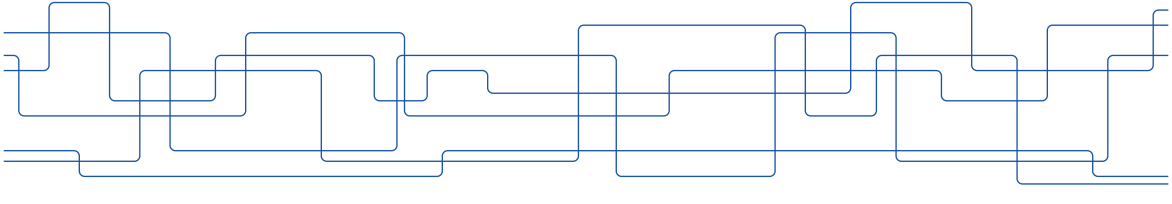
News article segmentation using multimodal input using Mask R-CNN and Sentence-BERT

Gustav Henning



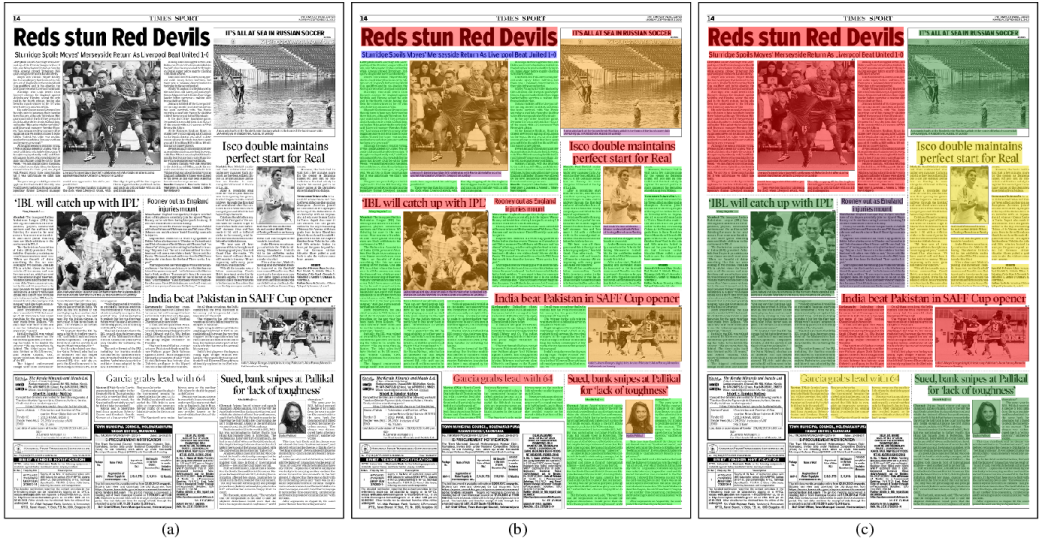
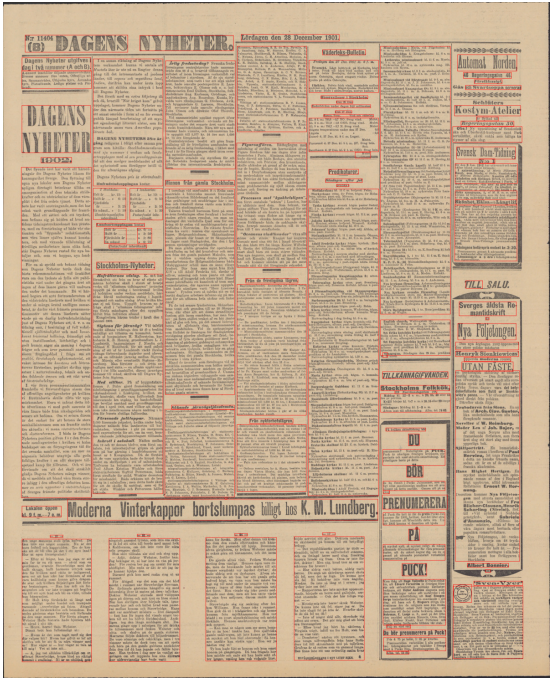


Figure: Example of newspaper article segmentation. [Bansal et al 2014]

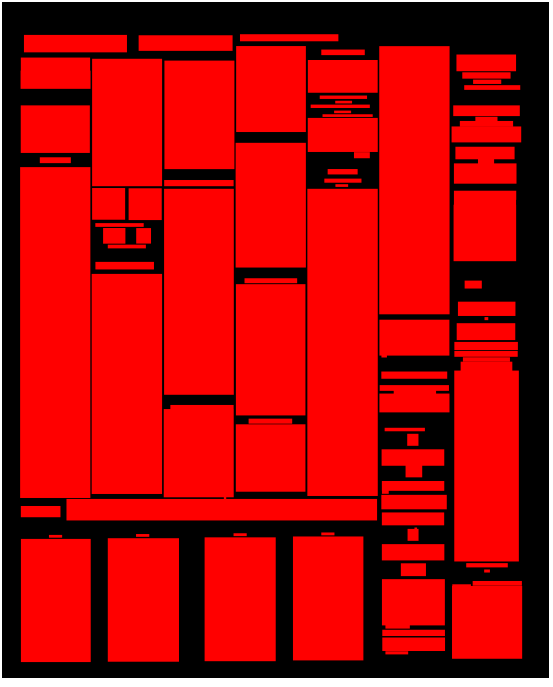
Figure: Mask R-CNN output for MS COCO (Common objects in Context). [He et al 2017]

Is this feasible?

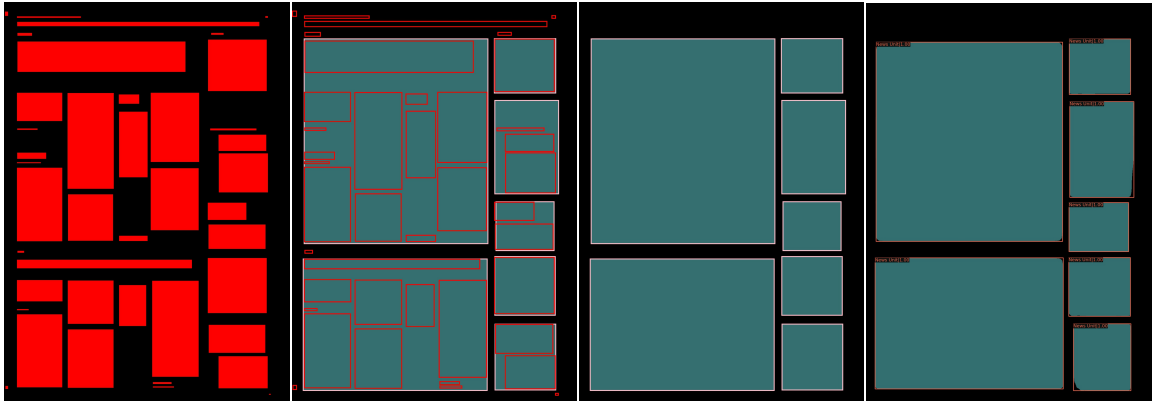
What we require in newspaper article segmentation	Can Mask R-CNN do this?	How?
Detecting multiple instances of the same class	Yes	Region Proposal Network
Detecting objects at different scales	Yes	Feature Pyramid Network
Representing several elements as one instance	? (Yes)	Convolutional Neural Network
Comparing texts based on semantics	No	-



(a) OCR bounding boxes visualized in red.



(b) OCR visualized in red without raw data.



(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page consisting of news articles.

Why newspaper article segmentation?

Focus on digitization of documents

- Pages are scanned on an image level
- Optical Character Recognition extracts boxes of text
- Document layout analysis: Determining reading order, meaningful article segmentation remains a challenge
- Meaningful segments could enable e.g. search on article level, instead of page level.

Problems and goals

Problem statement

How well can neural networks, developed for object detection and segmentation of real-world objects, perform in the domain of news article segmentation?

Goals

- Do multimodal ANNs outperform unimodal ANNs, using image and text input versus only using image input, in the field of instance segmentation?
- Is there a difference in how well ANNs perform in periods of changing typographic design with respect to modality?
- What effect does increasing the amount of annotated data have on model performance and its ability to generalize on previously unseen data?

What is Mask R-CNN?

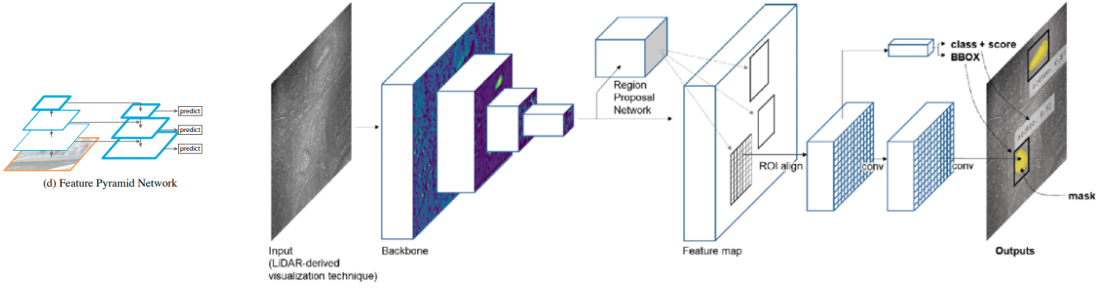
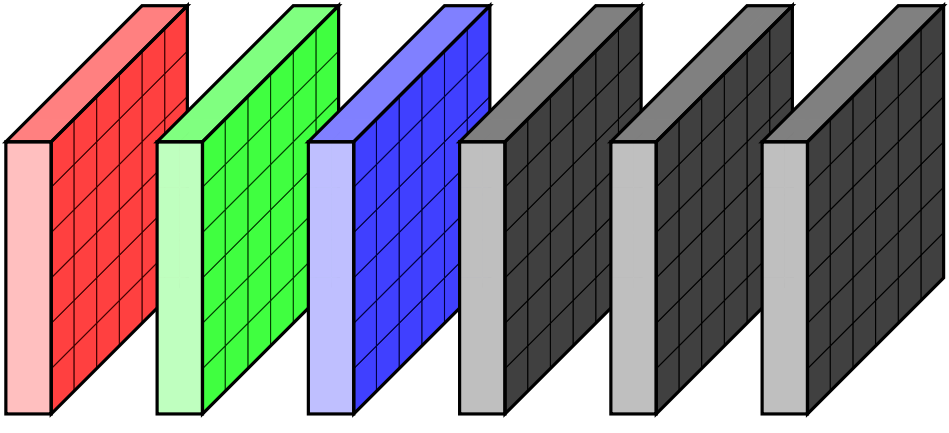


Figure: Mask R-CNN architecture. [Guyot et al 2021]

How do we combine the visual and textual modalities?



3 native RGB channels + 3 text embedding channels

Sentence transformers used in this thesis

Table: Sentence transformers

Model name	**Dimensions**
all-mpnet-base-v2	384
all-MiniLM-L6-v2	384
multi-qa-mpnet-base-dot-v1	384
all-distilroberta-v1	384
KBLab/sentence-bert-swedish-cased	768
KB/bert-base-swedish-cased	768

What is (Sentence-) BERT?

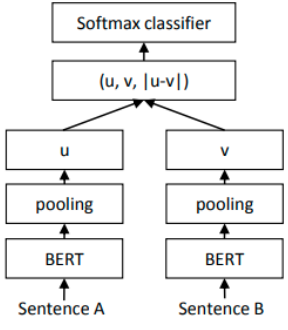


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

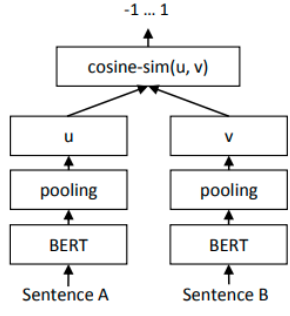
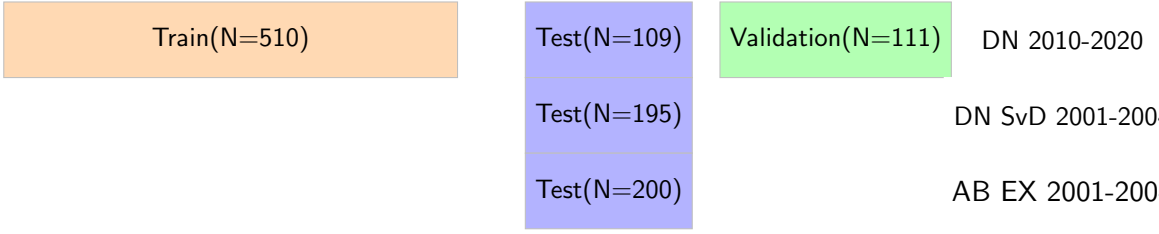


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

Figure: Sentence BERT extension of the original BERT. [Reimers et al 2019]

Datasets created in this thesis



Morning newspapers: DN = Dagens Nyheter, SvD = Svenska Dagbladet
Evening newspapers: AB = Aftonbladet, EX = Expressen
DN 2010-2020 is split (70/15/15) using random sampling.

Labeling Strategy

Classes:

- News Unit
- Advertisement
- Listing
- Weather
- Death Notice
- Game
- Publication Unit (All of the above as 1 class, for testing class confusion)

Caveats:

- Rectangles to polygons
- Spatial blocking of content
- Per page annotations (Implied content)

Experiments

- Unimodal vs multimodal (Image vs image + text)
- Impact of class labels (6 classes vs 1 class)
- Impact of dataset size (Random sampling different % of train data)

Performance Metrics

$$IoU = \frac{Area\ Overlap}{Area\ Union} \tag{1}$$

A prediction is considered positive when above a certain threshold $\tau \in [0, 1]$ of IoU.

$$Precision@\tau = P@\tau = \frac{TP}{(TP + FP)} \tag{2}$$

mean Average Precision:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{3}$$

mean Average Precision at different IoU thresholds:

$$mAP^{50} = mAP^{(IoU=0.5)} \tag{4}$$

Metrics are reported on the segmentation mask (m) and bounding box (bb) respectively.

Major results

Impact of dataset size - Varying a % of the training data using random sampling. Each experiment was repeated 5 times.

Table: Gradually increasing dataset size - 6 classes - In-domain test set

% Train	mAP_m	mAP_m^{50}	mAP_m^{75}	mAP_{bb}	mAP_{bb}^{50}	mAP_{bb}^{75}	σ
25%	0.723	0.848	0.790	0.688	0.847	0.781	0.027
50%	0.883	0.964	0.945	0.850	0.965	0.943	0.003
75%	**0.939**	**0.989**	**0.981**	**0.927**	**0.989**	**0.980**	0.004
90%	0.934	0.976	0.963	0.921	0.975	0.962	0.006
100%	0.901	0.960	0.935	0.895	0.960	0.934	0.005

The standard deviation σ denotes the average of the standard deviations of each cell in the row. All results in this table derive from the model resnet-50-FPN. All prediction visualizations in this presentation are produced by the top scoring model (75%).

Major Results 2

Table: Top 5 results - 6 classes - In-domain test set - 100% of train data

Final name	mAP_m	mAP_m^{50}	mAP_m^{75}	mAP_{bb}	mAP_{bb}^{50}	mAP_{bb}^{75}
x101-32x8d	0.910	0.964	0.936	0.906	0.964	0.937
x101-32x4d	0.906	0.957	0.929	0.901	0.957	0.932
x101-64x4d	0.903	0.954	0.934	0.893	0.954	0.934
r101	0.901	0.951	0.928	0.885	0.953	0.927
x101-64x4d-all-mpnet	0.881	0.952	0.927	0.868	0.954	0.929

Major Results 3

Multimodal models tended to outperform unimodal models on the out-of-domain test set when all classes are set to the same class: Publication Unit.

Table: Top 5 results - 1 class - Out-set

Final name	mAP_m	mAP_m^{50}	mAP_m^{75}	mAP_{bb}	mAP_{bb}^{50}	mAP_{bb}^{75}
r50-MiniLM-1c	0.465	0.622	0.490	0.458	0.622	0.488
r50-all-mpnet-1c	0.455	0.612	0.480	0.449	0.611	0.483
x101-64x4d-all-mpnet-1c	0.447	0.588	0.474	0.441	0.585	0.474
r50-1c	0.438	0.560	0.460	0.432	0.556	0.462
x101-all-mpnet-1c	0.433	0.564	0.459	0.426	0.564	0.456

Minor Results

Table: Top performing class choice

Test set	**Number of classes**
In-set	6 classes
Near-set	1 class
Out-set	6 classes

Discussion and conclusion

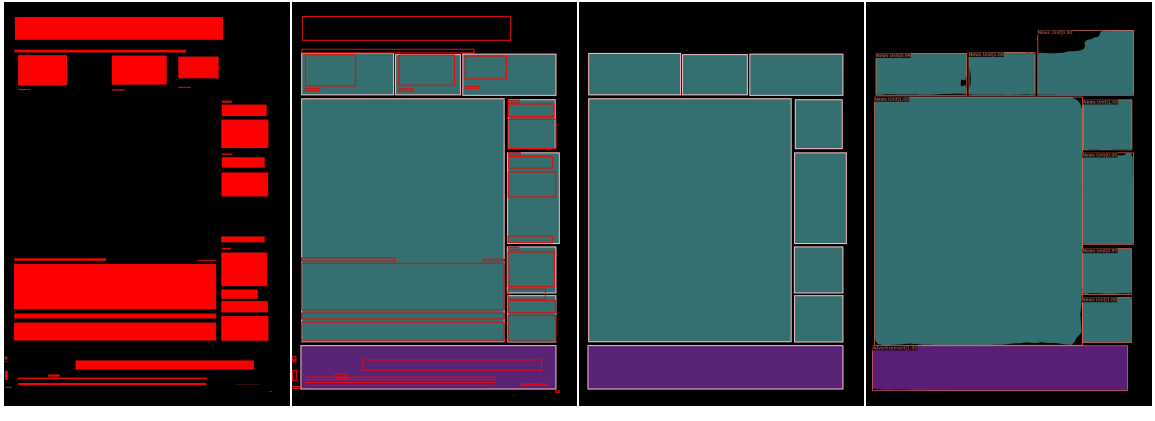
- Higher mAP is required in this domain for usefulness (simpler masks)
- Vanilla Mask R-CNN has been shown to perform well on in-domain test sets.
- Increase in error in dataset size experiment? Small dataset size + more complex examples left out in 75%, and not represented in test set.
- The greater the size of the training dataset, the better performance on near- and out-of-domain test sets (for both 6 classes and 1 class).
- 6 classes vs 1 class differs in best performance between in-, near- and out-of-domain test sets.
- We make no claims on the ability to generalize across typologies, but we see a consistent drop in performance proportional to 'distance' in typology and timespan.
- Multimodal performance could be explained by large and sparse dimensions, causing the model to ignore those dimensions. (Future work)

Future work

- Domain-specific architecture
- Normalization of textual embeddings
- Dimensionality reduction of textual embeddings
- Dataset specific image normalization values and training without pretrained weights.
- Text granularity (sentence level, word level, one hot encoding)

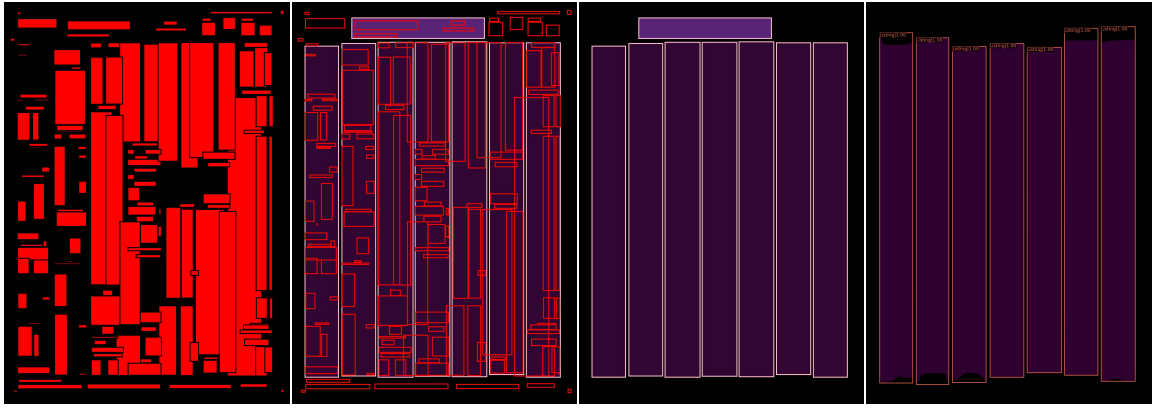
Demo

- Left: First 3 dimensions of all-mpnet-base-v2 vector representations normalized as RGB colors.
- Right: Labels produced by vanilla Mask R-CNN, trained on 75% of the training data. Close to ground truth in segmentation mask. (0.939 mAP_m)



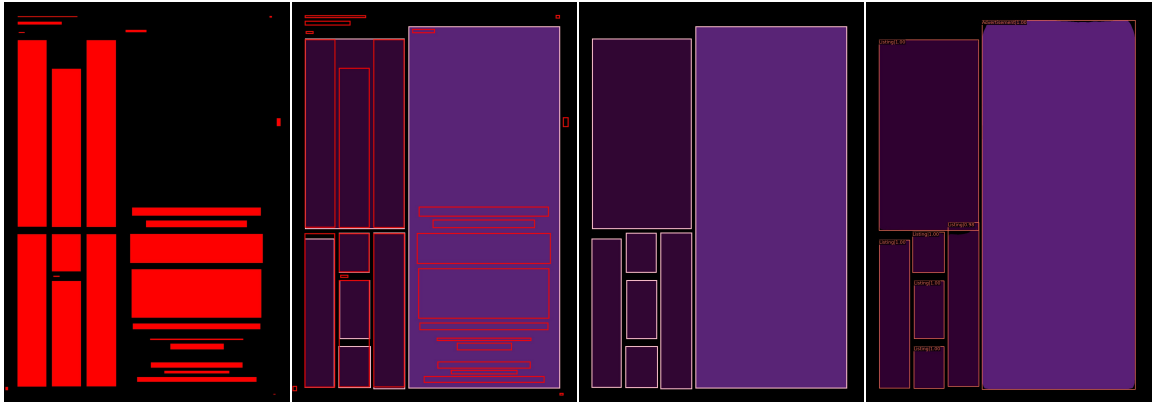
(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A front page newspaper page from the DN 2010-2020 Test subset.



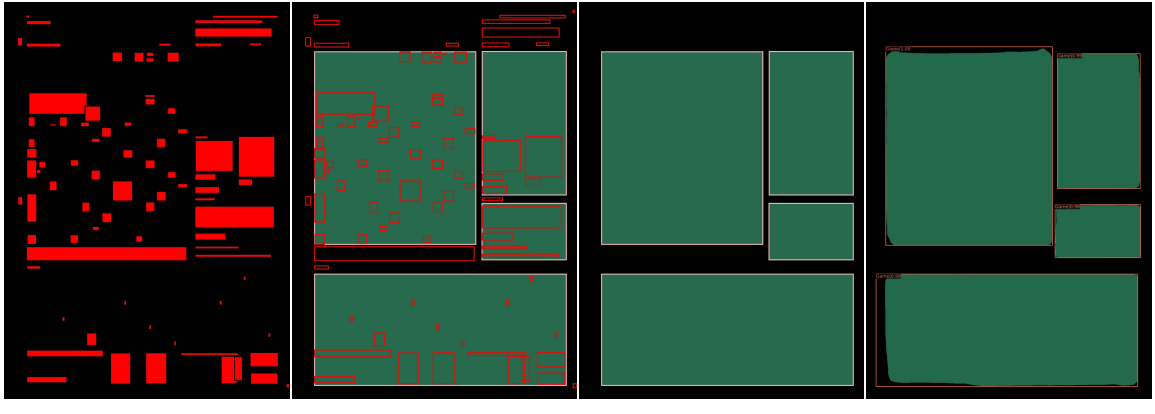
(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page with listing prices of funds and stocks.



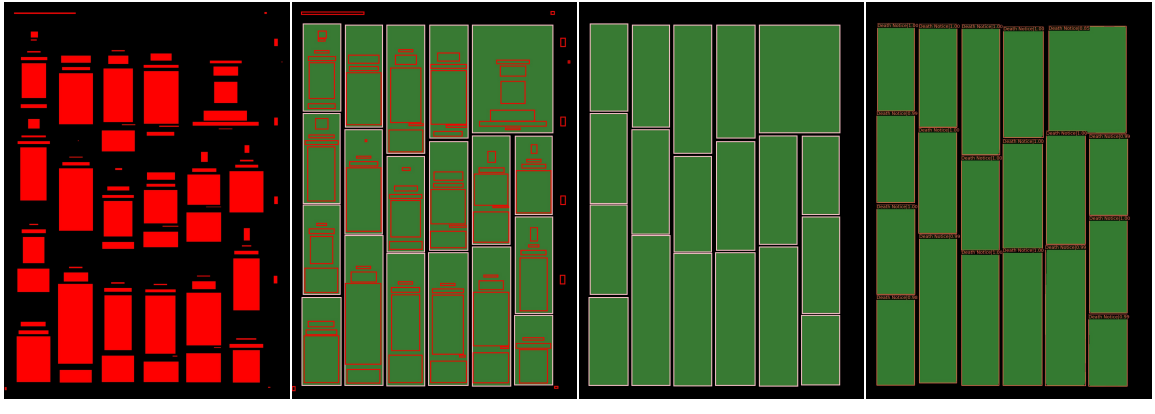
(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page with listings of TV/Radio timetables + advertisement.



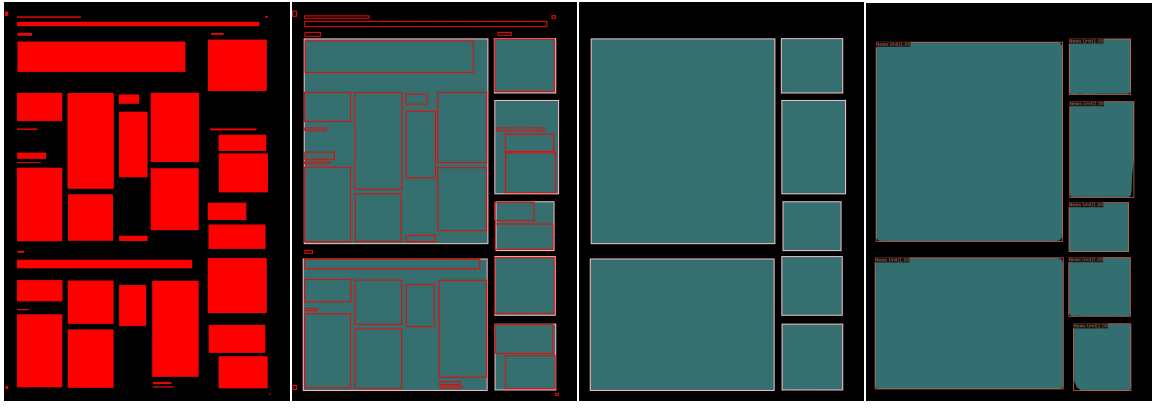
(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page with different types of games.



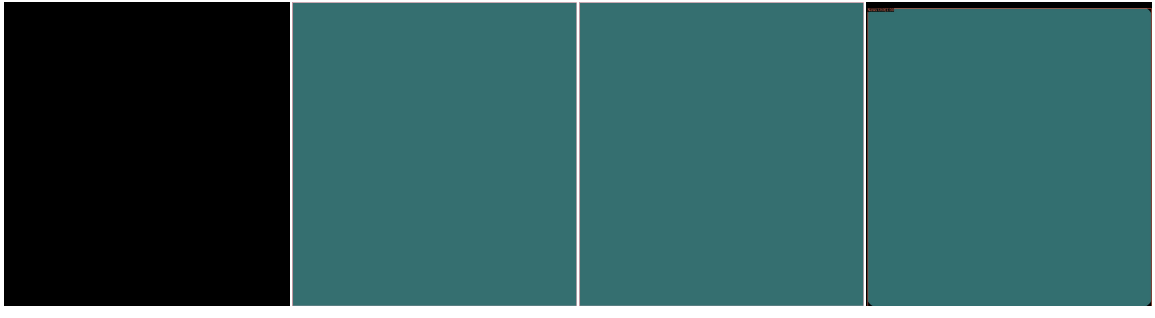
(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page with obituaries.



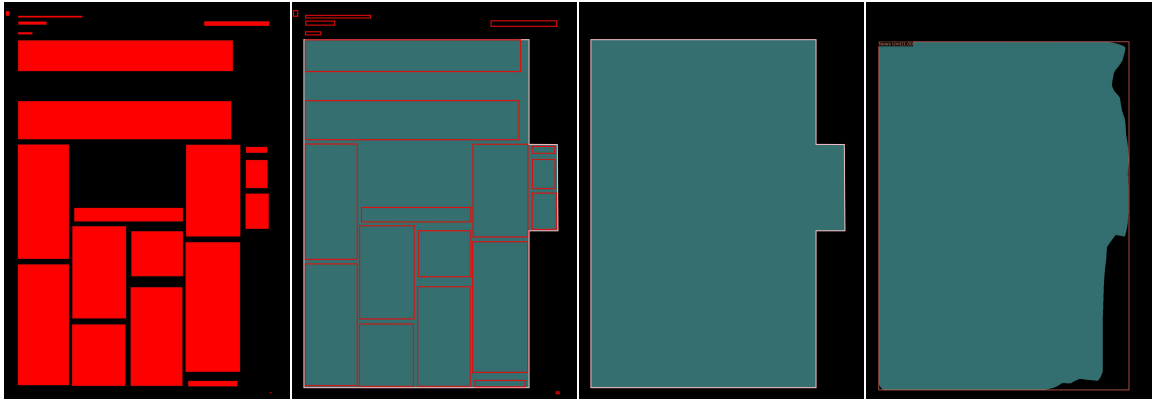
(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page consisting of news articles.



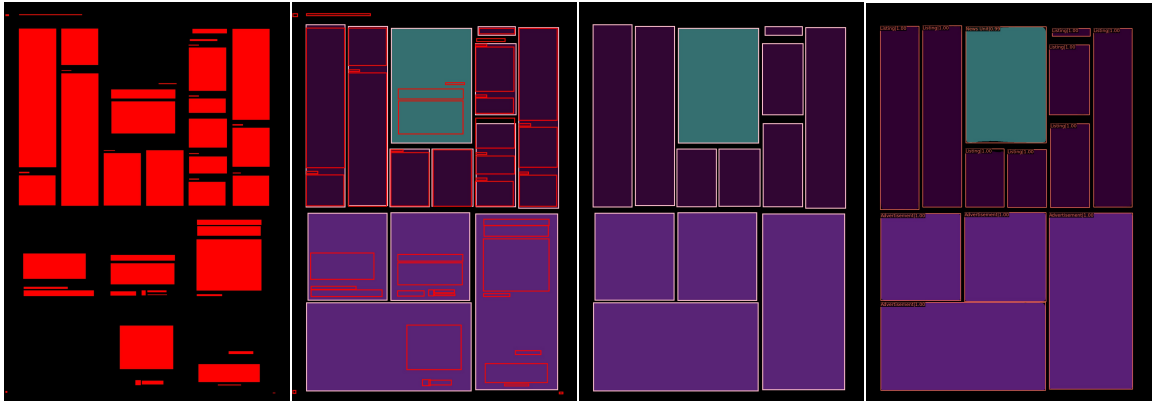
(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page consisting of a single large image, part of a multi-page article



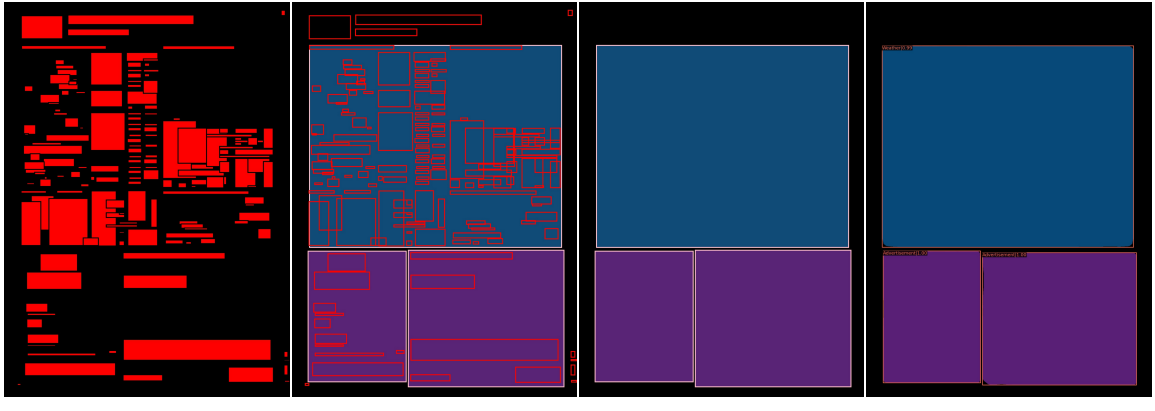
(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page consisting of a large news article: label extension demonstration.



(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page with event listings.



(a) OCR (b) GT Label + OCR (c) GT Label (d) Model prediction

Figure: A newspaper page consisting of weather reports and advertisement.

Thanks for listening.