# Context-based decision may help for interactive learning and domain adaptation

Adrien Chan-Hon-Tong

# Context-based decision may help for interactive learning and domain adaptation

Adrien Chan-Hon-Tong

December 30, 2024

Deep networks trained in a supervised way can achieve impressive performance in the initial distribution while performing poorly in close-but-different distributions. Although there is already a large literature on domain adaptation and interactive learning aiming to deal with such performance drop, the use of few labeled data from target distribution as context has not been widely studied. This technical report points out that using a context-like framework can slightly improve the performance of the standard deep network under moderate domain shift.

## 1 Introduction

In-context learning, which is the ability to learn how to answer a query by using a set of relevant documents, is a major trend in machine learning. From our human perspective, this is natural to combine direct mobilizable knowledge on subjects we are familiar with (in-weight ability), and, the ability to find the answers from external source for more distant subjects (in-context ability). Yet, this learn-to-learn ability has only reached relevant performance with the rise of large language model ([9, 10] and more precisely with transformer-backed [12] masked-auto-encoding [2] methods (MAE) for natural language processing. By learning to fill texts using far correlation thank to attention, MAE learn to use far pieces of text, and, indirectly, context texts.

However, this ability is today mainly used in generative model but not to improve supervised tasks. This technical report focuses on the possibility to add context into a classifier: the goal is to create a model $f(x, X, Y)$ where $x$ is a sample to classify and $X, Y$ is a set of labeled data (empty when relying the supervised part only).

Such setting has been studied for few shot learning where $x$ is classified only though $X, Y$ like [11]. But, the performances reached in few shot learning are significantly lower than the ones of a baseline supervised pipeline dedicated to the related classes. Inversely, this technical report focuses on the question of the possibility to increase the performance of a strong baseline. This way, this paper is closer to domain adaptation [13] but with the goal of being fine-tune-free, or,

to interactive learning[1] like [6] but with the goal of using a global context rather than a tricky encoding like in [6] restricting oracle actions to have very local effects.

## 2 Method

Unsurprisingly, one can implement the desired setting - $f(x, X, Y)$ which classifies $x$ and updates it decision in the light of context data $X, Y$ - by

- using a first head for baseline classification on the top of any encoder $\theta$ i.e. $\alpha(x) = \text{softmax}(w^T\theta(x) + b)$
- using cross-attention in a second head (still at the top of $\theta$) to express the query $x$ as function of the keys $X$ with values $Y$ to get a context-based decision i.e. $\beta(x, X, Y) = \text{softmax}\left(\theta(X)Q^TQ\theta(x)\right)Y$
- using an expert rule to decide how much the context should modify the baseline decision on a sample $x$ - for example, one may rely on the uncertainty of the baseline $\tau(x, X, Y) = \text{sigmoid}\left(\mathcal{H}(\alpha(x)) - \gamma\right)$ where $\mathcal{H}$ is the entropy of the distribution and $\gamma$ a constant
- combining those two decision with relative importance $\tau$ i.e. $f(x, X, Y) = \tau(x, X, Y) \times \alpha(x) + (1 - \tau(x, X, Y)) \times \beta(x, X, Y)$

This way, the model can both classify a sample $x$ without context with $\alpha(x)$, but, given a context, it can also produce a modified decision taking into account the proximity of $x$ and $X$ (known to have label $Y$) - proximity which may generalize better than raw classification decision.

Surprisingly, training directly this architecture performs poorly and raises many questions on what should be the correct implementation (for example, how should one select $X, Y$). Yet, just combining supervised training of $w, b$ (the baseline head) and a contrastive training of $Q$ (the context head) allows to train an efficient model $f(x, X, Y)$.

Precisely, several contrastive losses have been evaluated, and the best one has been found to be

$$l(Q, x_1, ...x_K, y_1, ..., y_K) = \frac{1}{K}\sum_{i=1}^{K}\text{relu}\left(\max_{j\neq i}\theta(x_i)^TQ^TQ(\theta(x_j) - \hat{x}_{y_i}) + \mu\right)$$

where $\mu$ is a constant and $\hat{x}_{y_i}$ is the average of samples with label $y_i$ i.e. $\hat{x}_{y_i} = \frac{1}{|j, y_j = y_i|}\sum_{j, y_j = y_i}\theta(x_j)$.

[1] Interactive learning is a setting where an operator tries to improve a model by acting as an oracle to guide the model toward better prediction

# 3 Experiments

## 3.1 Datasets

The framework presented in this technical report has found to be *not significantly worse* than the baseline when trained and tested on the same distribution i.e. when trained on *CIFAR train* [4] and tested on *CIFAR test*, or when trained on two third of *Oxford PET dataset* [8] and tested on the last third. On such setting, it is not surprising to not improve the baseline: context is not relevant for a network specialized for the problem processed.

In order to measure context influence in transfer, one relevant dataset is *FLAIR1* [3] a remote sensing dataset where testing data are clustered by unknown geographical areas where visual aspects of images tend to be similar compared to intra-region variation which can include variation in vegetation, in building roof material...

Both baseline and baseline+context are trained on *FLAIR1 train* and tested area per area on *FLAIR1 test* while subtracting deterministically around 10 images per class as context (area per area). Those context images are just discarded for the baseline method[2] or used as context in baseline+context method.

Precisely, the dataset, which is originally designed for segmentation, is converted into a classification dataset by keeping for each image, only the main label (discarding any image where the main label covers less than 50% of the pixel of the image) like in [1].

## 3.2 Results

The purpose of this technical report is the result summarized in table 1: both baseline and baseline+context are evaluated on *FLAIR1* as described previously on the top of a standard encoder (here a ConvNext tiny [7]) and context is found to improve the baseline. This result shows that, in this dataset, using a few contextual data can be relevant to deal with the domain shift (here geographical shift). Importantly, the improvement due to the context is consistent over the different areas i.e. the improvement is not just due to a single specific area better processed.

| methods | mean accuracy |
|---|---|
| baseline | 78,93 % |
| baseline+context | 81,12% |

Table 1: Mean accuracy across all testing areas of *FLAIR1 test* dataset for both ConvNext tiny (baseline) and ConvNext tiny with one standard head and one contextual head (as described in section 2). Results are averaged over 5 trials.

[2]Fine-tuning the baseline on the context images without regularization performs poorly. Regularization could have helped like in [5]. Yet, fine-tuning large model can be impossible in some settings as it may requires dedicated hardware.

Despite this technical report should be significantly consolidated on other datasets, with other backbones, with other potential settings, and better compared to the state of the art, it still points out a potential way to improve performance of a deep network after domain shift without modifying the model by just adding few new labeled data of the target domain as context (when the model has been prepared for using a context but this is somehow cost-free in the offered framework).

# References


[1] Marie-Ange Boum, Stéphane Herbin, Pierre Fournier, and Pierre Lassalle. Continual learning in remote sensing: Leveraging foundation models and generative classifiers to mitigate forgetting. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 8535–8540. IEEE, 2024.

[2] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Anatol Garioud, Stéphane Peillet, Eva Bookjans, Sébastien Giordano, and Boris Wattrelos. Flair# 1: semantic segmentation and domain adaptation dataset. *arXiv preprint arXiv:2211.12979*, 2022.

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[5] Gaston Lenczner, Adrien Chan-Hon-Tong, Bertrand Le Saux, Nicola Luminari, and Guy Le Besnerais. Dial: Deep interactive and active learning for semantic segmentation in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3376–3389, 2022.

[6] Gaston Lenczner, Bertrand Le Saux, Nicola Luminari, Adrien Chan Hon Tong, and Guy Le Besnerais. Disir: Deep image segmentation with interactive refinement. *arXiv preprint arXiv:2003.14200*, 2020.

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[8] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[9] Alec Radford. Improving language understanding by generative pre-training. 2018.

[10] Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*, 2022.

[11] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[12] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[13] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.