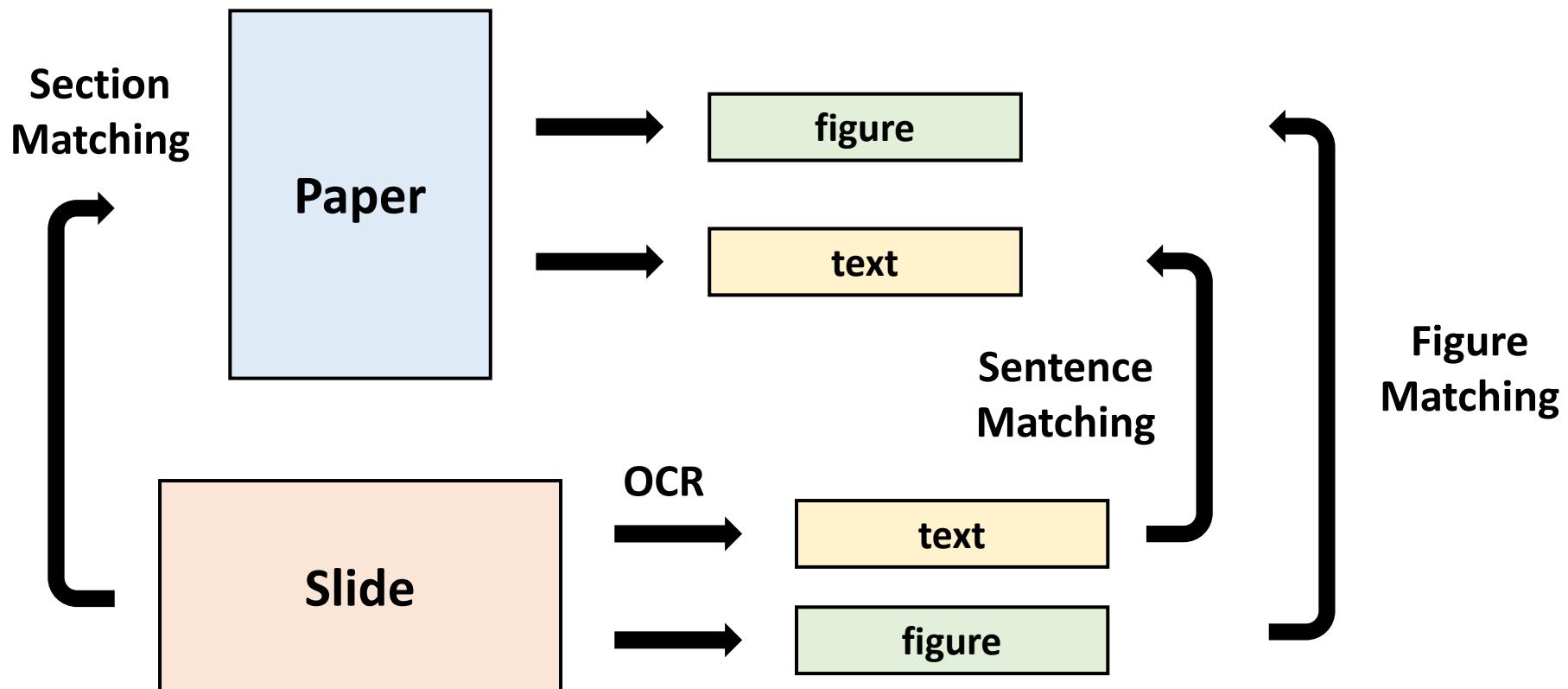
Dataset Building



**Progressive
Removing**

-
-
-

Dataset Building



**Progressive
Removing**

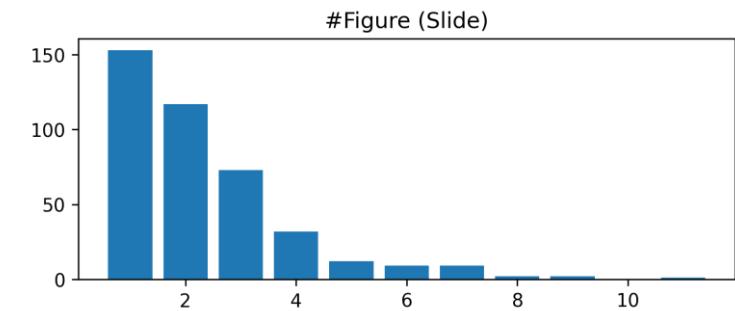
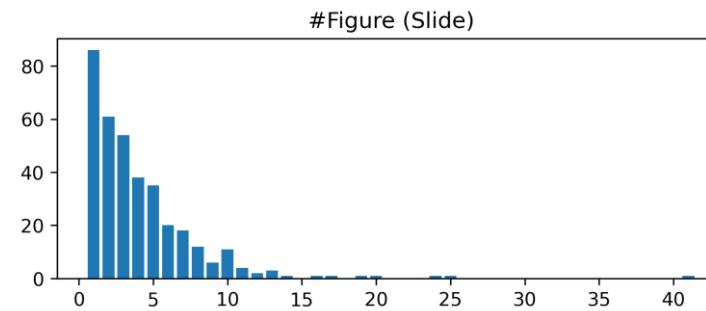
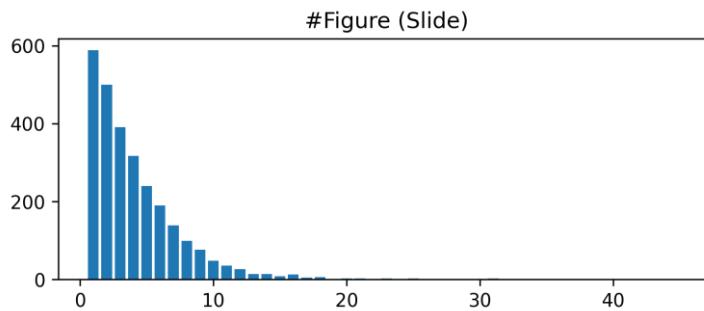
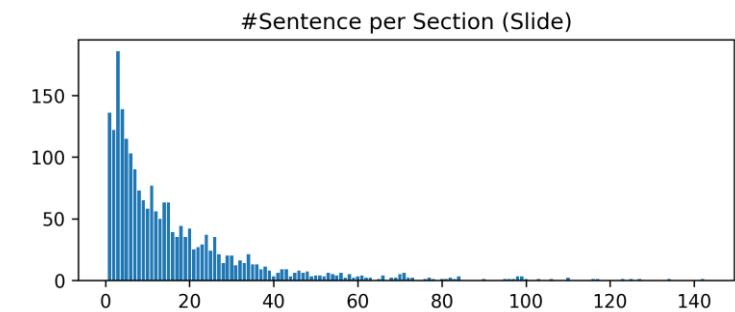
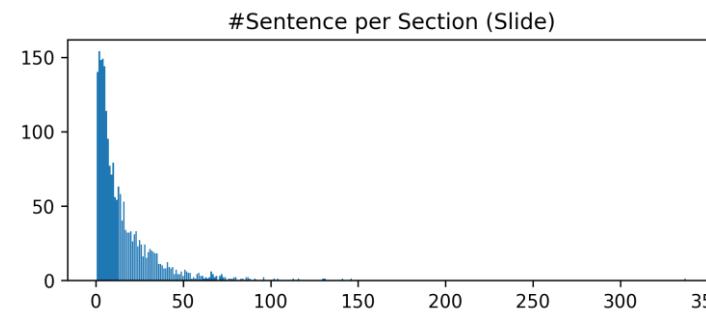
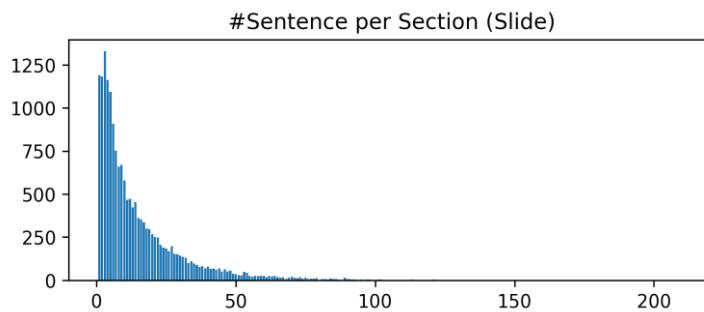
-
-
-

Dataset Building

| Paper | | | | | Slide | | |
|-----------------|-------|----------|----------------------------|---------|-------|----------------------------|---------|
| | num | #section | #sentence (per section) | #figure | #page | #sentence (per section) | #figure |
| Train | 4,686 | 6.9 | 42.9 | 8.3 | 16.9 | 8.1 | 2.4 |
| Val | 592 | 6.9 | 42.6 | 8.3 | 16.8 | 8.1 | 2.5 |
| Test | 595 | 6.9 | 42.4 | 8.4 | 16.5 | 8.1 | 2.6 |
| Test (Human) | - | | | | | | 2.3 |

Dataset Building

- Distribution of **#sentence** and **#figure** in slide
 - Similar between train, val, and test



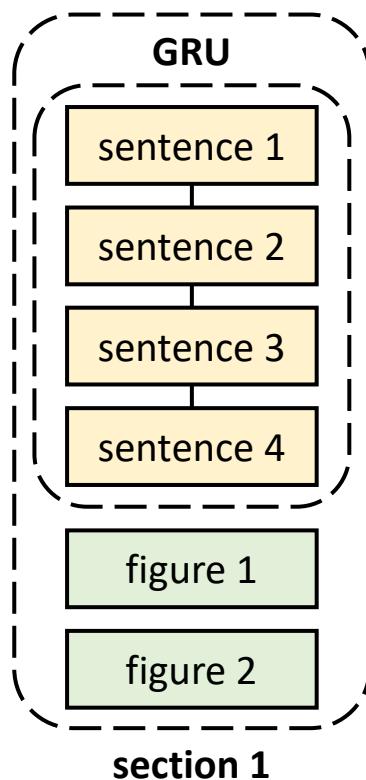
Train

Val

Test (Human)

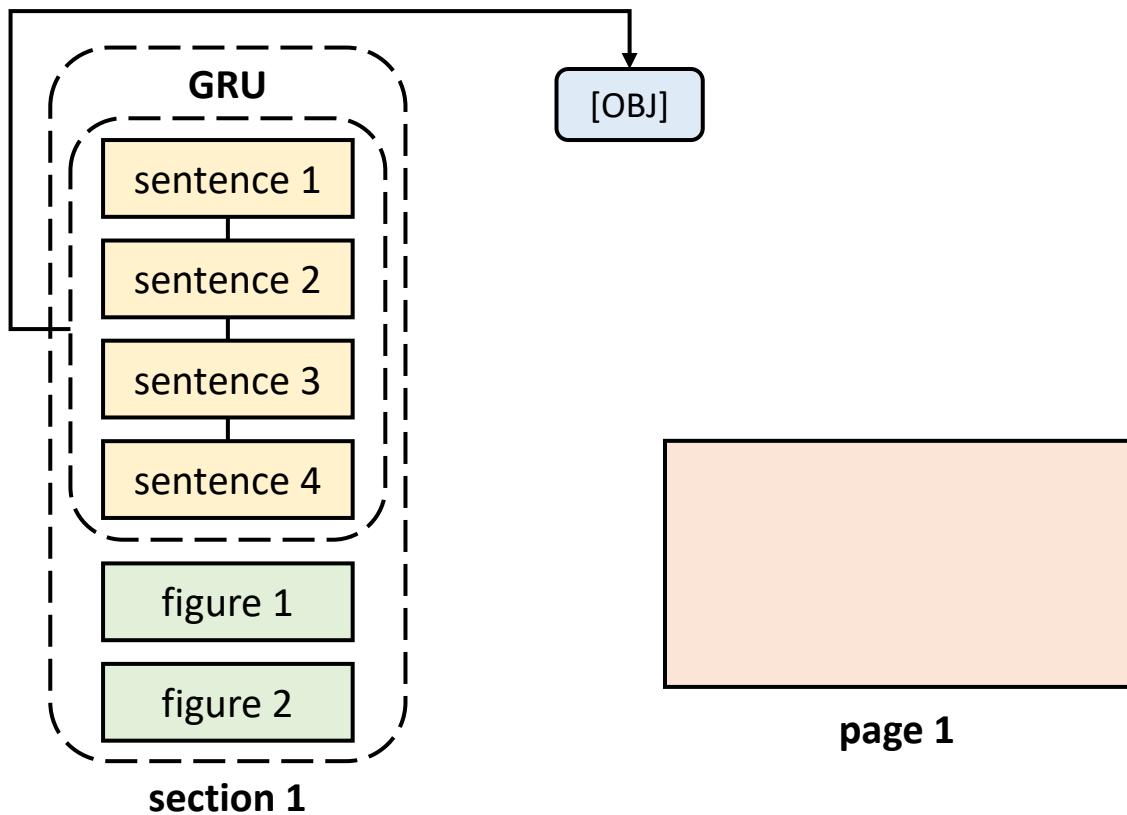
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
 - [OBJ], [PAGE], [SECTION] token
 - **Section-based** generation and **classification** for extraction



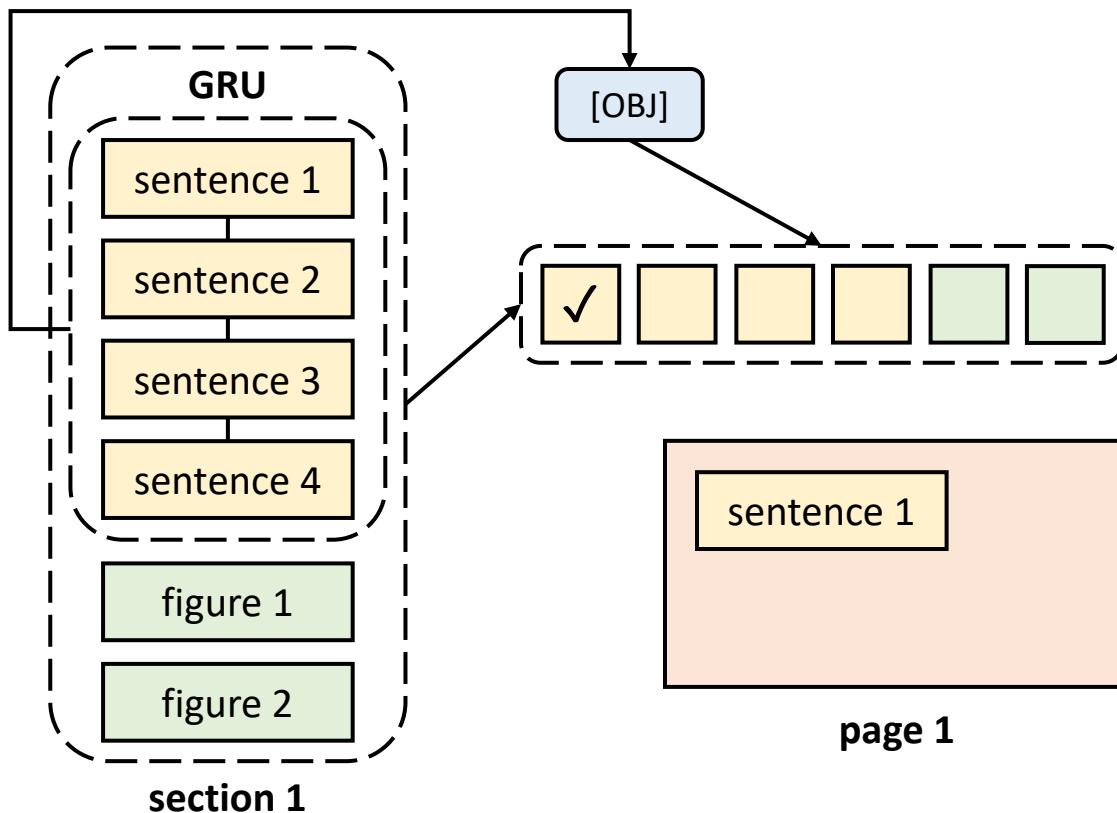
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
 - [OBJ], [PAGE], [SECTION] token
 - **Section-based** generation and **classification** for extraction



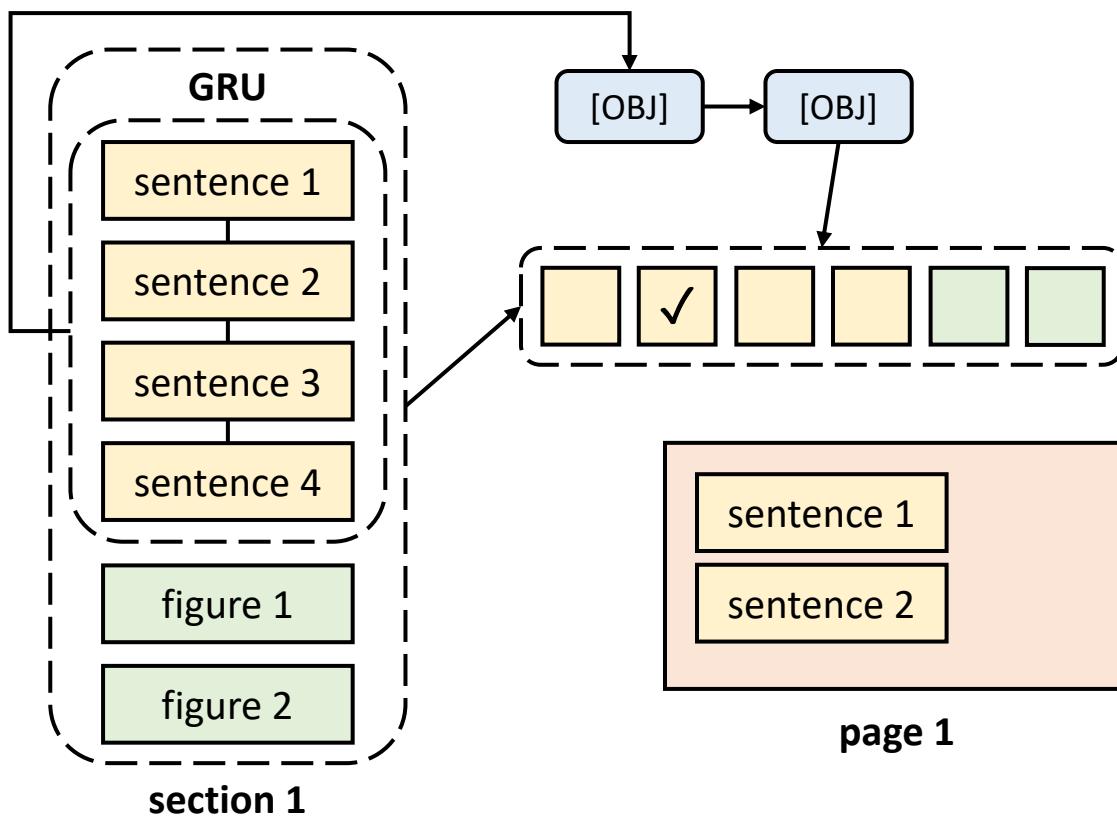
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
 - [OBJ], [PAGE], [SECTION] token
 - **Section-based generation and classification** for extraction



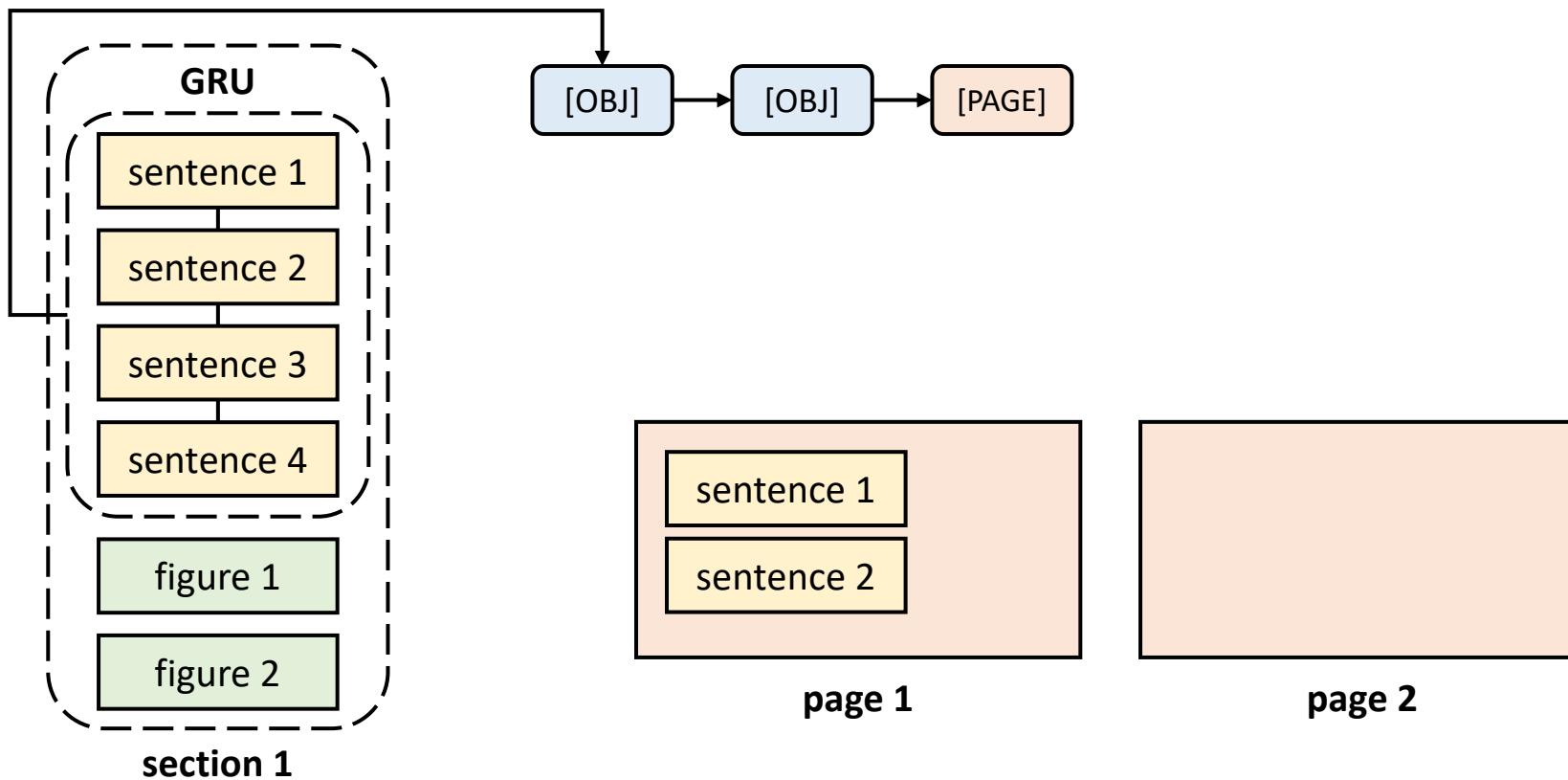
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
 - [OBJ], [PAGE], [SECTION] token
 - **Section-based generation and classification** for extraction



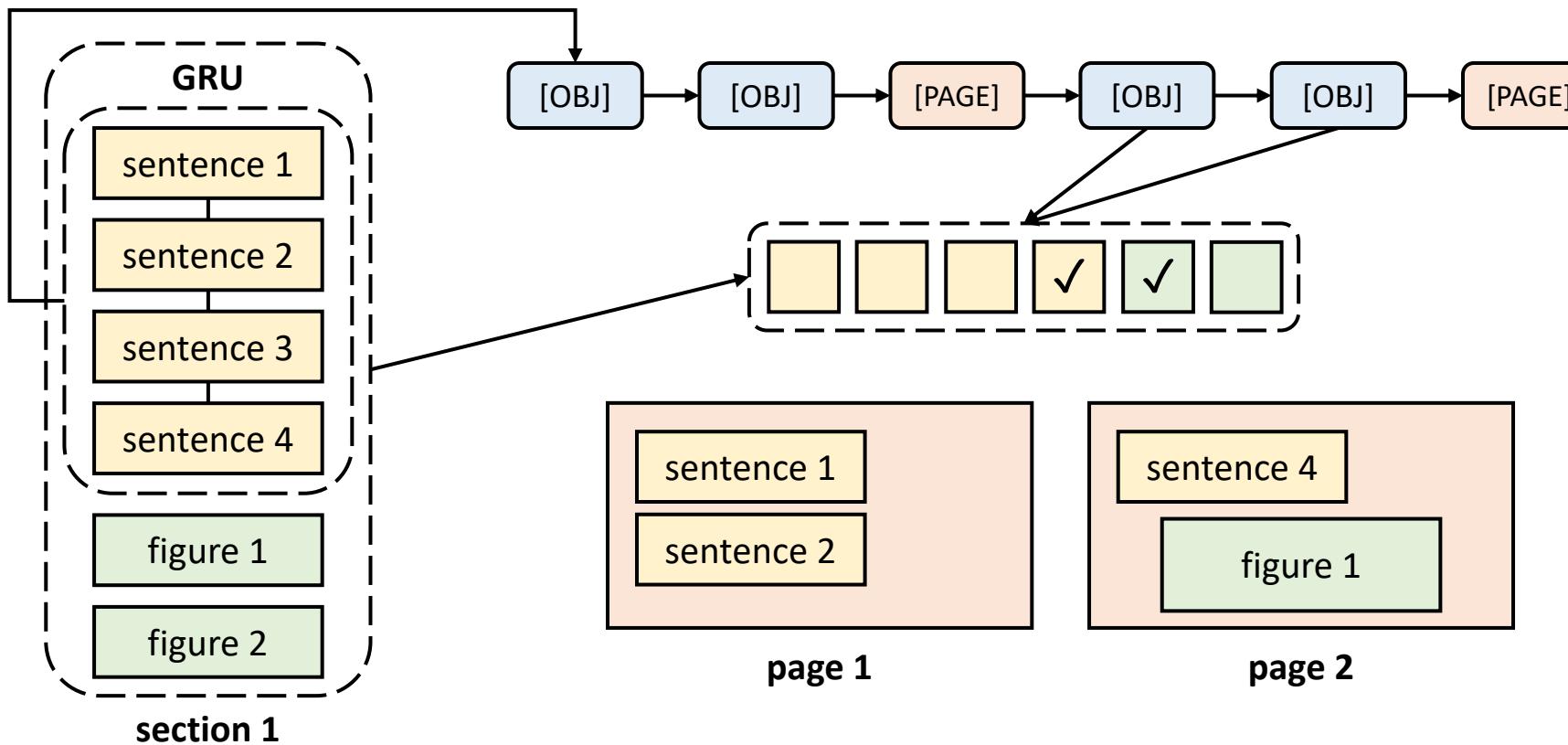
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
 - [OBJ], [PAGE], [SECTION] token
 - **Section-based generation and classification** for extraction



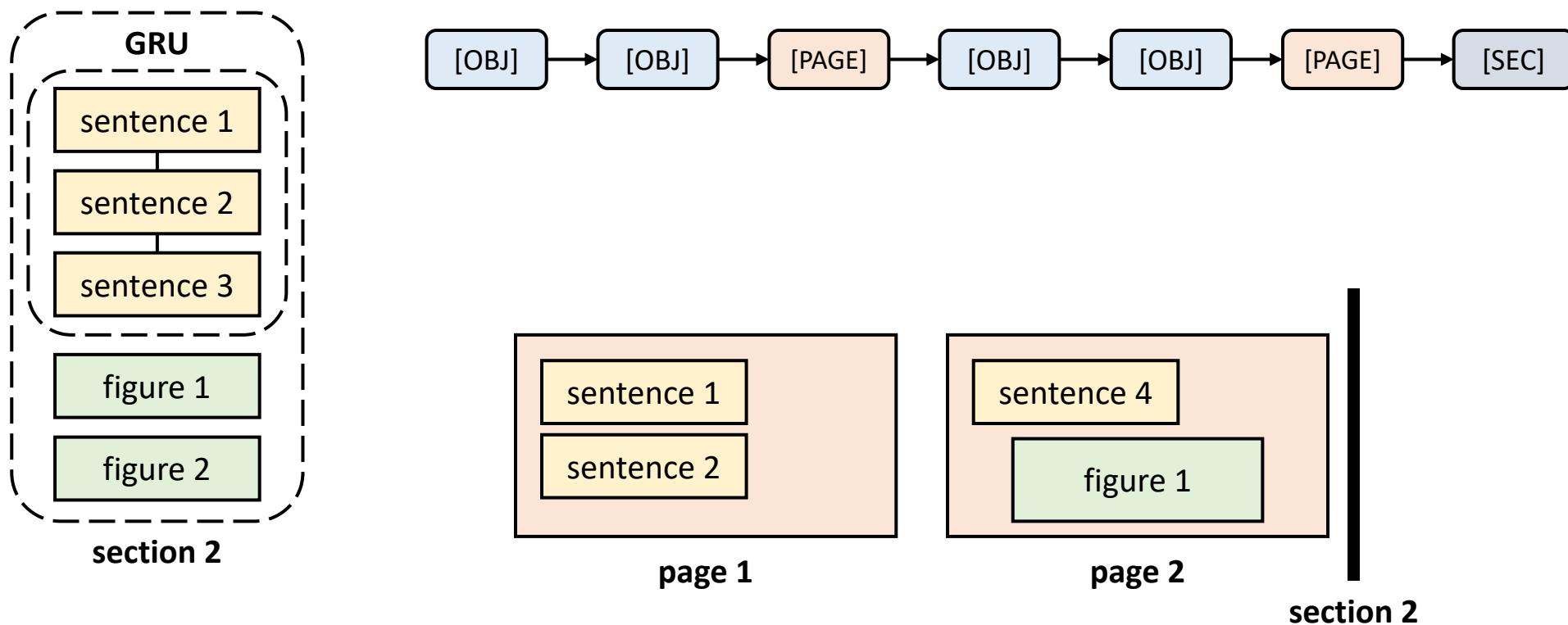
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
 - [OBJ], [PAGE], [SECTION] token
 - **Section-based generation and classification** for extraction



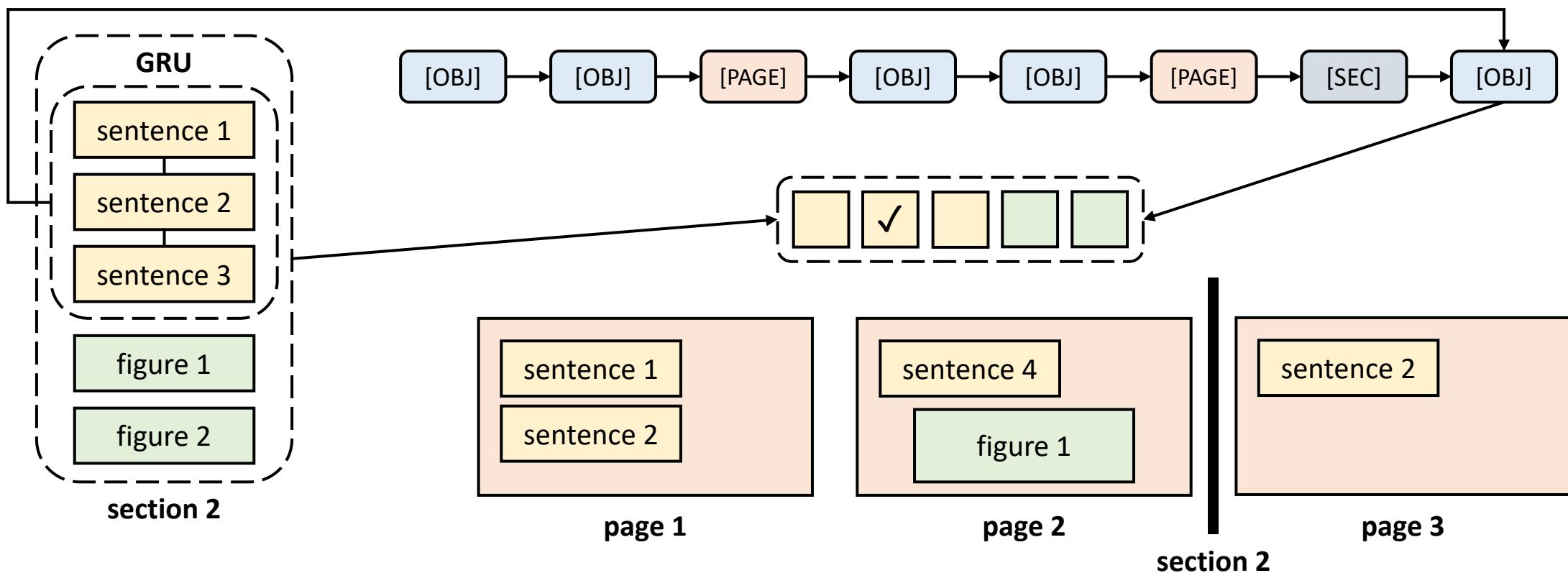
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
 - [OBJ], [PAGE], [SECTION] token
 - **Section-based generation and classification** for extraction



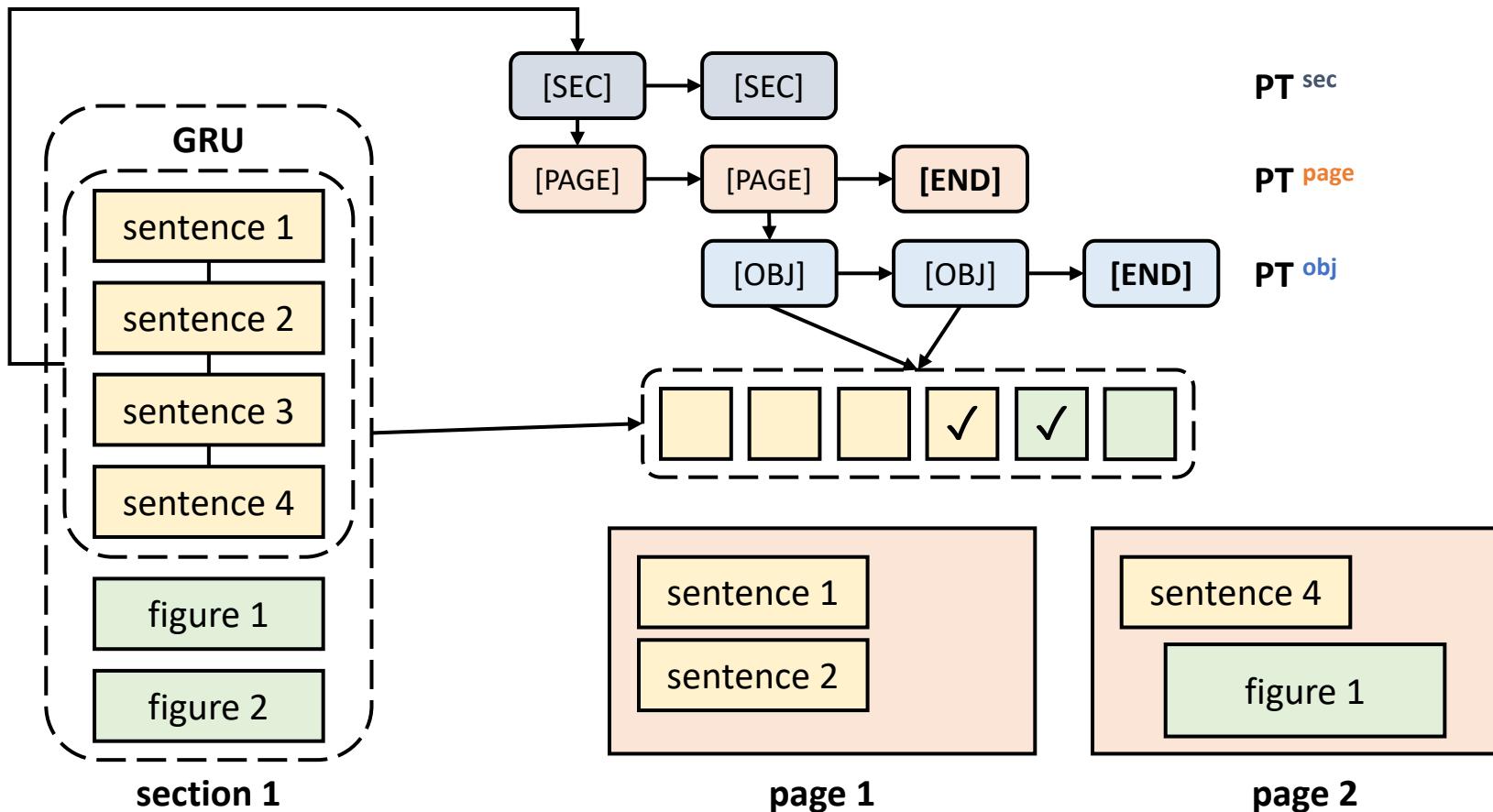
Model (Baseline)

- Recurrent extractor to build the slide step-by-step
 - [OBJ], [PAGE], [SECTION] token
 - **Section-based generation and classification** for extraction



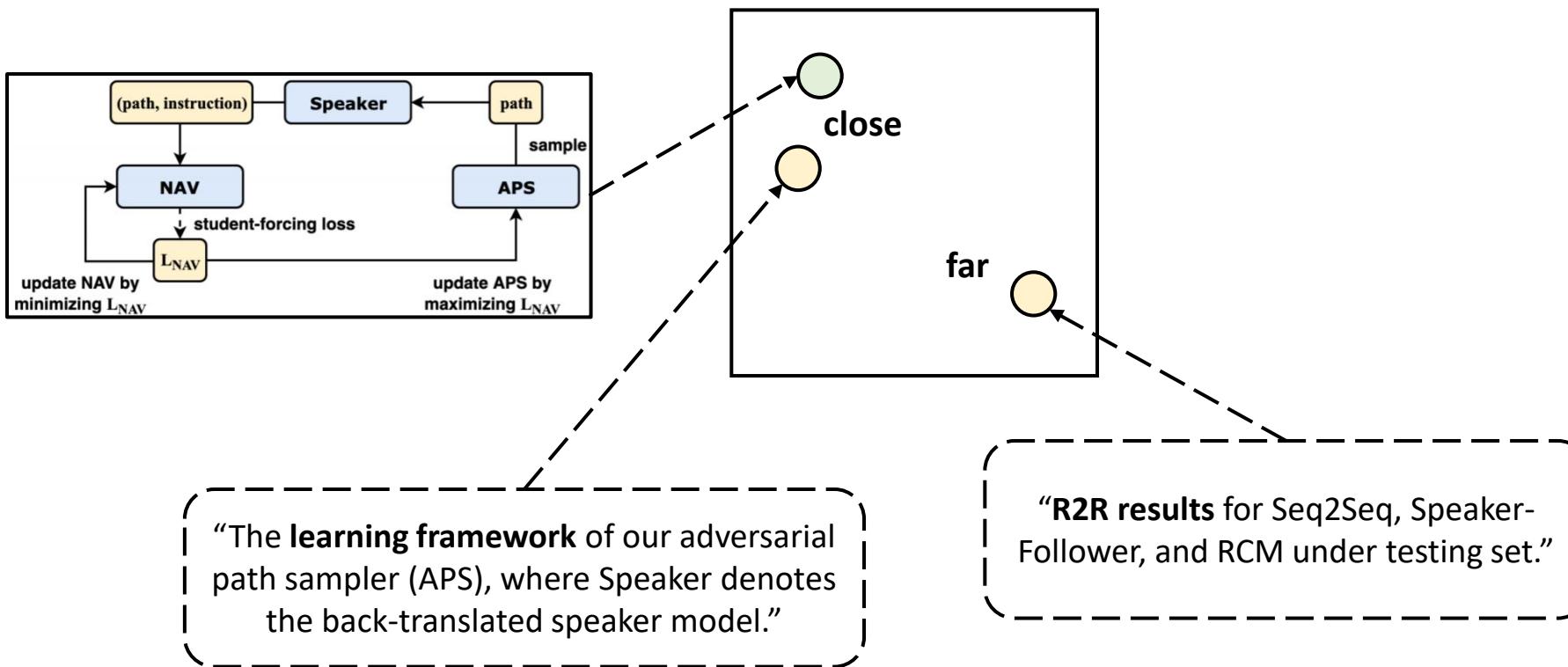
Model (HSE)

- Hierarchical Slide Extractor (HSE)
 - Different RNNs for **section**-, **page**-, and **object**- level



TextFigure Module

- Constrain the **coherence** between figure-text
 - Co-train with HSE
 - Related figure-text should be **close on embedding space**

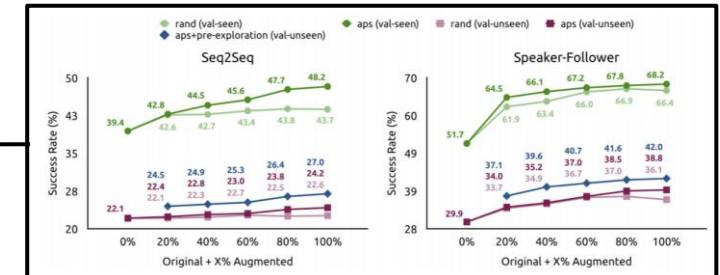
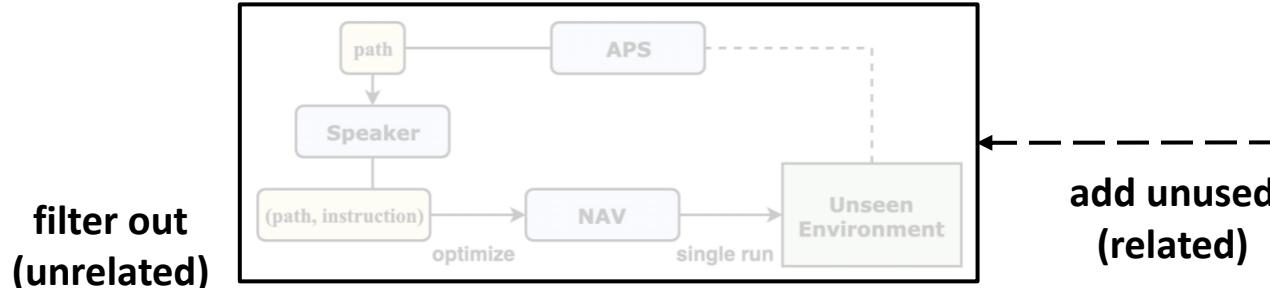


TextFigure Module

- Right figures put with right texts
 - **Filter out unrelated and add unused related figures**

Result

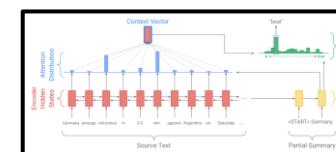
- Randomly sampled stop improving when using more than 60%
- APS sampled helps both seen and unseen
- Pre-Exploration further helps unseen environments



Paraphrasing Module

- Rewrite extracted sentences as **slide-style**
 - Seq2seq model (w/ copy attention)

“to understand the spread of individual judgements on a sentence , we compute the standard deviation of **ratings for each sentence** and then **take the mean** over all sentences .”



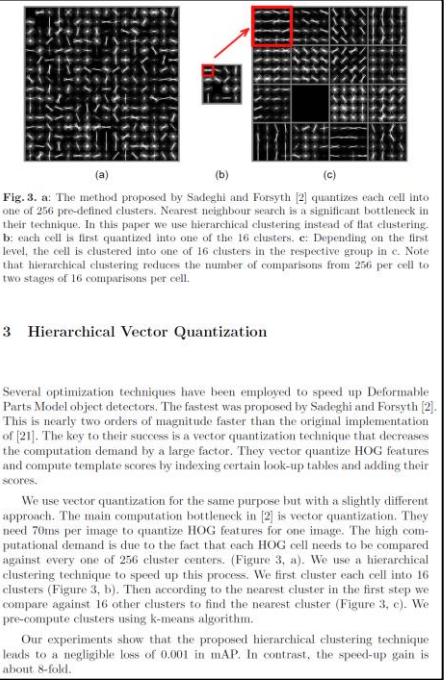
“we collect multiple ratings for a sentence and take the mean .”

“we perform **empirical evaluation** and analysis of a variety of **classification methods** for the above task .”

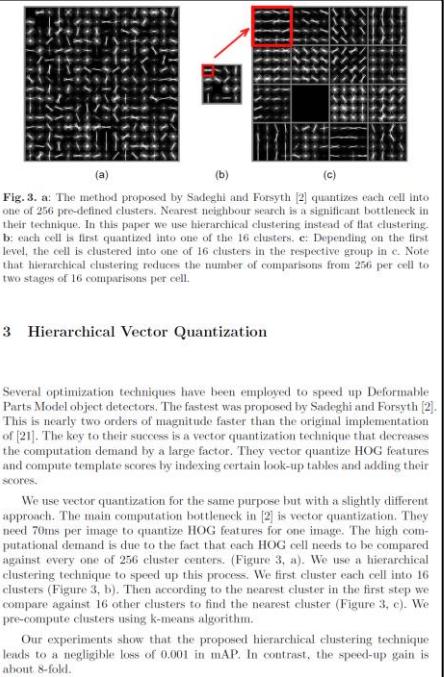
paraphrase

“**empirical evaluation** of **classification methods**”

HSE w/ TextFigure & Paraphrasing



HSE w/ TextFigure & Paraphrasing

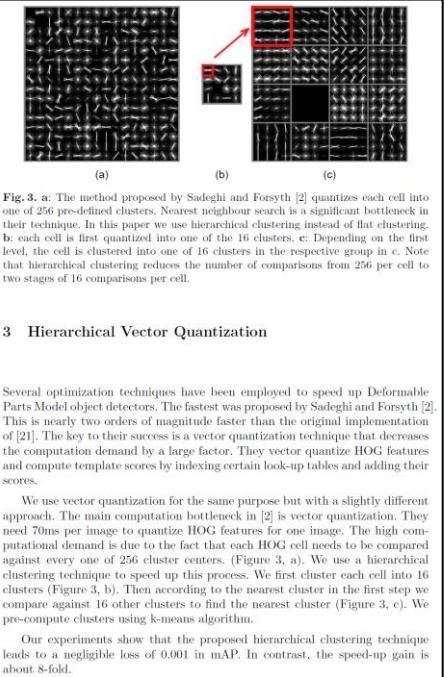


→
HSE

Hierarchical Vector Quantization

- We use a hierarchical clustering technique to speed up this process.
- We use vector quantization for the same purpose but with a slightly different approach.
- Then according to the nearest cluster in the first step we compare against 16 other clusters to find the nearest cluster (Figure 3, c).
- We pre-compute clusters using k-means algorithm.
- Our experiments show that the proposed hierarchical clustering technique leads to a negligible loss of 0.001 in mAP.

HSE w/ TextFigure & Paraphrasing



→
HSE

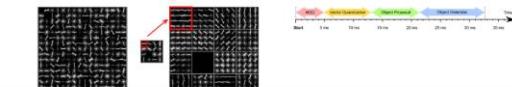
Hierarchical Vector Quantization

- We use a hierarchical clustering technique to speed up this process.
- We use vector quantization for the same purpose but with a slightly different approach.
- Then according to the nearest cluster in the first step we compare against 16 other clusters to find the nearest cluster (Figure 3, c).
- We pre-compute clusters using k-means algorithm.
- Our experiments show that the proposed hierarchical clustering technique leads to a negligible loss of 0.001 in mAP.

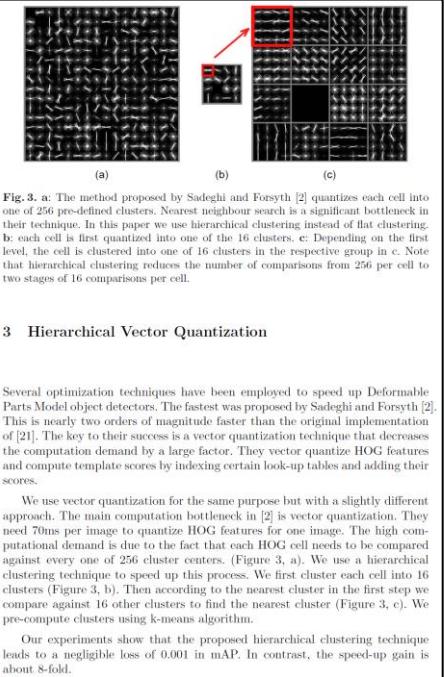
↓
**TextFigure
Module**

Hierarchical Vector Quantization

- We use a hierarchical clustering technique to speed up this process.
- We use vector quantization for the same purpose but with a slightly different approach.
- Then according to the nearest cluster in the first step we compare against 16 other clusters to find the nearest cluster (Figure 3, c).
- We pre-compute clusters using k-means algorithm.
- Our experiments show that the proposed hierarchical clustering technique leads to a negligible loss of 0.001 in mAP.



HSE w/ TextFigure & Paraphrasing



→
HSE

Hierarchical Vector Quantization

- We use a hierarchical clustering technique to speed up this process.
- We use vector quantization for the same purpose but with a slightly different approach.
- Then according to the nearest cluster in the first step we compare against 16 other clusters to find the nearest cluster (Figure 3, c).
- We pre-compute clusters using k-means algorithm.
- Our experiments show that the proposed hierarchical clustering technique leads to a negligible loss of 0.001 in mAP.

↓
**TextFigure
Module**

Hierarchical Vector Quantization

- We use a hierarchical clustering technique to speed up this process.
- We use vector quantization for the same purpose but with a slightly different approach.
- Then according to the nearest cluster in the first step we compare against 16 other clusters to find the nearest cluster (Figure 3, c).
- We pre-compute clusters using k-means algorithm.
- Our experiments show that the proposed hierarchical clustering technique leads to a negligible loss of 0.001 in mAP.

(a) (b) (c)

→
**Paraphrasing
Module**

Hierarchical Vector Quantization

- We use a hierarchical clustering technique to speed up this process .
- Vector quantization for the same feature space .
- How to find the nearest cluster in the first step ?
- Results : k-means algorithm using .
- Our method leads to a negligible loss of 0.001 in map

(a) (b) (c)

Experiments

- Evaluation metrics

Text

**the cat is sleeping
on bed**

**the brown cat is
sitting on bed**

Rouge-L: 83.3 / 71.4 / **76.9**

Experiments

- Evaluation metrics

Text

the cat is sleeping
on bed

the brown cat is
sitting on bed

Rouge-L: 83.3 / 71.4 / 76.9

$$\text{Rouge} \times e^{-\frac{|P-Q|}{Q}}$$

- consider Page Difference
- P: #Page_{pd}
- Q: #Page_{gd}

Experiments

- Evaluation metrics

Text

the cat is sleeping
on bed

the brown cat is
sitting on bed

Figure

A / D / C / F / E

A / F / B / E

LC-P/R/F: 60.0 / 75.0 / 66.7

Rouge-L: 83.3 / 71.4 / **76.9**

$$\text{Rouge} \times e^{-\frac{|P-Q|}{Q}}$$

- consider **Page** difference
- P: #Page_{pd}
- Q: #Page_{gd}

Experiments

- Evaluation metrics

Text

the cat is sleeping
on bed

the brown cat is
sitting on bed

Rouge-L: 83.3 / 71.4 / **76.9**

$$\text{Rouge} \times e^{-\frac{|P-Q|}{Q}}$$

- consider Page Difference
- P: #Page_{pd}
- Q: #Page_{gd}

Figure

A / D / C / F / E

A / F / B / E

LC-P/R/F: 60.0 / 75.0 / 66.7

TextFigure

the cat is sleeping
on bed

A

the brown cat is
sitting on bed

A

Rouge-L

the fast fox
jumped over

B

fast brown fox
jumped up

B

Experiments

1st / 2nd

| Model | Co-Train | | | w/ Module | | Text | | Figure | | TextFigure |
|----------|------------|------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|
| | TextFigure | Paraphrase | TextFigure | Rouge-L | w/ Page | LC-P | LC-R | LC-F1 | Rouge-L | |
| Baseline | X | X | X | 27.2 | 21.8 | 13.2 | 21.9 | 16.5 | 3.6 | |
| HSE | X | X | X | 27.7 | 22.9 | 14.6 | 23.7 | 18.1 | 4.3 | |
| | X | ✓ | X | 32.3 | 26.7 | 14.6 | 23.7 | 18.1 | 4.7 | |
| | ✓ | X | X | 28.7 | 24.0 | 14.8 | 32.4 | 20.3 | 7.9 | |
| | ✓ | X | ✓ | 28.7 | 24.0 | 24.6 | 40.5 | 30.6 | 13.8 | |
| | ✓ | ✓ | X | 33.6 | 28.2 | 14.8 | 32.4 | 20.3 | 8.2 | |
| | ✓ | ✓ | ✓ | 33.6 | 28.2 | 24.6 | 40.5 | 30.6 | 15.5 | |

Experiments

| Model | Co-Train | w/ Module | Text | | Figure | | TextFigure | | |
|-----------------|------------|------------|------------|-------------|-------------|-------------|-------------|-------------|------------|
| | TextFigure | Paraphrase | TextFigure | Rouge-L | w/ Page | LC-P | LC-R | LC-F1 | Rouge-L |
| Baseline | X | X | X | 27.2 | 21.8 | 13.2 | 21.9 | 16.5 | 3.6 |
| | X | X | X | 27.7 | 22.9 | 14.6 | 23.7 | 18.1 | 4.3 |
| | X | ✓ | X | 32.3 | 26.7 | 14.6 | 23.7 | 18.1 | 4.7 |
| | ✓ | X | X | 28.7 | 24.0 | 14.8 | 32.4 | 20.3 | 7.9 |
| | ✓ | X | ✓ | 28.7 | 24.0 | 24.6 | 40.5 | 30.6 | 13.8 |
| | ✓ | ✓ | X | 33.6 | 28.2 | 14.8 | 32.4 | 20.3 | 8.2 |
| | ✓ | ✓ | ✓ | 33.6 | 28.2 | 24.6 | 40.5 | 30.6 | 15.5 |

- **Hierarchical architecture** extracts slide
 - Helps both **text quality** and **figure retrieval**

Experiments

| Model | Co-Train w/ Module | | | Text | | Figure | | | TextFigure |
|----------|--------------------|------------|------------|-------------|-------------|--------|------|-------|------------|
| | TextFigure | Paraphrase | TextFigure | Rouge-L | w/ Page | LC-P | LC-R | LC-F1 | Rouge-L |
| Baseline | X | X | X | 27.2 | 21.8 | 13.2 | 21.9 | 16.5 | 3.6 |
| | X | X | X | 27.7 | 22.9 | 14.6 | 23.7 | 18.1 | 4.3 |
| | X | ✓ | X | 32.3 | 26.7 | 14.6 | 23.7 | 18.1 | 4.7 |
| HSE | ✓ | X | X | 28.7 | 24.0 | 14.8 | 32.4 | 20.3 | 7.9 |
| | ✓ | X | ✓ | 28.7 | 24.0 | 24.6 | 40.5 | 30.6 | 13.8 |
| | ✓ | ✓ | X | 33.6 | 28.2 | 14.8 | 32.4 | 20.3 | 8.2 |
| | ✓ | ✓ | ✓ | 33.6 | 28.2 | 24.6 | 40.5 | 30.6 | 15.5 |
| | | | | | | | | | |

- **Paraphrasing module** rewrites sentences into slide-style
 - Better **text** as a slide

Experiments

| Model | Co-Train | | | w/ Module | | Text | | Figure | | TextFigure |
|----------|------------|------------|------------|-------------|-------------|-------------|-------------|-------------|------------|------------|
| | TextFigure | Paraphrase | TextFigure | Rouge-L | w/ Page | LC-P | LC-R | LC-F1 | Rouge-L | |
| Baseline | X | X | X | 27.2 | 21.8 | 13.2 | 21.9 | 16.5 | 3.6 | |
| HSE | X | X | X | 27.7 | 22.9 | 14.6 | 23.7 | 18.1 | 4.3 | |
| | X | ✓ | X | 32.3 | 26.7 | 14.6 | 23.7 | 18.1 | 4.7 | |
| | ✓ | X | X | 28.7 | 24.0 | 14.8 | 32.4 | 20.3 | 7.9 | |
| | ✓ | X | ✓ | 28.7 | 24.0 | 24.6 | 40.5 | 30.6 | 13.8 | |
| | ✓ | ✓ | X | 33.6 | 28.2 | 14.8 | 32.4 | 20.3 | 8.2 | |
| | ✓ | ✓ | ✓ | 33.6 | 28.2 | 24.6 | 40.5 | 30.6 | 15.5 | |

- Co-train with **TextFigure** constrain
 - Learns the **correlation** between text and figure

Experiments

| Model | Co-Train w/ Module | | | Text | | Figure | | | TextFigure |
|----------|--------------------|------------|------------|---------|---------|-------------|-------------|-------------|-------------|
| | TextFigure | Paraphrase | TextFigure | Rouge-L | w/ Page | LC-P | LC-R | LC-F1 | Rouge-L |
| Baseline | X | X | X | 27.2 | 21.8 | 13.2 | 21.9 | 16.5 | 3.6 |
| HSE | X | X | X | 27.7 | 22.9 | 14.6 | 23.7 | 18.1 | 4.3 |
| | X | ✓ | X | 32.3 | 26.7 | 14.6 | 23.7 | 18.1 | 4.7 |
| | ✓ | X | X | 28.7 | 24.0 | 14.8 | 32.4 | 20.3 | 7.9 |
| | ✓ | X | ✓ | 28.7 | 24.0 | 24.6 | 40.5 | 30.6 | 13.8 |
| | ✓ | ✓ | X | 33.6 | 28.2 | 14.8 | 32.4 | 20.3 | 8.2 |
| | ✓ | ✓ | ✓ | 33.6 | 28.2 | 24.6 | 40.5 | 30.6 | 15.5 |

- **TextFigure module** removes unrelated or adds related
 - Benefits **figure retrieval** a lot

Experiments

| Model | Co-Train w/ Module | | | Text | | Figure | | | TextFigure |
|----------|--------------------|------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | TextFigure | Paraphrase | TextFigure | Rouge-L | w/ Page | LC-P | LC-R | LC-F1 | Rouge-L |
| Baseline | X | X | X | 27.2 | 21.8 | 13.2 | 21.9 | 16.5 | 3.6 |
| HSE | X | X | X | 27.7 | 22.9 | 14.6 | 23.7 | 18.1 | 4.3 |
| | X | ✓ | X | 32.3 | 26.7 | 14.6 | 23.7 | 18.1 | 4.7 |
| | ✓ | X | X | 28.7 | 24.0 | 14.8 | 32.4 | 20.3 | 7.9 |
| | ✓ | X | ✓ | 28.7 | 24.0 | 24.6 | 40.5 | 30.6 | 13.8 |
| | ✓ | ✓ | X | 33.6 | 28.2 | 14.8 | 32.4 | 20.3 | 8.2 |
| | ✓ | ✓ | ✓ | 33.6 | 28.2 | 24.6 | 40.5 | 30.6 | 15.5 |

- Combines both **Paraphrasing** and **TextFigure** module
 - Fully improves on **all aspects of metrics**

Qualitative Examples

Introduction

- We propose a novel multi-label conditional alignment methodology to bridge domain divergence while preserving the discriminability of the features .
- Mcar : multi-label conditional distribution alignment and detection regularization model
- Minimize the cross-domain feature distribution gaps .
- A whole image can have complex multimodal structures .
- Global (image-level) feature alignment (image-level)

Category Prediction based Regularization

- What is the structure of the graph ?
- We propose a novel category prediction mechanism for object detection .
- Each proposal will be classified as a regressor r .
- Region proposal network (rpn) 23-28 august zA,zA,
- Loss function : $= + (x)$

Adaptation from Clear to Foggy Scenes.

- Cross-domain detection from real to virtual image scenarios
- Domain adaption from normal / clear images to foggy image
- Pascal voc , pascal voc to comic

Ablation Study

- Qualitative results : quantitative results 23-28 august zA,zA,
- Adacose : adaptive feature visualization with mutual regularization--p. zhao et al .
- We use the foggy cityscapes dataset as the target domain .
- Train on labeled data in the target domain
- Multiple auxiliary loss terms in the proposed learning objective

| Method | IM07 PASCAL3D+ mIoU | our work mIoU | train multiview (zA,zA) | | | | | | |
|-----------------|---------------------|---------------|-------------------------|------|------|------|------|------|------|
| Neural-only [1] | 25.1 | 32.7 | 31.0 | 32.5 | 31.3 | 20.4 | | | |
| HRD-Feature [3] | 26.1 | 37.2 | 31.2 | 31.2 | 32.6 | 26.9 | 37.5 | | |
| DA-Model [2] | 30.5 | 38.0 | 36.3 | 36.3 | 39.0 | 35.0 | 38.5 | | |
| SC-DA [4] | 33.5 | 38.0 | 44.5 | 36.3 | 39.0 | 35.0 | 38.5 | | |
| MAP [7] | 39.7 | 38.3 | 43.3 | 39.0 | 40.8 | 39.7 | 41.0 | | |
| DA-Net [10] | 36.2 | 40.0 | 39.1 | 39.1 | 40.2 | 37.8 | 41.2 | | |
| DD-MIL [18] | 30.8 | 40.7 | 44.1 | 37.7 | 39.4 | 34.5 | 39.4 | 32.2 | 34.6 |
| MT-DA [19] | 34.8 | 44.2 | 44.0 | 39.2 | 39.2 | 36.5 | 39.2 | 34.8 | 36.5 |
| Deep-DA [16] | 33.2 | 44.2 | 44.0 | 39.2 | 39.2 | 36.5 | 39.2 | 36.5 | 36.5 |
| SCL [24] | 31.6 | 41.0 | 41.4 | 36.1 | 36.8 | 34.7 | 37.0 | 36.2 | 37.0 |
| MCAR (ours) | 32.7 | 42.3 | 43.3 | 41.3 | 44.3 | 43.4 | 37.4 | 36.6 | 36.6 |
| Train-on Target | 30.0 | 36.2 | 37.8 | 31.7 | 32.0 | 33.9 | 37.4 | 35.0 | 35.0 |

Introduction

- What is a good emotion classification task ?
- We use the context principle for emotion recognition .
- Context 1 : incorporating cues from different modalities
- Multimodal emotion recognition (cvpr 2020)
- Not asking for the meaning of a word in isolation and instead of finding the meaning in isolation .

Network Architecture

- How to train your neural network ?
- To train the soft margin loss function :
- We combine the two loss functions , Imultiplicative (from eq . 1) .
- \rightarrow 2classification

Datasets

- We present a comparison with other datasets .
- The apparent emotional states of the people
- How do we evaluate the annotation process ?
- How do we evaluate the friendliness ?

| Data type | Dataset | Dataset Size | Agents Annotated | Setting | Emotion Labels | Context | Emoticon |
|-----------|---------------|-------------------------|------------------|--------------|----------------|---------|----------|
| Images | EMOTIC [1] | 183.6 images | 34,320 | Web | 2 Categories | Yes | 48.32 |
| | CASME-S [2] | 120 images | 64,000 | Web | 3 Categories | No | |
| | CASME [3] | 70,000 images | 70,000 | TV Shows | 3 Categories | Yes | |
| Videos | AFW [4] | 1,809 clips | 1,809 | TV Show | 4 Categories | No | 16.46 |
| | EMOCAF [5] | 1,120 clips | 1,120 | TV Show | 4 Categories | No | 16.45 |
| | GroupWalk [6] | 45 clips (10 mins each) | 3544 | Real Setting | 4 Categories | Yes | 72.12 |

Analysis and Discussion

- Emotic dataset . emotic dataset was collected for
- Two-stream network (two-stream) [2]
- Gcn (ours) depth-based (ours)
- Groupwalk dataset was difficult to test on groupwalk .

| Labels | Kou et al. [1] | Zhang et al. [2] | Lee et al. [3] | Emoticon |
|--------------|----------------|------------------|----------------|----------|
| Affection | 29.85 | 46.39 | 79.0 | 68.65 |
| Anger | 09.49 | 10.87 | 11.5 | 14.92 |
| Annoyance | 14.06 | 11.25 | 16.4 | 18.45 |
| Anticipation | 38.04 | 62.04 | 33.05 | 68.17 |
| Confidence | 07.48 | 8.91 | 7.4 | 12.45 |
| Disapproval | 17.0 | 17.0 | 16.0 | 21.21 |
| Fear | 21.37 | 29.91 | 22.30 | 34.12 |
| Happiness | 16.89 | 16.94 | 17.19 | 16.41 |
| Neutral | 70.0 | 80.0 | 78.0 | 80.75 |
| Pain | 03.18 | 1.94 | 15.68 | 11.25 |
| Engagement | 87.53 | 88.56 | 96.58 | 90.45 |
| Interest | 17.73 | 13.03 | 19.26 | 22.37 |
| Surprise | 27.16 | 11.53 | 23.7 | 38.4 |
| Fatigue | 09.70 | 13.26 | 13.04 | 19.15 |
| | | | | 16.23 |

SELF-ADVERSARIAL LEARNING

- For a training set with n real samples , we have
- Sal (ours) (a) sal
- How to suffer from the reward sparsity ?
- Sal (ours) 3 (ours)

TRAINING

- The comparative discriminator can offer more informative learning signals from the comparative discriminator .
- How to enhance the generalization ability of the comparative discriminator ?
- $E (p_z (z) , m)$

COMPARATIVE DISCRIMINATOR

- The self-improvement mechanism corresponds to the comparative discriminator .
- How to construct the model to supervise the model ?
- (goodfellow et al . , 2014)

RESULTS IN REAL DATA

- Table 3 . the results of coco image caption .

| Metric | MLE | SeqGAN | MilGAN | RecurrentGAN | SAL |
|---------|---------------|---------------|---------------|---------------|----------------------|
| BLEU-2 | 0.494 ± 0.010 | 0.514 ± 0.010 | 0.491 ± 0.010 | 0.503 ± 0.010 | 0.521 ± 0.010 |
| BLEU-3 | 0.157 ± 0.005 | 0.175 ± 0.005 | 0.155 ± 0.005 | 0.167 ± 0.005 | 0.187 ± 0.005 |
| BLEU-4 | 0.057 ± 0.001 | 0.071 ± 0.001 | 0.055 ± 0.001 | 0.067 ± 0.001 | 0.077 ± 0.001 |
| CIDEr | 0.153 ± 0.005 | 0.171 ± 0.005 | 0.151 ± 0.005 | 0.163 ± 0.005 | 0.173 ± 0.005 |
| SPICE | 0.759 ± 0.002 | 0.694 ± 0.002 | 0.703 ± 0.002 | 0.715 ± 0.002 | 0.754 ± 0.002 |
| ROUGE-L | 0.422 ± 0.002 | 0.442 ± 0.002 | 0.431 ± 0.002 | 0.451 ± 0.002 | 0.453 ± 0.002 |
| BLEU-1 | 0.723 ± 0.02 | 1.003 ± 0.02 | 1.072 ± 0.02 | 1.147 ± 0.02 | 1.147 ± 0.02 |
| ROUGE-L | 0.640 ± 0.008 | 0.695 ± 0.018 | 0.676 ± 0.018 | 0.707 ± 0.018 | 0.717 ± 0.018 |

Qualitative Examples

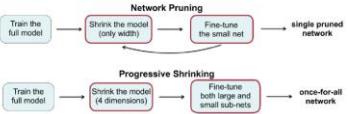
- TextFigure Module (w/o vs w/)

INTRODUCTION

- Decouple the model training stage and search stage
- Specialized sub-nets + sub-nets + sub-nets
- We extensively evaluated the effectiveness of ofa on imagenet
- How to deploy different hardware efficiency constraints ?
- We propose a progressive shrinking algorithm for once-for-all .

INTRODUCTION

- Decouple the model training stage and search stage
- Specialized sub-nets + sub-nets + sub-nets
- We extensively evaluated the effectiveness of ofa on imagenet
- How to deploy different hardware efficiency constraints ?
- We propose a progressive shrinking algorithm for once-for-all .

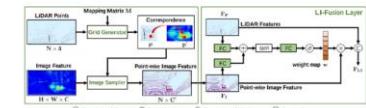


Method

- Enet : enhancing point features with image semantics
- Image feature in a point-wise manner .
- Li-fusion rpn effectively combines the lidar point feature
- We combine the point features si with the aid of our li-fusion module .

Method

- Enet : enhancing point features with image semantics
- Image feature in a point-wise manner .
- Li-fusion rpn effectively combines the lidar point feature
- We combine the point features si with the aid of our li-fusion module .



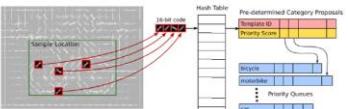
Priority Lists

- Each cell is linked to a list of templates
- Each root has a limited budget of locations
- How to balance proposals among all locations ?
- We propose a score adjustment process .

| Method | Ours | Ours | Ours | FTVQ [2] | DPM V5 [21] |
|-------------|--------|--------|--------|----------|-------------|
| Frequency | 100Hz | 30Hz | 15Hz | 2Hz | 0.07Hz |
| aeroplane | 0.1689 | 0.2695 | 0.3029 | 0.3320 | 0.3318 |
| bicycle | 0.3563 | 0.5735 | 0.5946 | 0.5933 | 0.5878 |
| boat | 0.0363 | 0.0909 | 0.0909 | 0.0927 | 0.0909 |
| boat | 0.0363 | 0.0363 | 0.1141 | 0.1568 | 0.1861 |
| bottle | 0.0969 | 0.1938 | 0.2126 | 0.2094 | 0.2535 |
| bus | 0.2989 | 0.4130 | 0.4720 | 0.5129 | 0.5056 |
| car | 0.2505 | 0.4240 | 0.4996 | 0.5373 | 0.5271 |
| cat | 0.1729 | 0.1729 | 0.1931 | 0.2253 | 0.1904 |
| chair | 0.0969 | 0.0969 | 0.1934 | 0.2890 | 0.3460 |
| cow | 0.0969 | 0.1062 | 0.1994 | 0.2432 | 0.2444 |
| diningtable | 0.1743 | 0.2500 | 0.2510 | 0.2685 | 0.2750 |
| dog | 0.0507 | 0.1159 | 0.1159 | 0.1260 | 0.1238 |
| horse | 0.2724 | 0.4735 | 0.5539 | 0.5651 | 0.5709 |

Priority Lists

- Each cell is linked to a list of templates
- Each root has a limited budget of locations
- How to balance proposals among all locations ?
- We propose a score adjustment process .



Evaluation on Various Target Tasks

- Meta-networks trained by our meta-training scheme (meta-networks)
- The first source model is trained on tinyimagenet .
- We use 34 - layer resnet as a source and target model , respectively .
- Ours needs only 50 samples per class

Evaluation on Various Target Tasks

- Meta-networks trained by our meta-training scheme (meta-networks)
- The first source model is trained on tinyimagenet .
- We use 34 - layer resnet as a source and target model , respectively .
- Ours needs only 50 samples per class

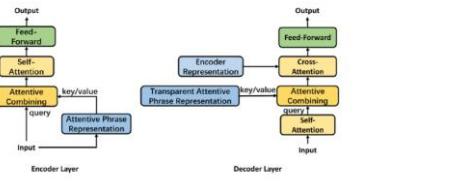
| Source task | TinyImageNet | | | | | | ImageNet |
|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|----------|
| | CIFAR-100 | STL-10 | CUB200 | MIT67 | Stanford Dogs | Stanford Dogs | |
| Scratch | 67.69 ± 0.02 | 65.18 ± 0.02 | 42.15 ± 0.02 | 48.91 ± 0.02 | 36.93 ± 0.02 | 58.08 ± 0.02 | |
| LwF | 69.23 ± 0.08 | 68.64 ± 0.08 | 45.52 ± 0.08 | 53.73 ± 0.08 | 39.73 ± 0.08 | 60.33 ± 0.08 | |
| AT (one-to-one) | 70.00 ± 0.02 | 70.00 ± 0.02 | 46.00 ± 0.02 | 54.00 ± 0.02 | 39.00 ± 0.02 | 60.00 ± 0.02 | |
| LwF+AT (one-to-one) | 68.75 ± 0.02 | 75.00 ± 0.02 | 58.90 ± 0.02 | 61.42 ± 0.02 | 60.20 ± 0.02 | 72.67 ± 0.02 | |
| FM (single) | 69.40 ± 0.02 | 75.00 ± 0.02 | 47.60 ± 0.02 | 55.15 ± 0.02 | 42.93 ± 0.02 | 66.05 ± 0.02 | |
| FM (pair) | 70.00 ± 0.02 | 75.00 ± 0.02 | 47.60 ± 0.02 | 55.15 ± 0.02 | 42.93 ± 0.02 | 66.05 ± 0.02 | |
| L2T-w (single) | 70.27 ± 0.02 | 74.35 ± 0.02 | 51.85 ± 0.02 | 60.41 ± 0.02 | 46.25 ± 0.02 | 69.16 ± 0.02 | |
| L2T-w (one-to-one) | 70.02 ± 0.02 | 74.42 ± 0.02 | 56.64 ± 0.02 | 59.78 ± 0.02 | 48.19 ± 0.02 | 69.84 ± 0.02 | |
| L2T-w (all-to-all) | 70.96 ± 0.01 | 78.31 ± 0.02 | 65.05 ± 0.01 | 64.85 ± 0.02 | 63.08 ± 0.02 | 78.08 ± 0.02 | |

Qualitative Examples

- Paraphrasing Module (w/o vs w/)

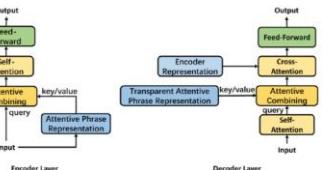
Introduction

- Since the phrase representations are produced and attended at each encoder layer, the encoding of each layer is also enhanced with phrase-level attention computation;



Introduction

- The phrase representations are produced by phrase-level attention .



Related work

- Action proposals is an essential part of many methods for action detection, explored by a number of recent
- More related to our work, previous methods [9, 18, 24, 34, 35] explore the temporal order, either by predicting the exact order of consecutive frames [18, 35] or verifying their partial order [9, 24, 34].
- In the video domain, motion has been used as a cue for learning video representations in [1, 26, 33, 7].
- The notion of actionness was first introduced in [5] as a confidence measure of intentional bodily movement of biological agents.



Related work

- Action detection is a key tool for action detection
- Predicting the exact order of consecutive frames :
- Motion has been used as a cue .
- Actionness was first introduced in biological agents



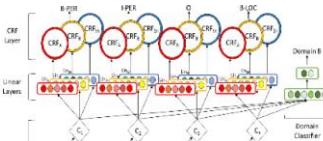
Base Architecture

- The feed-forward layers capture the domainspecific and -independent information by using private output layers for each domain and one shared output layer.
- Word embeddings are derived from a combination GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017) pre-trained word embeddings, as used in (Ma and Hovy, 2016).
- The global objective function is the combination of the NER loss function and domain loss:
- The domain classification objective is to minimize the crossentropy loss $L_{\text{domain}}(x, y_d)$ for an input x with domain label y_d .
- We propose a new architecture based on the BiLSTM \oplus CRF model tailored to the three proposed experimental setups.



Base Architecture

- Use private output layers for each domain
- (ma et al . , 2017)
- The global objective function is the combination of the ner loss
- The domain classification objective is to minimize the crossentropy loss .
- We propose a new architecture based on the bilstm model .



Conclusions

- In this paper, we proposed two parameterized benchmark games in which EFCE exhibits interesting behaviors.
- We also provided an alternative saddle-point formulation of EFCE and demonstrated its merit with a simple subgradient method which outperforms standard LP based methods.
- We analyzed those behaviors both qualitatively and quantitatively, and isolated two ways through which a mediator is able to compel the agents to follow the recommendations.
- We hope that our analysis will bring attention to some of the computational and practical uses of EFCE, and that our benchmark games will be useful for evaluating future algorithms for computing EFCE in large games.

Conclusions

- We proposed two parameterized games in which efce exhibits interesting behaviors .
- We propose a saddle-point formulation of efce
- Two ways to compel the agents to follow the recommendations
- We show that our analysis will be useful for computing efce in large games

Qualitative Examples

- Applying Design Ideas

Introduction

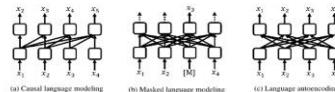
- Weak supervision in text classification has the burden of human experts
- How to train a deep neural network?
- We have performed experiments on real-word datasets
- Identify words that are discriminative and highly label-indicative

Introduction

- Weak supervision in text classification has the burden of human experts
- How to train a deep neural network?
- We have performed experiments on real-word datasets
- Identify words that are discriminative and highly label-indicative

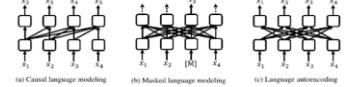
Language Model Baselines

- Feeding the token in the input sequence
- We can obtain the contextualized embeddings



Language Model Baselines

- Feeding the token in the input sequence
- We can obtain the contextualized embeddings



Face photo to drawing generator G

- No pairings need to exist between two domains
- Require the inverse generator to reconstruct a face photo
- A strict loss function for cycle-consistency loss
- $P(p, s)$



Face photo to drawing generator G

- No pairings need to exist between two domains
- Require the inverse generator to reconstruct a face photo
- A strict loss function for cycle-consistency loss
- $P(p, s)$

Detection Results on CityPersons

- Ablation study of PBM
- Breaking the curse of many agents with events

| Method | PPFE | R2NMS | R | HO |
|----------|--------|-------|-------------|-------------|
| Baseline | - | - | 13.8 | 59.0 |
| PBM | concat | - | 12.5 | 57.3 |
| PBM | concat | ✓ | 12.1 | 57.0 |
| PBM | mask | - | 12.3 | 54.9 |
| PBM | mask | ✓ | 11.1 | 53.3 |

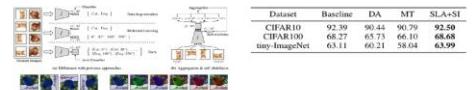
Detection Results on CityPersons

- Ablation study of PBM
- Breaking the curse of many agents with events

| Method | PPFE | R2NMS | R | HO |
|----------|--------|-------|-------------|-------------|
| Baseline | - | - | 13.8 | 59.0 |
| PBM | concat | - | 12.5 | 57.3 |
| PBM | concat | ✓ | 12.1 | 57.0 |
| PBM | mask | - | 12.3 | 54.9 |
| PBM | mask | ✓ | 11.1 | 53.3 |

Multi-task Learning with Self-supervision

- Depending on the type of training samples, the statistical characteristics of the augmented training samples
- Remove unnecessary invariant property of the classifier
- Aggregate the corresponding conditional probabilities to improve the classification accuracy



Multi-task Learning with Self-supervision

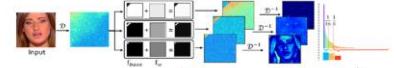
- Depending on the type of training samples, the statistical characteristics of the augmented training samples
- Remove unnecessary invariant property of the classifier
- Aggregate the corresponding conditional probabilities to improve the classification accuracy



| Dataset | Baseline | DA | MT | SLA+SI |
|---------------|----------|-------|-------|--------------|
| CIFAR10 | 92.39 | 90.44 | 90.79 | 92.50 |
| CIFAR100 | 69.43 | 67.47 | 68.48 | 68.48 |
| tiny-ImageNet | 63.11 | 60.21 | 58.04 | 63.99 |

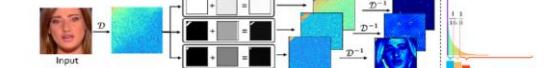
FAD Frequency-Aware Decomposition

- The frequency of frequency-aware components can be inversely transformed
- Number of filters (n=512)
- The frequency filtering is a special case of the input image



FAD Frequency-Aware Decomposition

- The frequency of frequency-aware components can be inversely transformed
- Number of filters (n=512)
- The frequency filtering is a special case of the input image



Conclusion

- DOC2PPT serves as a **multi-modal summarizer** to generate slide from academic documents
- We propose **hierarchical architecture**, **text-figure constrain**, and **paraphrasing module** to improve the quality of slide generation
- DOC2PPT **provides useful outline and flow** to make building a slide more efficiency

thanks!

