HYCEDIS: HYbrid Confidence Engine for Deep Document Intelligence System

Sinh Nguyen; Bach Tran; Tuan Anh Nguyen Dang; Duc Nguyen; Hung Le

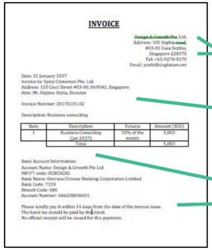
Document Intelligence Workshop
@ KDD 2021

cinnamon

Confidence in AI models: the notorious problem

- AI models have shown **impressive predictive performance** on many problems, but they are poorly calibrated in term of confidence.
- To have widespread real-world adoption, we need to know when **we can trust** the model output.
- Lot of **mission-critical** use-cases require strict estimation of models' confidence.
- Unfortunately current AI models are **hard to understand** their behaviour / correctness.

Input data



Invoice
(Free Format)

Output data

Item	Data
Invoicing Address	20 Anson Rd, Singapore 079912
Company name	Cinnamon Pte. Ltd.
Atten	Miku Hirano
Original Address	150 Cecil St, Singapore 069543
Account Name	Joseph Tan
Account Number	68430875
Content	Business Consulting
Amount	¥300,000
Tax	¥24,000
Total Amount	¥324,000
Payment Due	30th September 2017

Extract information to insert into database

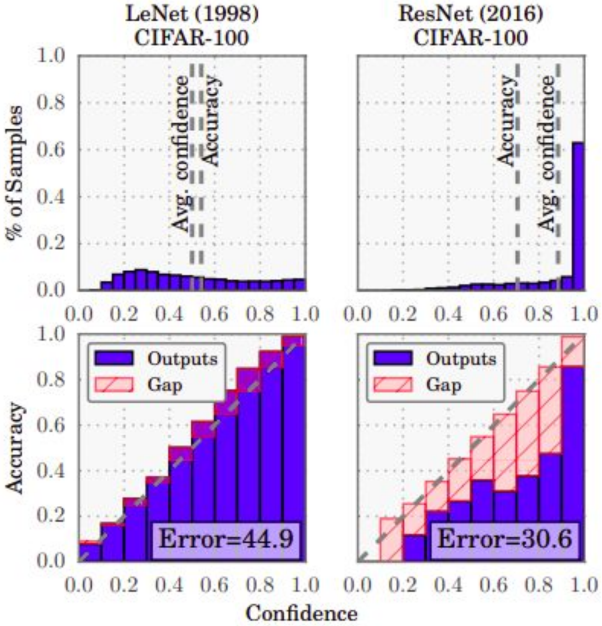


Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

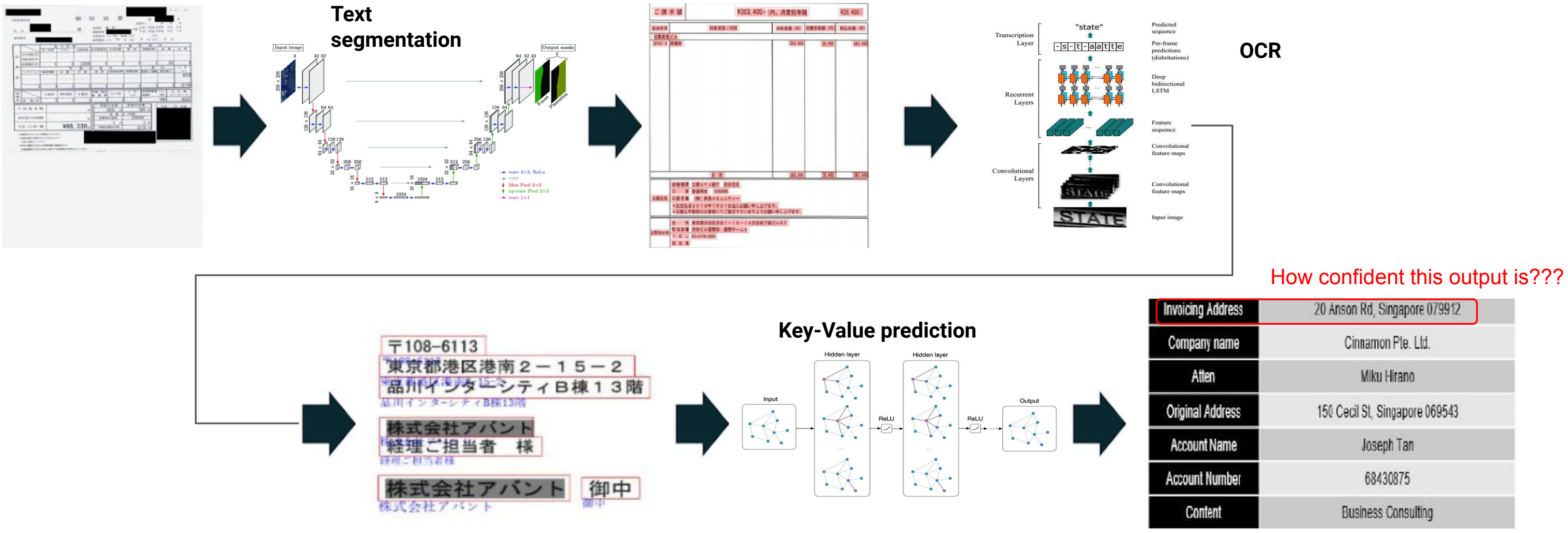
[1] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks

Document Intelligence System confidence estimation: problem definition

- Document Intelligence System (DIS): consists of 3 **IE networks**
    - Texline segmentation: U-net based
    - OCR: CRNN + CTC loss
    - **Key-value** prediction: Graph convolution NN

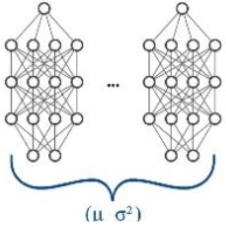
- Output a confidence score for each prediction (field) in DIS output: Reflect the likelihood of correctness (higher confidence score ~ higher chance of prediction being correct)

- **Binary classification problem:** Separate the *correct / in-correct* prediction



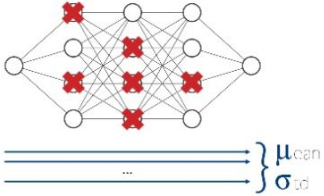
Related literature

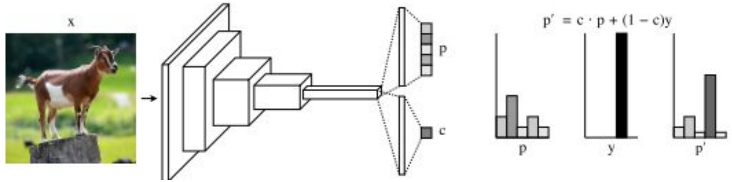
- **Rich literature** in confidence / uncertainty estimation of AI models
- Main approaches:
    - Variational inference [1]
    - Bayesian model
        - Deep ensemble [2]
        - MC Dropout [3]
    - Logits calibration [5]
    - Confidence estimators [6]
    - Out-of-Distribution detector [4]
    - .. and more



$$\mu_c = \frac{1}{M}\sum_{i=1}^{M} \mu_i$$

$$\sigma_c^2 = \frac{1}{M}\sum_{i=1}^{M} (\sigma_i^2 + \mu_i^2) - \mu_c^2$$

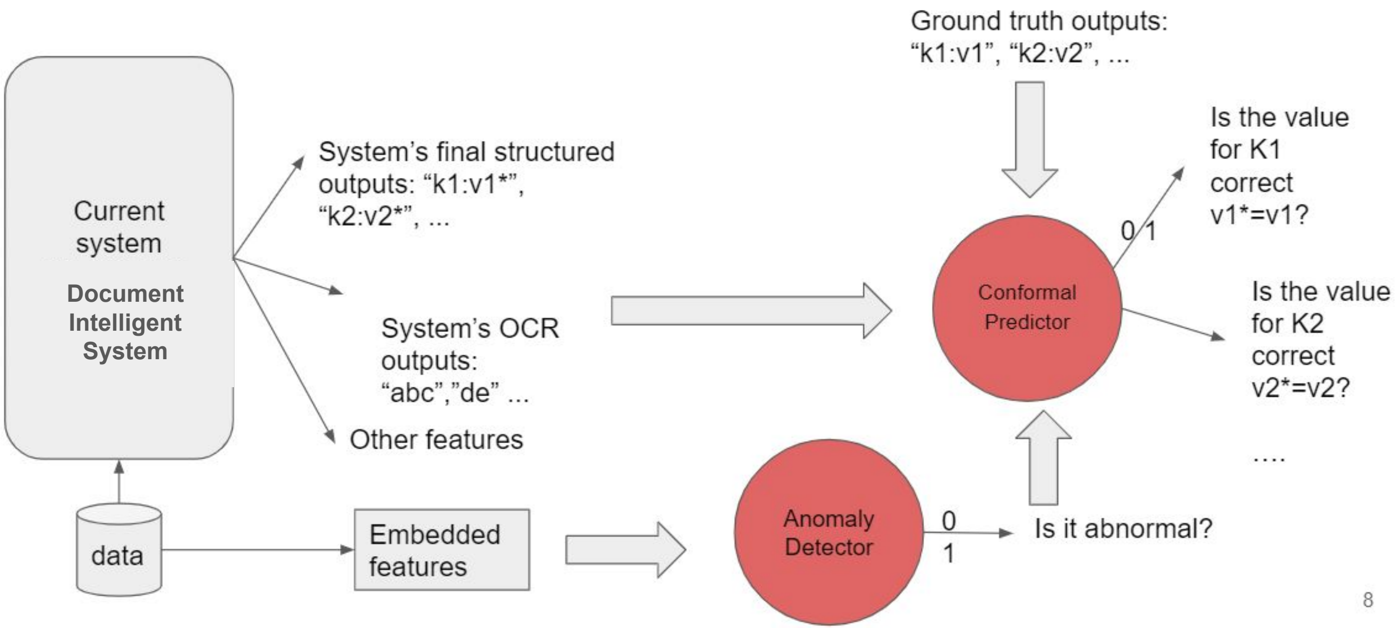




[1] Posch, Konstantin, Jan Steinbrener, and Jürgen Pilz. "Variational inference to measure model uncertainty in deep neural networks."
[2] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles."
[3] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." international conference on machine learning. PMLR, 2016.
[4] Hendrycks, Dan, Mantas Mazeika, and Thomas Dietterich. "Deep anomaly detection with outlier exposure."
[5] Guo, Chuan, et al. "On calibration of modern neural networks." International Conference on Machine Learning. PMLR, 2017.
[6] Mor, Noam, and Lior Wolf. "Confidence prediction for lexicon-free ocr." 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

Our proposed solution

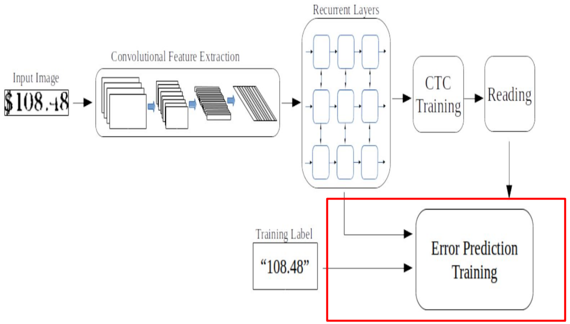
- Based on these observations, we design a holistic solution that utilizes all sources of information and takes advantage of both confidence predictor and anomaly (OOD) detector.



Conformal Predictor: Motivation

[1] attach an auxiliary head to the model to predict the correctness of OCR model.

-> Achieve at least 2x AUC compared to other previous methods.

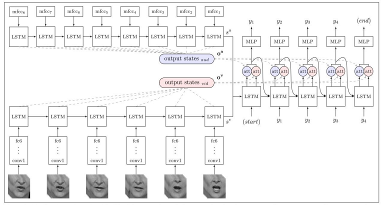


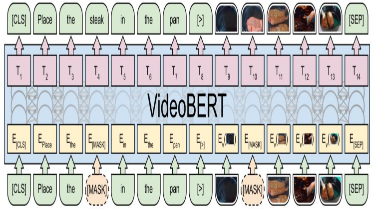
[1]: Mor, N., & Wolf, L. (2018). *Confidence Prediction for Lexicon-Free OCR*. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, *2018-January*, 218–225. https://doi.org/10.1109/WACV.2018.00030

[2], [3]

- **Learn joint visual-linguistic representation in self-supervised manner.**
- **Combine Image encoder and Audio encoder to recognise phrases and sentences being spoken by human.**

-> Efficiently leverage cross-modality information.



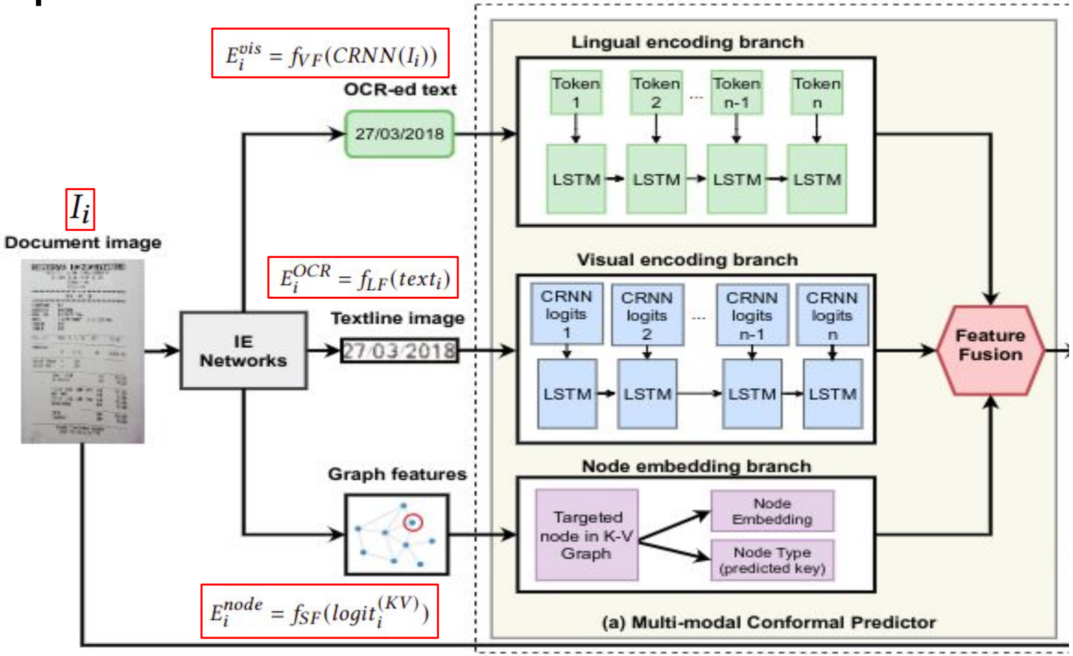


[2]: Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: *A joint model for video and language representation learning*. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*, 7463–7472. https://doi.org/10.1109/ICCV.2019.00756

[3]: Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman: *Lip Reading Sentences in the Wild*

Conformal Predictor: Architecture

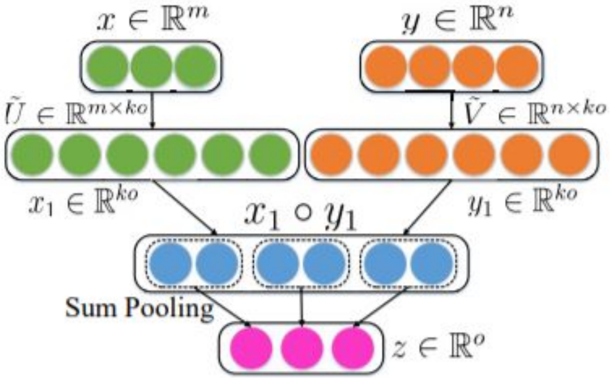
Multi-modal conformal predictor (MCP)



Feature fusion: $F_i = f_{Fusion}(E_i^{vis}, E_i^{OCR}, E_i^{node})$

Conformal Predictor: Feature Fusion

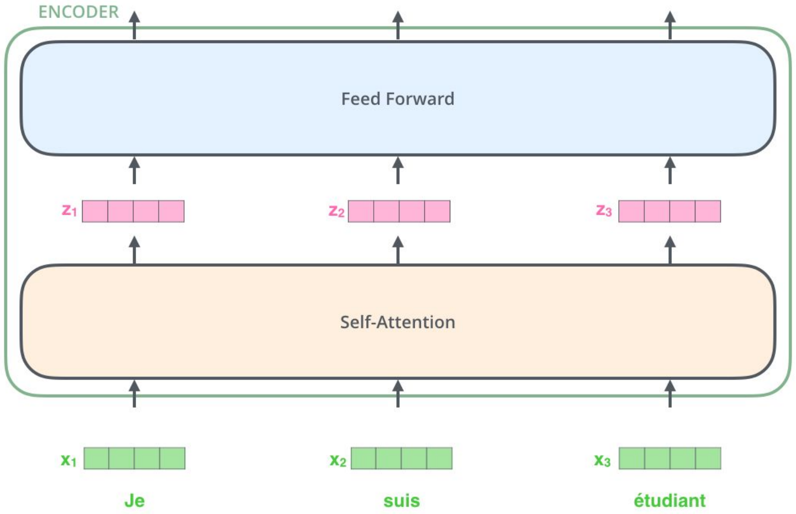
Bilinear pooling



(a) Multi-modal Factorized Bilinear Pooling

Yu, Zhou, et al. "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering." *Proceedings of the IEEE international conference on computer vision*. 2017.

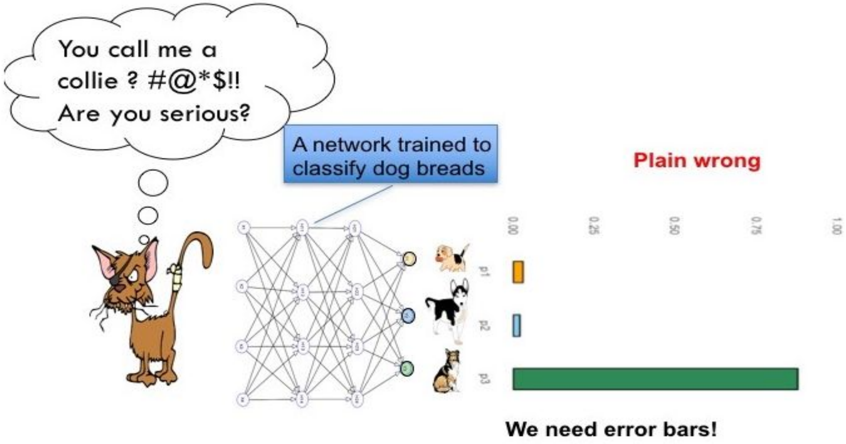
Attention-based pooling (Transformer)



Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Anomaly Detector: Motivation

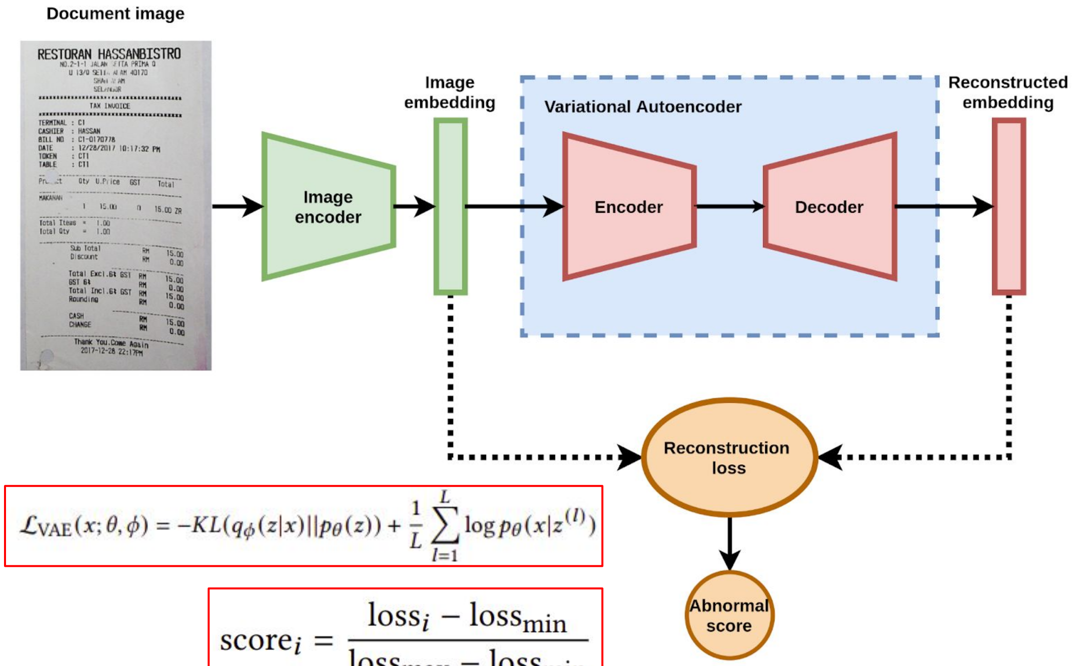
- Problem with out-of-distribution data: neural network yields very high confidence for outliers
- To overcome: anomaly detection



[1] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. 2016.
[2] Hein, Matthias, Maksym Andriushchenko, and Julian Bitterwolf. "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Anomaly Detector: Variational AutoEncoder

- VAE [1,2] to detect anomaly
- Represent image by lower-dimensional representation



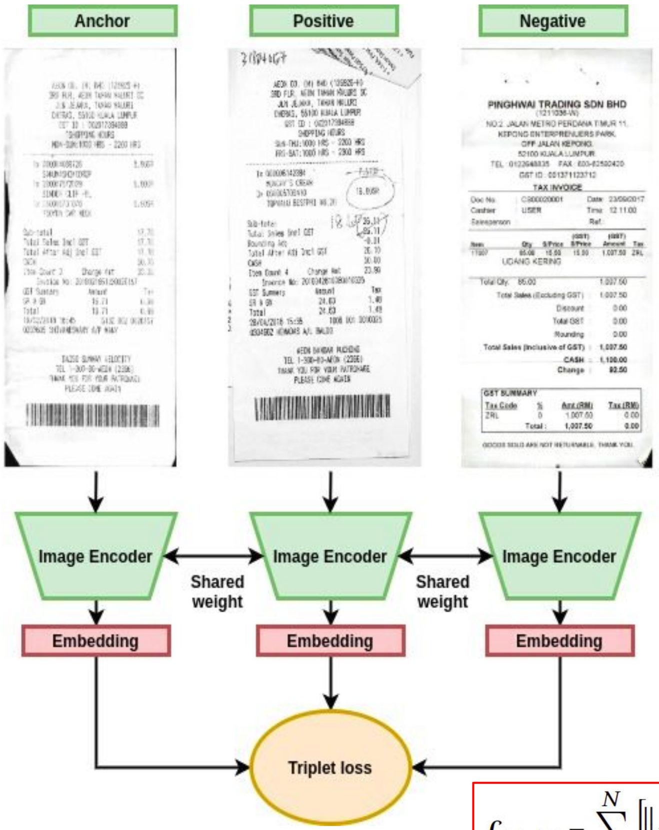
$$\mathcal{L}_{\mathrm{VAE}}(x;\theta,\phi) = -KL(q_\phi(z|x)||p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|z^{(l)})$$

$$\mathrm{score}_i = \frac{\mathrm{loss}_i - \mathrm{loss}_{\mathrm{min}}}{\mathrm{loss}_{\mathrm{max}} - \mathrm{loss}_{\mathrm{min}}}$$

[1] Stochastic backpropagation and approximate inference in deep generative models

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Anomaly Detector: Cluster-oriented Encoder

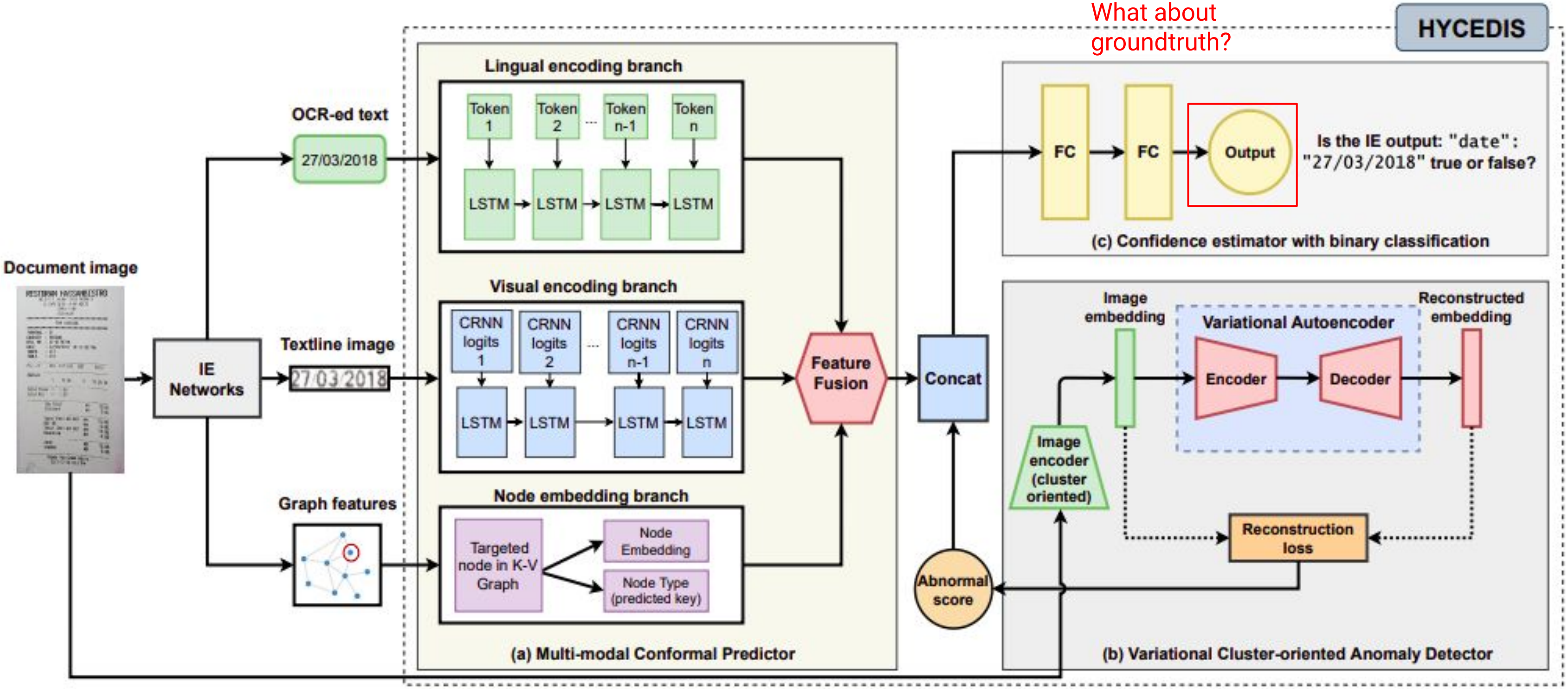


t-SNE visualization of the learnt embeddings

$$\mathcal{L}_{\text{triplet}} = \sum_{i}^{N} \left[\left\| f_I(x_i^a) - f_I(x_i^p) \right\|_2^2 - \left\| f_I(x_i^a) - f_I(x_i^n) \right\|_2^2 + \alpha \right]$$

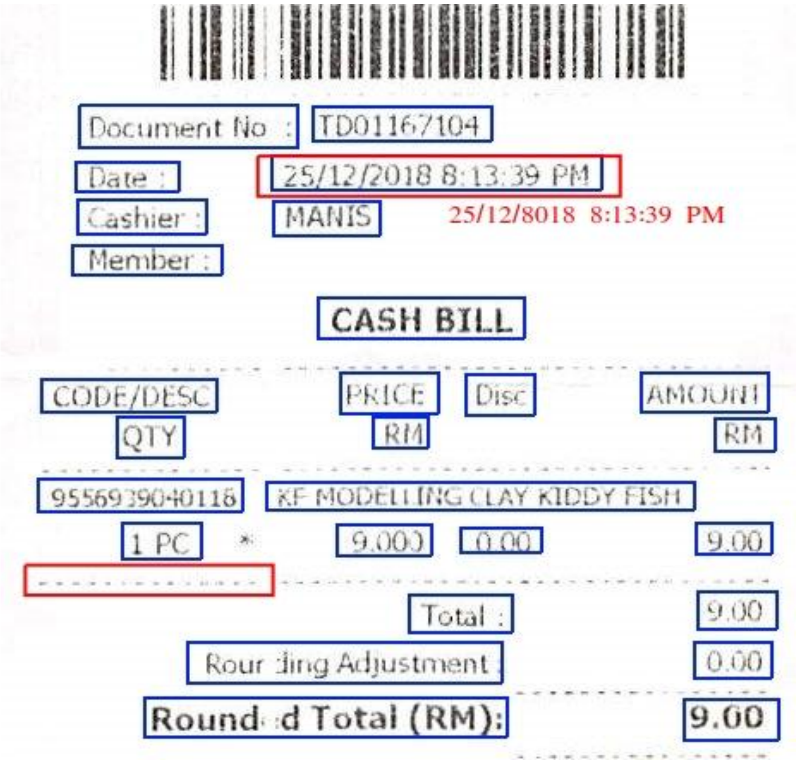
HYCEDIS architecture: Conformal Predictor and Anomaly Detector

Cross-entropy loss.
What about
groundtruth?



HYCEDIS architecture: ground truth

False: wrong OCR/wrong box/wrong key-value



True: correct OCR, correct box (IoU>thresh), correct key-value

Experiment setup: Japanese task (in-house dataset)

- **In house-1:** combination of our company's data, which was mainly used for training and testing purpose.
- **In house-2**: Invoice-like dataset with visually distinct format (***Out-of-Distribution data***)

	In house-1	**In house-2**
Training	835 (original) + 535 (augmented) files	
Testing	338 files	68 files
Keyword (bold is important keywords)	account_name, account_number, account_type, amount_excluding_tax, amount_including_tax, bank_name, **branch_name**, **company_address**, company_department_name, **company_fax**, **company_name**, **company_tel**, **company_zipcode**, delivery_date, document_number, invoice_number, issued_date, item_line_number, **item_name**, **item_quantity**, item_total_amount, item_unit, **item_unit_amount**, payment_date, tax	**branch_name**, **company_address**, **company_fax**, **company_name**, **company_tel**, **company_zipcode**, **item_name**, **item_quantity**, **item_unit_amount**
Number of formal keys	25 (9 common keys)	12 (9 common keys)

Experiment setup: English task

- **SROIE:** a variant of Task 3 of "Scanned Receipts OCR and Information Extraction" (SROIE) that consists of a set of store receipts with 4 semantic fields: Company, Date, Address, and Total price.
- **Consolidated Receipt Dataset (CORD)**: a set of store receipts with 800 training, 100 validation, and 100 testing examples with more 30 semantic entities including menu name, menu price, and so on. (***Out-of-distribution data***)

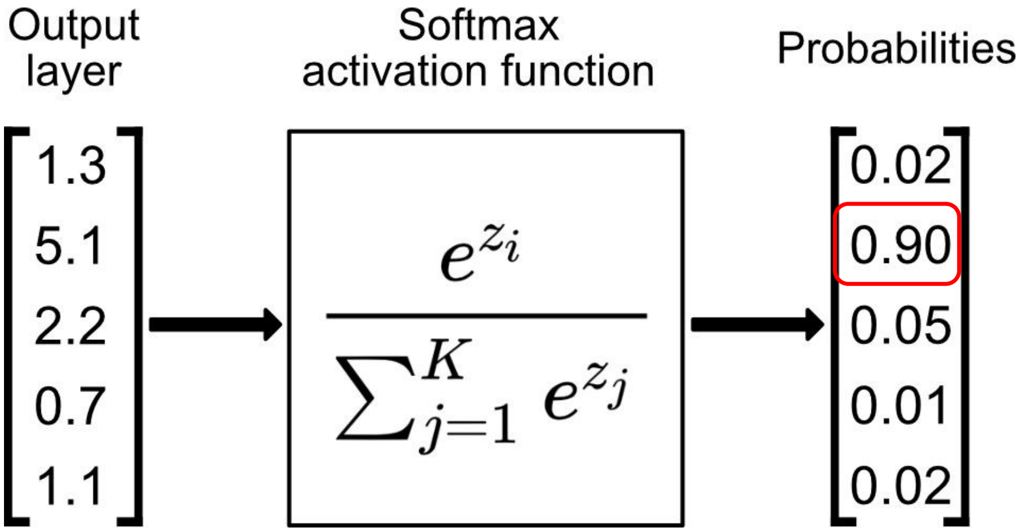
	SROIE [1]	**CORD [2]**
Training	626 files	
Testing	341 files	100 files
Description	● a dataset of scanned receipts. ● 4 keys: ○ Address ○ Company ○ Date ○ Total	● Receipts collected from Indonesian shops and restaurants. ● Noisy and low in quality. ● Key: ○ Total

[1]. Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu,and CV Jawahar. ICDAR 2019 competition on scanned receipt ocr and information extraction. In 2019 International Conference on Document Analysis and Recognition(ICDAR), pages 1516–1520. IEEE, 2019.
[2]. Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing.2019

Baselines

Baselines:

- **Softmax threshold [1]**
  combining both softmax probabilities from OCR and KV models using multiplication (i.e: $p_{final} = p_{OCR} * p_{KV}$
- **Softmax classifier**
  $p_{final} = \mathrm{MLP}([p_{OCR} \mid p_{KV}])$ where MLP is learned
  classifier
- **Temperature Scaling [2]**
  $p_{final} = \max(\mathrm{softmax}(\mathrm{logit}/T))$ where T is learned temperature (on validation set)
- **MC-Dropout [3]**
  Run n=128 times KV predictions to get variance of softmax probabilities
  $p_{final} = 1 - \mathrm{sqrt}(\boldsymbol{\sigma}_i - \max(\boldsymbol{\sigma}_i))$ to normalize variance of i-th sample

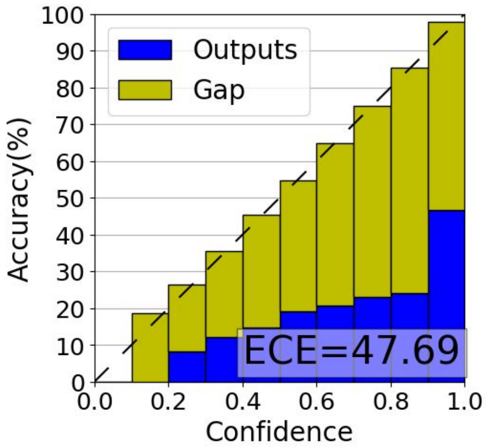


[1] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks.arXiv preprint arXiv:1610.02136,2016
[2] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks.arXiv preprint arXiv:1706.04599, 2017
[3] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning, pages 1050–1059, 2016.

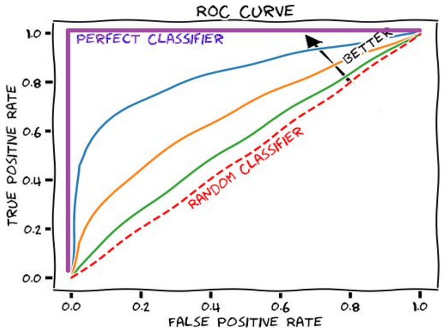
Evaluation metrics

Expected Calibration Error (ECE) [1]

- Compare the confidence score with the actual model accuracy.
- Partitioning predictions into N equally-spaced bins and taking a weighted average of the bins' accuracy and confidence difference
- citation



Area under the ROC Curve (AUC) [2]



- ROC curve (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at various classification thresholds

[1]. Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 29, 2015

[2]. Noam Mor and Lior Wolf. Confidence prediction for lexicon-free ocr. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 218–225.IEEE, 2018

Results

Japanese data

Methods	In-house 1		In-house 2	
	ECE	AUC	ECE	AUC
Softmax threshold	0.1285	68.79	0.5885	53.38
Softmax classifier	0.2810	71.43	0.3945	51.22
MC Dropout	0.3733	66.14	03621	48.20
Temperature scaling	0.1728	64.00	0.5879	58.18
MCP	0.0782	86.32	0.3348	60.12
HYCEDIS	**0.0712**	**90.12**	**0.3019**	**61.90**

Table 3: Performance comparison of baselines and proposed methods on In-house datasets

Results

English data

Methods	SROIE		CORD	
	ECE	AUC	ECE	AUC
Softmax threshold	0.1525	83.75	0.1731	66.91
Softmax classifier	0.1400	85.50	0.3289	54.91
MC Dropout	0.1175	86.90	0.5446	43.52
Temperature scaling	0.1385	84.37	0.3787	74.58
MCP	0.1124	86.40	0.1432	75.12
HYCEDIS	**0.1002**	**88.12**	**0.1259**	**77.45**

Table 2: Performance comparison of baselines and proposed methods on SROIE and CORD datasets

Methods	ECE	AUC
MCP (concatenation)	0.1525	83.75
MCP (bilinear pooling)	0.1175	86.90
MCP (concatenation) + VCAD	0.1385	84.37
MCP (bilinear pooling) + VCAD	**0.1002**	**88.12**

Table 1: Ablation study on SROIE dataset

Conclusion

- Achievements
  - We have presented about our solution: **HYCEDIS**, its motivation, design and current result on the practical datasets.
  - Experiment result shows that our model provides significant improvement compare to baselines, in term of confidence vs accuracy relation, errors detection and recall of output at high accuracy.

- Remaining challenges:
  - Learning from highly unbalanced data
  - Better combination of features
  - Upper limit of Anomaly Detector on In-distribution data

- Future directions
  - Extend confidence model to other applications
  - Support human-in-the-loop processing flow

Q&A

Thanks for your listening!