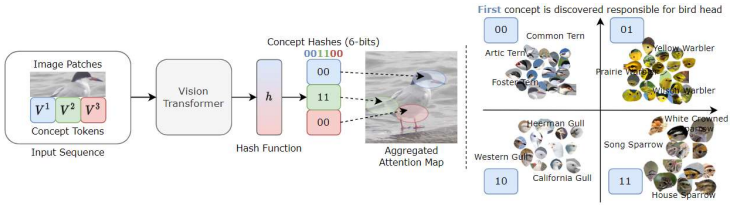
ConceptHash: Interpretable Fine-Grained Hashing via Concept Discovery

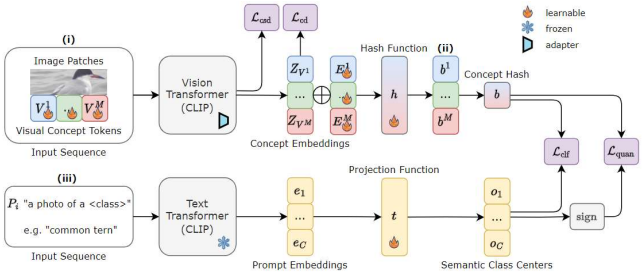
Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, Tao Xiang
University of Surrey

Research questions

- ❑ Can each hash bit/code represent meaningful/interpretable concept?
- ❑ Can we visualize how a bit/code is computed through attention map?
- ❑ Can we utilize language to make interpretable bit/code more semantic?



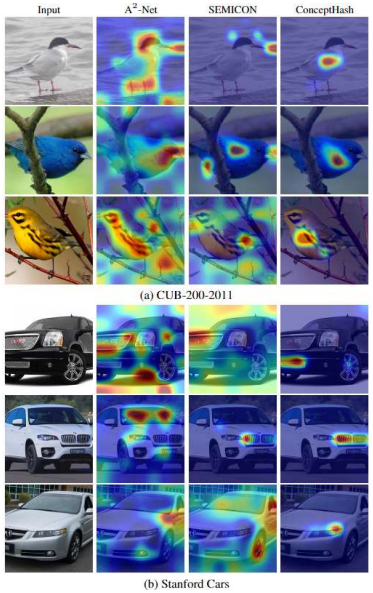
Architecture



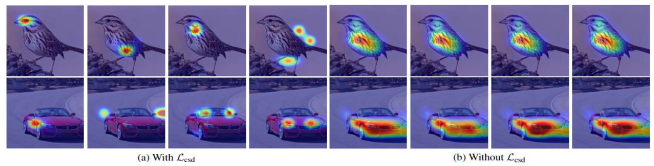
- ❑ Visual concept tokens $V^1 \dots V^M$ are appended to image patches and passed into vision transformer to compute its' concept embedding $Z_{V^1} \dots Z_{V^M}$.
- ❑ Concept embeddings $Z_{V^1} \dots Z_{V^M}$ are shifted by concept specificity embeddings $E^1 \dots E^M$ into common space to compute each sub-code $b^1 \dots b^M$ by a shared concept-generic hash function (a linear projection).
- ❑ All sub-codes $b^1 \dots b^M$ are concatenated to form a full hash code b.

L_{clf} : classification loss with semantic class center $(o_1, \dots, o_C)$
L_{quan} : classification loss with binarized center to minimize quantization error
L_{cd} : concept discrimination loss on concept embeddings
L_{csd} : concept spatial diversity – enhance attention map diversity

Where does a specific bit/code attend?

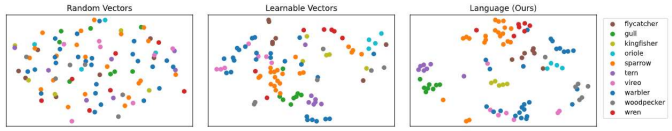


Why concept spatial diversity constraint?



- ❑ Without L_{csd}, the attention always tends to focus on redundant regions.

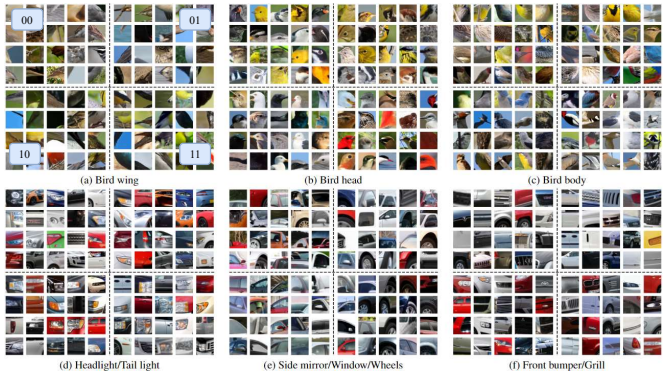
How language information helps?



- ❑ The hash centers $(o_1, \dots, o_C)$ has consistent family structures
- ❑ E.g, ***tern*** and ***gull*** are both seabirds, staying away from non-seabird families

Visualization of toy examples:

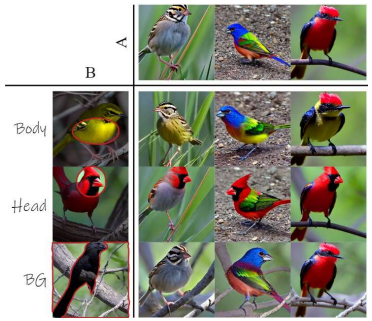
- ❑ Setting: 6-bits hash codes
- ❑ Each 2-bit encodes interpretable concept (e.g., wing, head and body).



Summary

- ❑ We propose ConceptHash, a framework that aim to encode meaningful concept into sub-codes for fine-grained retrieval.
- ❑ We also leverage language information to improve both performance and semantics of hash codes
- ❑ Without manual part labels, ConceptHash can identify meaningful object parts and encode them into corresponding sub-code space.

Trailer (Controllable image generation with parts)



Email:
kamwoh.ng@surrey.ac.uk
kamwoh@gmail.com