Feature Embedding by Template Matching as a ResNet Block

Ada Görgün, Yeti Z. Gürbüz and A. Aydın Alatan
{ada.gorgun, yeti, alatan}@metu.edu.tr

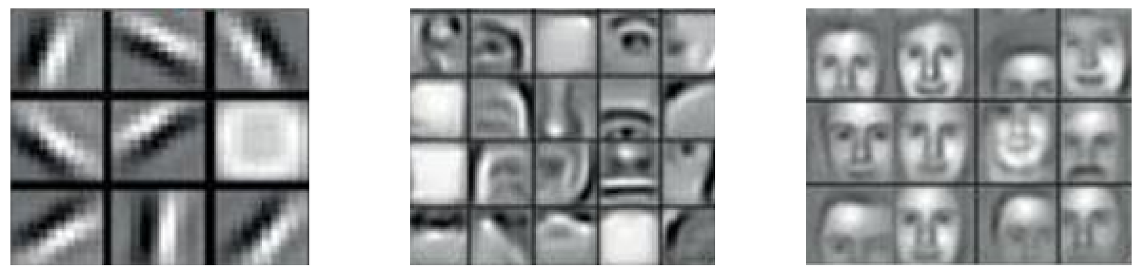
METU · OGAM · BMVC 2022 · 21st - 24th November 2022, London, UK

Contributions

- Reformulating convolution block based local feature embedding as feature assignment through best matching kernel
- Repurposing *soft-max* as a batch-statistics-free replacement of BN-ReLU
- Exploiting mixture of class labels to shape the intermediate features of CNN

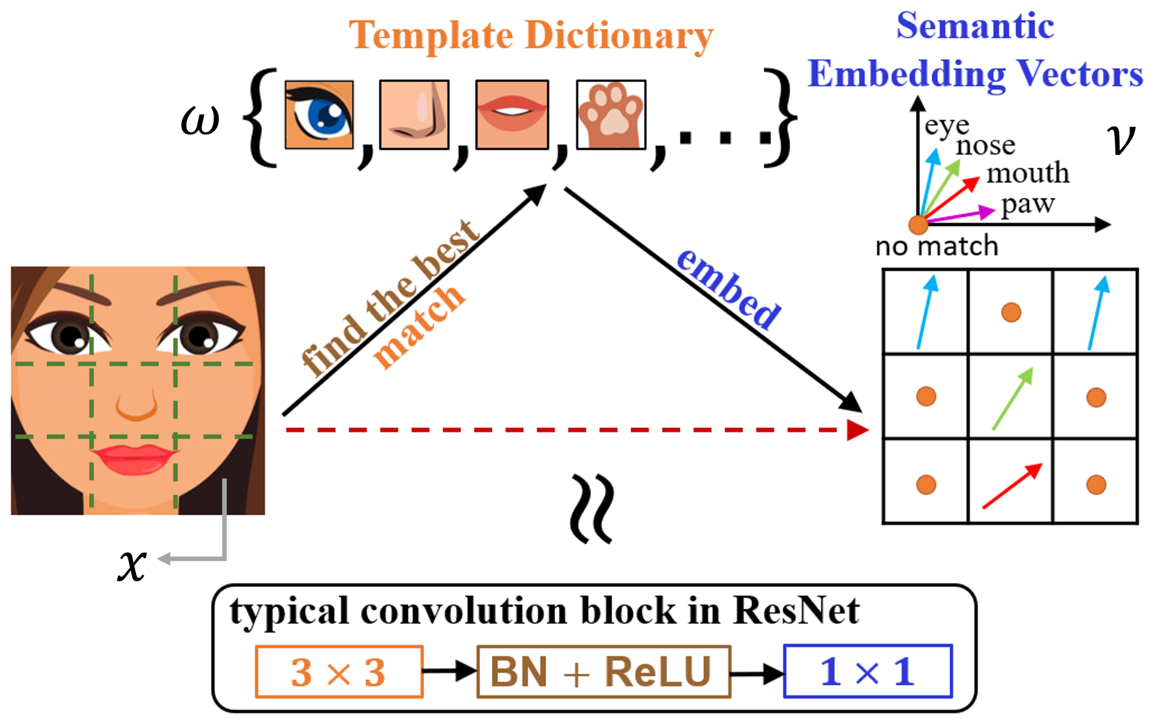
Approach

CNNs: Extracting hierarchical features through stacked convolution blocks whose parameters are learned in **top-down** manner via the feedback from a class-supervised loss function



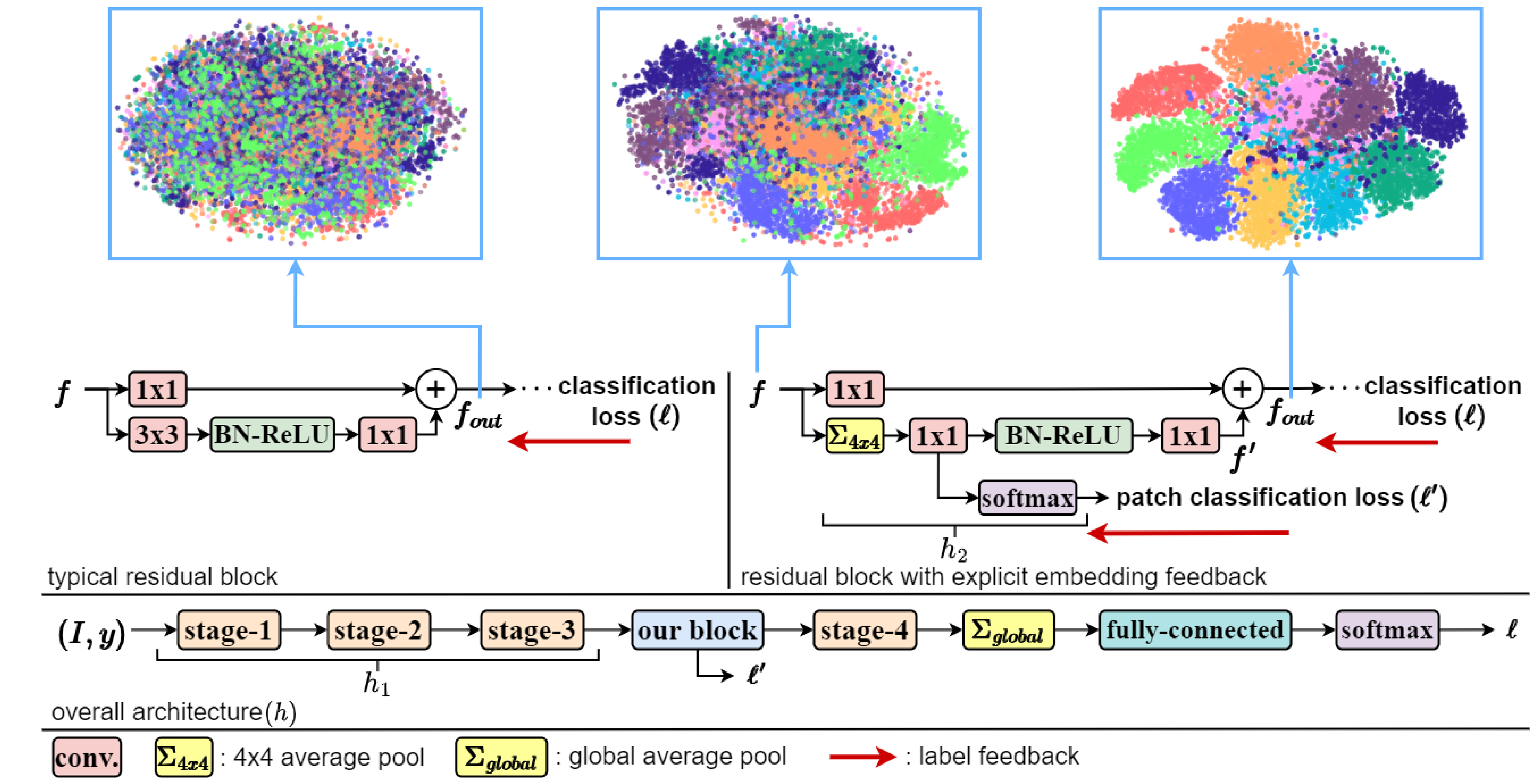
Low level features | Mid level features | High level features

- Differently, we want to use the **mixture of class labels** to synthesize new labels that can **explicitly supervise lower level feature extraction** (*e.g. plane* + *bird* ≈ *wing*).
- We first formulate an optimization problem to analytically express selecting the best matching kernel and assigning its semantic vector.



- We then show its resemblance to a typical ResNet block.

Method Overview



1 We augment the training loss with an auxiliary per patch ($x_{\square}$) classification loss and enable label supervision in lower layers:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{(I,y)\in\mathcal{D}}\Big[(1\text{-}\lambda)\ell(h(I),y) + \frac{1}{wh}\sum_{x\in h_1(I)}\lambda\ell(h_2(x_{\square}),y)\Big]$$

2 With our new block, we yield novel semantic entities as the convex combination of the class features and explicitly shape the feature embeddings according to these semantics.

Theory to Practice

- Given a set of matching kernels ω_k and 3x3 features as $x_{3\times3}$, we define the problem as:

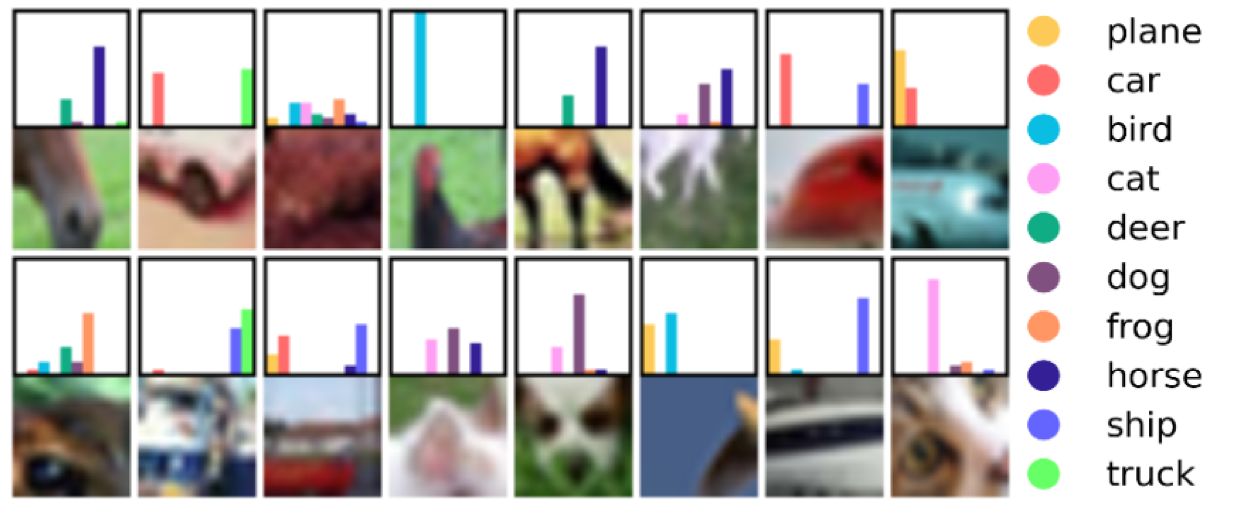
$$p^* = \underset{\substack{p,q\geqslant 0\\ q+\sum_k p_k=1}}{\arg\max}\; q\,\mu + \textstyle\sum_k p_k\,\omega_k^{\intercal} x_{3\times3}$$

- We make the above operation differentiable by smoothing the objective with entropy.
- We represent 3x3 pattern with $x' = \Sigma_k p_k^* \nu_k$.
- Relating BN-ReLU to *soft-max*, we show that 3x3-BN-ReLU-1x1 inherently perform these operations.

Implications

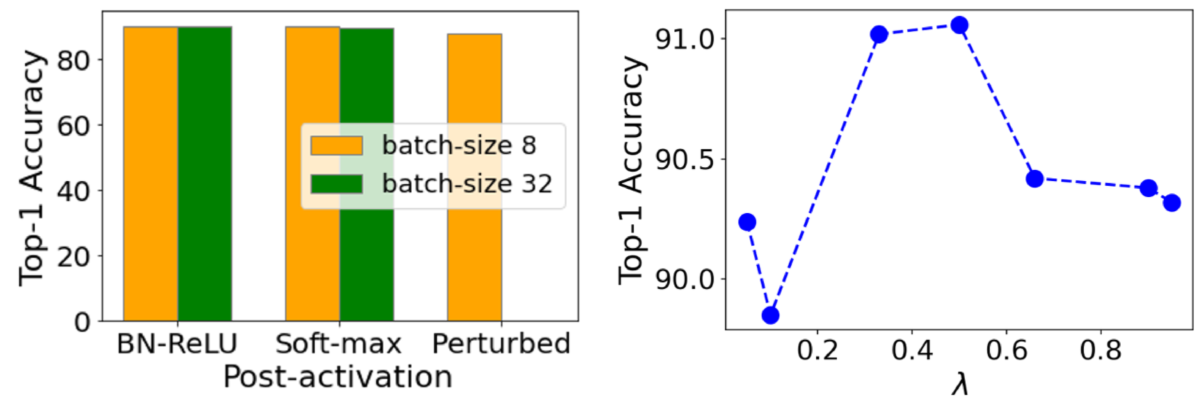
- We observe better clustering effects in which a clear distinction between *animals* and *vehicles* exists.
- Our method enables mixture of class labels and creates fine labels by using existing classes with the help of our embedding vectors which semantically reshape the overall geometry.
- Our formulation mimicks *cross-attention* between $x_{3\times3}$ (*queries*) and ω_k (*keys*) in which the final representation is given as the convex combination of ν_k (*values*).

Analysis of Behaviour



- We generate new semantic entities from the combination of class vectors such as *wing* from *bird* and *plane*, *tire* from *car* and *truck*.
- We observe discriminative patches specific to certain classes.
- We observe generic entities as the mixture of many classes such as *fur* for *animal* classes.

Ablation Studies



(a) (b)

(a) BN-ReLU ≈ *soft-maximizer*

(b) The inclusion of our mechanism through auxiliary loss boosts the performance.

Quantitative Results

Dataset →		Cifar10	Cifar100	Mini-ImageNet
Architecture ↓	Params	top-1 acc.	top-1 acc.	top-1 acc.
RN26	0.96M+257C	89.52	65.94	60.43
RN26-aux.	0.96M+386C	90.57	66.21	60.70
RN26-Ours	0.98M+516C	**91.06**	**66.78**	**61.23**
RN38	1.42M+257C	90.78	68.15	60.72
RN38-Ours	1.44M+516C	**91.36**	**69.01**	**63.83**
WRN16	1.28M+129C	90.52	67.11	60.73
WRN16-Ours	1.30M+388C	**91.10**	**67.36**	**62.92**
DN100	1.20M+535C	92.62	71.65	65.03
DN100-Ours	1.32M+1222C	**92.92**	71.25	68.86
DN100-Ours-C	1.36M+1264C	92.71	**72.14**	**68.93**