

Simple Dialogue System with AUDITED

Yusuf Tas
Piotr Koniusz*

Data61/CSIRO, Australia &
the Australian National University

<http://users.cecs.anu.edu.au/~koniusz>

Abstract

We devise a multimodal conversation system for dialogue utterances composed of text, image or both modalities. We leverage Auxiliary Unsupervised vISual and TExtual Data (AUDITED). To improve the performance of text-based task, we utilize translations of target sentences from English to French to form the assisted supervision. For the image-based task, we employ the DeepFashion dataset in which we seek nearest neighbor images of positive and negative target images of the MMD data. These nearest neighbors form the nearest neighbor embedding providing an external context for target images. We form two methods to create neighbor embedding vectors, namely Neighbor Embedding by Hard Assignment (NEHA) and Neighbor Embedding by Soft Assignment (NESA) which generate context subspaces per target image. Subsequently, these subspaces are learnt by our pipeline as a context for the target data. We also propose a discriminator which switches between the image- and text-based tasks. We show improvements over baselines on the large-scale Multimodal Dialogue Dataset (MMD) and SIMMC.

1 Introduction

Deep learning is popular in many areas *e.g.*, object detection [17], speech recognition [19], image super-resolution [24], text and natural language processing [10], domain adaptation [26, 28, 44], few-shot learning [24, 57, 58, 60], and even arts recognition [28, 59]. Realistic problems such as Visual Question Answering (VQA) are often multimodal. Image Captioning (IC) [54] learns from text and images to generate image captions. VQA [53] answers questions about a video by leveraging the spatio-temporal visual data and the accompanying text. Multimodal conversation systems use text and images used together as chat bots [35], autonomous retail agents [57] and task-specific dialogue systems [50]. Saha *et al.* [57] introduced one of the largest multimodal conversation datasets called Multimodal Dialogue (MMD) dataset, containing over 150K shopper-retail agent dialogues. Figure 1a shows dialogues of shoppers asking about/referring to items or asking for items from a given image. MMD contains the image- and text-based tasks. In the image-based task, the model has to retrieve/rank the correct image from given positive and negative images in response to the multimodal context. The text-based task predicts the agent’s response within the context.

In this paper, we go beyond separate protocols of Saha *et al.* [57] by introducing a discriminator whose role is to learn/predict an appropriate task. As a limited number of utterances contain images, we leverage external visual and textual knowledge via the so-called assisted supervision. Figure 1a shows our pipeline. Our contributions are listed below:

* The corresponding author.

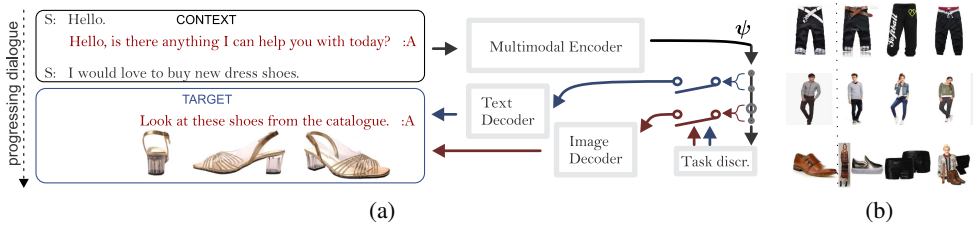


Figure 1: Our pipeline includes the Multimodal Encoder, Text Decoder, Image Decoder (Feature Matching Head) and the Task Discriminator (Fig. 1a). The MMD dataset contains dialogues between shoppers (S) and retail agents (A) which progress in time. Dialogues are split by the sliding window (default protocol) to form the input ($CONTEXT$) fed to the Multimodal Encoder. The output ($TARGET$) may contain text, images, or both modalities, which are imposed via dedicated losses on the Text and/or Image Decoders. The switches indicate that one half of the Context Descriptor ψ may be passed to the Text Decoder and the other half to the Image Decoder depending on the Task Discriminator. The details of Multimodal Encoder, Text Decoder and Image Decoder are shown in Figures 2a, 2b and 3, respectively. Figure 1b shows 3 nearest neighbors (columns 2–4) retrieved (decreasing similarity order) from DeepFashion [60] for query samples (column 1) from the MMD dataset [57]. Feature descriptors were encoded by ResNet-50, the approximate nearest neighbor search was performed by the FAISS library [23]. Such images form an external context for target images.

- i. We propose a novel assisted supervision to create a context for target images and thus implicitly incorporate more images in unsupervised manner into the learning process of image-based task. The DeepFashion dataset [60] is used to search for closest matching images to given positive and negative target images. Through the perspective of sampling the natural manifold of images, we capture context images for target images.
- ii. We design two embeddings for neighbor images: Neighbor Embedding by Hard Assignment (NEHA) and Neighbor Embedding by Soft Assignment (NESA). NEHA retrieves η nearest neighbors for positive/negative target images to encode them into subspace descriptors by SVD. NESA also reweights the contribution of each context image by the membership probability in a GMM-like model [25, 27] spanned on target images.
- iii. For the text-based task, we propose an assisted supervision that uses translation decoders to generate predictions of text in multiple languages to learn a universal representation of conversations by limiting ambiguities of a single language model [51].
- iv. Finally, we introduce a discriminator whose role is to combine image- and text-based tasks by learning to predict an appropriate task in response given the multimodal context.

The above strategy of leveraging unsupervised data can be seen as capturing the variance of linguistic and visual data to help the network capture how each utterance may vary.

2 Related Work

Below we describe popular dialogue systems, and detail the Multimodal Hierarchical Encoder Decoder (M-HRED) [57] on which we build.

Conversation Systems. Early conversation systems [3, 5] use scripts and subtitles for retrieval of responses in a dialogue. Ritter *et al.* [66] uses generative probabilistic models for conversations on blogging websites. VQA approaches [4, 53] answer questions about

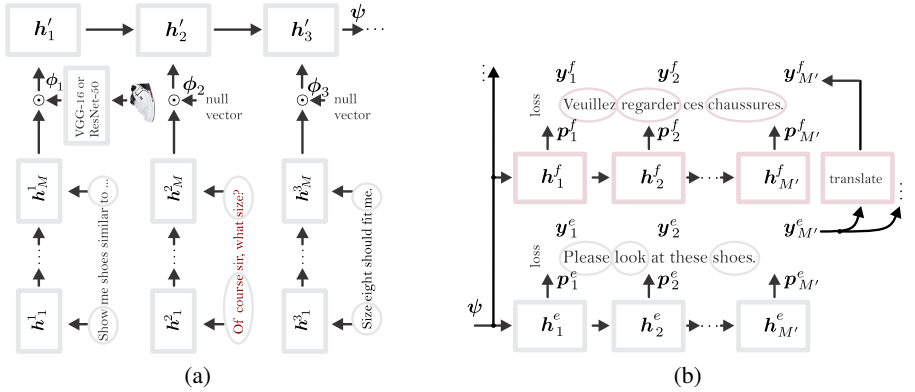


Figure 2: In Multimodal Encoder, shown in Fig. 2a, the text is processed by a first-level GRU while images are encoded by ResNet-50 to obtain compact embeddings. We concatenate text and avg-pooled image representations (if image is not present, we use a null vector) by \odot into utterance descriptors ϕ_1, \dots, ϕ_3 and process them with a second-level GRU to produce the Context Descriptor ψ , which we pass it to the Text Decoder with the assisted supervision in Figure 2b. The text is translated from English (ground truth) into French (and other languages). The losses (per language) encourage the network to absorb syntactic differences which implicitly helps capture the true dynamics of the dialogue better. Standard Text Encoder [37] consists of the gray blocks while pink blocks form our assisted supervision.

images. Approaches [10, 29] tackle visual dialogs about individual images. Approach [47] focuses on the visual dialog navigation. Bhattacharya *et al.* [0] retrieves images via textual queries. FashionIQ [20] is concerned with the NLP-based image retrieval.

Recent dialogue systems use an RNN encoder-decoder [41, 45]. Hierarchical Recurrent Encoder-Decoder [58] uses a two-level RNN to create a context-aware conversation system. Approach [37] predicts answers of a shopping assistant from natural conversations of the large scale MMD dataset, which we use. Below, we describe and build on models [39, 40].

Multimodal-Hierarchical Encoder Decoder. M-HRED [37] is an extension of Hierarchical Recurrent Encoder Decoder (HRED) models [39, 40]. HRED consists of two different levels of Recurrent Neural Network (RNN) [62] combined together, which represent an encoder which captures the so-called word and sentence context, respectively. The first RNN in HRED model learns to generate the next word in a given sentence by using the word context. The second RNN takes the final representation of a given sentence to generate the representation of next sentence by using the sentence context. An RNN decoder receives a sentence-level representation to decode it and generate a full sentence. Moreover, M-HRED and HRED use the interconnected encoder and decoder but M-HRED also uses images.

Multimodal Encoder (ME). ME receives a sequence of N utterances (so-called context) to produce the Context Descriptor via GRU [9]. An utterance contains a sentence, image or both modalities. Images are encoded by VGG-16 [44] (4096 ch. of the last FC layer [37]).

Multimodal Utterance Encoder (MUE) in Fig. 2a consists of two levels of GRU [9]. The first-level GRU (bottom) contains hidden states h_1^n, \dots, h_M^n , where M is the maximum number of input words per utterance, each word is one-hot encoded with a discrete vocabulary of size $V = 7457$, $n = 1, \dots, N$ and N is the context size *e.g.*, $N = 3$ utterances. The first-level GRU and ResNet-50 encode words and images, respectively. The last state and the output of ResNet-50 are concatenated by \odot into ϕ and padded with zeros if image or text is missing. Encoded

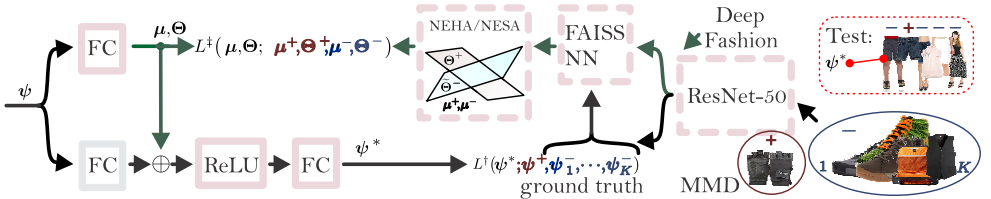


Figure 3: Our Image Decoder a.k.a. Feature Matching Head consists of the main stream (FC→ReLU→FC) whose role is to take Context Descriptors Ψ and produce visual features Ψ^* that are combined with loss L^\dagger in Eq. (1). The traditional head (older method) contains one FC layer (gray block). For each ground truth target positive and negative image descriptors Ψ^+ and Ψ^-, \dots, Ψ^K from the MMD dataset (encoded by ResNet-50), we find η and ηK approximate nearest neighbor image descriptor from DeepFashion [50] with the FAISS library [23]. Then we create positive and negative mean descriptors μ^+ and μ^- as well as subspaces Θ^+ and Θ^- with NEHA or NESA step. They capture the mean and variability of positive and negative images. The role of another FC layer is to learn positive visual context representations (μ, Θ) via the assisted supervision loss L^\ddagger in Eq. (2) which attracts (μ, Θ) towards (μ^+, Θ^+) and repels it from (μ^-, Θ^-) . Finally, (μ, Θ) are combined with the main stream via a residual link (operator \oplus). Blocks with dashed borders/losses are not used during testing. During testing, Ψ^* is matched against test images of an utterance. In the above example, the correct ground truth is ranked as second (R@1 fails but R@2 succeeds).

utterances are passed to the Context Encoder (CE), a second-level GRU, with hidden states $h'_1, \dots, h'_{N'}$ shown in Fig. 2a (top) to obtain a Context Descriptor Ψ per context. Fig. 1a shows examples of context and target utterances. We use encoder networks from the M-HRED model [57] (results in the same testbed, based on ResNet-50).

Multimodal Decoder (MD). MD receives the Context Descriptor Ψ from CE. For the text-based task, the target is a sentence. A GRU decoder [69] with hidden states $h^e_1, \dots, h^e_{M'}$ generates the target sentence word-by-word, starting with the start-of-sentence and ending with end-of-sentence token. Given the target ground truth sentence with one-hot representation of words \mathbf{y}^e and the final output predictions \mathbf{p}^e from the model (e indicates English), the combined Multimodal Encoder Decoder is trained via the cross-entropy loss. At the test time, the quality of generated utterance is evaluated against target ground truth sentences via so-called BLEU and NIST metrics [67]. Figure 2b shows our extended Text Decoder (pink plus gray blocks) and the baseline Text Decoder (gray block) [67]. The image-based task is the ranking-based task. Given a positive target image, and K negative images, Context Descriptor Ψ is ranked against these positive and negative images at the test time. During training, M-HRED uses the cosine similarity and the hinge loss:

$$L^\dagger(\Psi^*, \Psi^+, \Psi^-, \dots, \Psi^K) = \max\left(0, 1 - \Psi^{*T} \left(\Psi^+ - \frac{1}{K} \sum_{k=1}^K \Psi_k^- \right)\right), \quad (1)$$

where $\Psi^* \in \mathbb{R}^{4096}$ is a feature vector obtained by passing the Context Descriptor $\Psi \in \mathbb{R}^{1024}$ from CE via an FC layer, and $\Psi^+ \in \mathbb{R}^{4096}$ and $\Psi^- \in \mathbb{R}^{4096}$ correspond to image descriptors (VGG-16) for the positive and negative ground truth images, resp. The Hinge loss encourages Ψ^* to be close to Ψ^+ and away from Ψ^- . L is minimized w.r.t. network parameters.

Self-supervised Learning. Pretext tasks such as sampling and predicting patch locations (left, right, top left, top right), rotations (0° , 90° , 180° , 270°) or other transformations are popular in self-supervision [13, 15, 16, 56, 58]. Note self-supervision by mutual information estimation [21], egomotion prediction [2], and multi-task self-supervised learning [12]. One

Input: $\eta' \leq \eta, K, L$
 $\Psi_1^+, \Psi^- \equiv \{\Psi_1^+, \dots, \Psi_K^-\} \leftarrow$ ground truth positive and negative target descriptors from MMD,
 $\{\Psi_1^-, \dots, \Psi_L^-\} \leftarrow$ unsupervised feature descriptors from DeepFashion [60].

- 1: $(\Psi_1^+, \dots, \Psi_{\eta'}^+) = \text{FAISS_NN}(\Psi^+, \eta; \{\Psi_1^-, \dots, \Psi_L^-\})$
- 2: **for** $n=1, \dots, \eta$:
- 3: $\Psi_n^+ \leftarrow s^+(\Psi_n^+, \Psi^+, \Psi^-) \cdot \Psi_n^+$
- 4: $\mu^+ = \frac{1}{\eta} \sum_{n=1}^{\eta} \Psi_n^+$
- 5: $(\Theta^+, \lambda^+) = \text{SVD}(\Psi_1^+ - \mu^+, \dots, \Psi_{\eta'}^+ - \mu^+; \eta')$
- 6: **for** $k=1, \dots, K$:
- 7: $(\Psi_{1k}^-, \dots, \Psi_{\eta k}^-) =$
- 8: $\text{FAISS_NN}(\Psi_k^-, \eta; \{\Psi_1^-, \dots, \Psi_L^-\})$
- 9: **for** $k=1, \dots, K$:
- 10: **for** $n=1, \dots, \eta$:
- 11: $\Psi_{nk}^- \leftarrow s^+(\Psi_{nk}^-, \Psi^+, \Psi^-) \cdot \Psi_{nk}^-$
- 12: $\mu^- = \frac{1}{\eta K} \sum_{n=1}^{\eta} \sum_{k=1}^K \Psi_{nk}^-$
- 13: $(\Theta^-, \lambda^-) = \text{SVD}(\Psi_1^- - \mu^-, \dots, \Psi_{\eta K}^- - \mu^-; \eta')$

Output: (μ^+, Θ^+) and (μ^-, Θ^-)

Algorithm 1: Neighbor Embedding by Hard Assignment (black color). Neighbor Embedding by Soft Assignment (black/blue colors).

may use Bag-of-Words on hand-crafted descriptors for an alignment task [49, 60], or form positive and negative sampling for a contrastive learning strategy [61, 62, 63]. GAN-based pipelines [18, 42, 43] also perform self-supervision by generator-discriminator competition.

Motivation from Cognitive Psychology. For the text-based task, we use a translating network [64] and decoders to predict target responses in several languages. This limits the quantization noise resulting from the single language syntactic thus helping capture universal concepts better. Cognitive psychology notes that multilingual babies exhibit better attention and conflict management, and adjust to new rules quicker than monolingual babies [65].

For the image-related task, we retrieve the η and ηK nearest neighbors from the DeepFashion [60] dataset for positive and negative target images to form subspace descriptors which represent the learning context of target images, and form the manifold of fashion images. From the psychological point of view, our approach is motivated by knowledge transfer, which is ‘the dependency of human conduct, learning or performance on prior experience’ a.k.a. ‘transfer of particle’ [66]. Notice that pre-training our visual task on the DeepFashion is impossible as DeepFashion dataset is not organised in the form of dialogue.

In conclusion, providing multiple translations and multiple positive images (subspaces are second-order statistics) helps our pipeline capture better the innate variance of data.

3 Our Approach

Notations. Bold lowercase symbols are vectors e.g., μ, ϕ, ψ . Regular lowercase/uppercase symbols are scalars e.g., η, K, N . Bold uppercase symbols are matrices or sets of parameters e.g., Θ . Symbols \odot and \oplus are the vector concatenation & summation (residual link).

Pipeline. Our pipeline in Figure 1a follows the baseline model [67] in that we use the Multimodal Encoder, Text Decoder and Image Decoder (Feature Matching Head only). The Multimodal Encoder receives the context, a collection of $N=3$ utterances which are snippets of dialogues between a shopper and a retail agent obtained by a sliding window, a standard

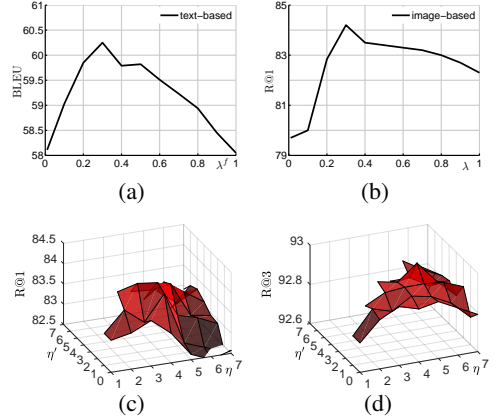


Fig. 4a & 4b are cross-validation results w.r.t. λ^f and λ for the text- and image-based tasks. Fig. 4c & 4d are cross-val. results (R@1 & R@3 metric) w.r.t. η' and η for the image-based task. If $\eta' = 0$, we use μ^- and μ^+ only.

protocol on the MMD dataset of retail dialogues. The context window may contain text, image, or both modalities. The Multimodal Encoder takes N utterances, based on a discrete vocabulary of size V , and ResNet-50 encoded images to produce the Context Descriptor which is fed to the Text Decoder and Image Decoder, whose roles are to predict a target ground truth text responses (within discrete vocabulary space) and/or generate ResNet-50 image features to retrieve a visual recommendation from the MMD (or SIMMC) dataset (also encoded with ResNet-50). As the baseline model [57] is formulated as two separate tasks, it requires ground truth test labels about the type of output task to perform. In contrast, we introduce the Task Discriminator (the pink box in Figure 1a which resolves this issue. To improve predictions, our Text and Image Decoders use the assisted supervision by leveraging the knowledge from the DeepFashion [50] dataset and the translation model [54] in an unsupervised way. Section 2 details the Multimodal Encoder. Below we detail our decoders.

Image-based Task. Figure 3 shows our Feature Matching Head (Image Decoder). The image-based task finds the closest match between a predicted image descriptor and one positive and K negative ground truth descriptor candidates per target utterance. The image-based task uses two losses, the standard loss L^\dagger given by Eq. (1) and our assisted supervision loss:

$$L^\ddagger(\boldsymbol{\mu}, \Theta; \boldsymbol{\mu}^+, \Theta^+, \boldsymbol{\mu}^-, \Theta^-) = \max\left(0, 1 - \boldsymbol{\mu}^T(\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-) - \sum_{n=1}^{\eta'} \mathbf{u}_n^T(\mathbf{u}_n^+ - \mathbf{u}_n^-)\right), \quad (2)$$

where $\boldsymbol{\psi}^* \in \mathbb{R}^D$ is a feature vector of size $D=2048$ obtained by passing the Context Descriptor $\boldsymbol{\psi} \in \mathbb{R}^{1024}$ from CE via an FC layer. Moreover, $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\Theta \equiv [\mathbf{u}_1, \dots, \mathbf{u}_{\eta'}] \in \mathbb{R}^{D \times \eta'}$ are the context feature vectors generated by an FC layer, indicated in Figure 3, which are encouraged by a Hinge loss to approach the mean $\boldsymbol{\mu}^+ \in \mathbb{R}^D$ and eigenvectors $\Theta^+ \equiv [\mathbf{u}_1^+, \dots, \mathbf{u}_{\eta'}^+] \in \mathbb{R}^{D \times \eta'}$ and stay repelled from the mean $\boldsymbol{\mu}^- \in \mathbb{R}^D$ and eigenvectors $\Theta^- \equiv [\mathbf{u}_1^-, \dots, \mathbf{u}_{\eta'}^-] \in \mathbb{R}^{D \times \eta'}$. Visual Feature Descriptors (VFD) $(\boldsymbol{\mu}^+, \Theta^+)$ and $(\boldsymbol{\mu}^-, \Theta^-)$ represent the positive and negative context for the ground truth positive and negative target descriptors $\boldsymbol{\psi}^+ \in \mathbb{R}^D$ and $\boldsymbol{\psi}_1^-, \dots, \boldsymbol{\psi}_K^- \in \mathbb{R}^D$ obtained from ResNet-50. Below we explain Neighbor Embedding by Hard Assignment (NEHA) and Neighbor Embedding by Soft Assignment (NESA) which produce VFDs.

NEHA is obtained by applying SVD to η and ηK nearest neighbors $\boldsymbol{\psi}'_1, \dots, \boldsymbol{\psi}'_{\eta'} \in \mathbb{R}^D$ and $\boldsymbol{\psi}'_{11}, \dots, \boldsymbol{\psi}'_{\eta K} \in \mathbb{R}^D$ found among images of DeepFashion [50] dataset encoded by ResNet-50, represented by L feature descriptors $\boldsymbol{\psi}'_1, \dots, \boldsymbol{\psi}'_L$. The search is performed by FAISS [23], an extremely efficient approximate nearest neighbor search library, by searching feature descriptors of DeepFashion against the ground truth positive/negative target descriptors $\boldsymbol{\psi}^+$ and $\boldsymbol{\psi}_1^-, \dots, \boldsymbol{\psi}_K^-$ from the MMD dataset, respectively. Figure 1b shows the quality of matching images from DeepFashion against ground truth images from MMD.

Algorithm 1 shows steps of NEHA. FAISS_NN($\boldsymbol{\psi}, \eta; \{\boldsymbol{\psi}'_1, \dots, \boldsymbol{\psi}'_L\}$) denotes the FAISS search which retrieves η approximate nearest neighbors of $\boldsymbol{\psi}$ from $\{\boldsymbol{\psi}'_1, \dots, \boldsymbol{\psi}'_L\}$. Moreover, SVD($\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{\eta'}; \eta'$) returns $\eta' \leq \eta$ leading eigenvectors and eigenvalues $(\Theta, \boldsymbol{\lambda})$. We note that NEHA does not take into account the effect of decreasing similarity between the ground truth positive/negative target descriptors and searched feature descriptors of DeepFashion as one progresses over consecutive $1, \dots, \eta$ nearest neighbors. Thus, Visual Feature Descriptors $(\boldsymbol{\mu}^+, \Theta^+)$ and $(\boldsymbol{\mu}^-, \Theta^-)$ may provide gradually worsening visual context for target descriptors of MMD. To his end, we introduce an improved strategy below.

NESA follows NEHA but it uses reweighting by so-called Soft Assignment applied prior to SVD steps. We use the two weighting functions for positive $\boldsymbol{\psi}'^+$ and negative $\boldsymbol{\psi}'^-$:

$$s^+(\boldsymbol{\psi}', \boldsymbol{\psi}^+, \boldsymbol{\Psi}^-) = \frac{1}{\tau(\boldsymbol{\psi}', \boldsymbol{\psi}^+, \boldsymbol{\Psi}^-)} e^{-\frac{\|\boldsymbol{\psi}' - \boldsymbol{\psi}^+\|_2^2}{2\sigma^2}} \quad \text{and} \quad s^-(\boldsymbol{\psi}', \boldsymbol{\psi}^+, \boldsymbol{\Psi}^-) = \frac{1}{\tau(\boldsymbol{\psi}', \boldsymbol{\psi}^+, \boldsymbol{\Psi}^-)} \max_{k=1, \dots, K} e^{-\frac{\|\boldsymbol{\psi}' - \boldsymbol{\psi}_k^-\|_2^2}{2\sigma^2}}, \quad (3)$$

where $\Psi^- \equiv \{\psi_k^-\}_{k=1}^K$. Expression $\tau(\psi', \psi^+, \Psi^-)$ given below normalizes probability partitions in Eq. (3):

$$\tau(\psi', \psi^+, \Psi^-) = e^{-\frac{\|\psi' - \psi^+\|_2^2}{2\sigma^2}} + \sum_{k=1}^K e^{-\frac{\|\psi' - \psi_k^-\|_2^2}{2\sigma^2}}, \quad (4)$$

while σ determines the steepness of likelihood partitions. The Soft Assignment step is performed by reweighting $\psi'_1, \dots, \psi'_\eta$ by $s^+(\psi'_1, \cdot, \cdot), \dots, s^+(\psi'_\eta, \cdot, \cdot)$ and $\psi'^-, \dots, \psi'^-_{\eta K}$ by $s^-(\psi'^-, \cdot, \cdot), \dots, s^-(\psi'^-_{\eta K}, \cdot, \cdot)$. Algorithm 1 (with steps highlighted in blue) realizes NESA. NEHA and NESA use combination losses: $L^\dagger(\psi^*; \psi^+, \Psi^-) + \lambda L^\ddagger(\mu, \Theta; \mu^+, \Theta^+, \mu^-, \Theta^-)$.

NNO. Nearest Neighbor Only (NNO) strategy is given for completeness. NNO simply encourages the standard head with one FC layer (gray block in Figure 3) to get closer not only to target samples of MMD but also to the positive approximate nearest neighbor(s) retrieved from DeepFashion. NNO uses combined losses: $L^\dagger(\psi^*; \psi^+, \Psi^-) + \lambda \|\psi^* - \frac{1}{\eta} \sum_{n=1}^{\eta} \psi_n^{*+}\|_2^2$.

Text-based Task. Figure 2b shows that apart from the standard GRU decoder (gray blocks), we use translating network [64] to translate [64] ground truth sentences from English into French, German and Russian, with one GRU per language. For English, we have a GRU with hidden states $h_1^e, \dots, h_{M'}^e$, output predictions $p_1^e, \dots, p_{M'}^e$ and ground truth one-hot vectors $y_1^e, \dots, y_{M'}^e$. By analogy, we use analogous streams for other languages. Moreover, every $p_m^e \in \mathbb{R}^{7457}$ is an output of an FC layer connected to the corresponding hidden state $h_m^e \in \mathbb{R}^{1024}$. The FC layer translates hidden states into word activation vectors corresponding to a 7457 dimensional dictionary. Note that for every language, the dictionary size differs. For French, we have 9519 words after considering words with the occurrence of at least $5 \times$ given the training data. Each sentence starts with the start-of-sentence token, ends with the end-of-sentence token and is padded to the maximum sentence length of $M' = 20$ with the pad-sentence token. Pink blocks realize the assisted supervision for the text-based task. At the test time, they are removed. The final loss for the Text Decoder becomes:

$$L\left(\{(p_m^e, y_m^e), (p_m^f, y_m^f), \dots\}_{m=1}^{M'}; \lambda^f, \dots\right) = \sum_{m=1}^{M'} y_m^{eT} \log(p_m^e) + \lambda^f y_m^{fT} \log(p_m^f) + \dots, \quad (5)$$

where λ^f is the relevance constant of the French translation task, and λ^g and λ^r are relevance constants for German/Russian but we omit them from notations for brevity.

Task Discriminator (TD). The Context Descriptor¹ $\psi \in \mathbb{R}^{2048}$ is passed to an FC layer (2048×3 size) following the cross-entropy loss with task labels: *text-based*, *image-based* and *text+image-based*. During training, we can access such labels. Thus, during testing, we can go beyond separate protocols of the baseline model [63]. Figure 1a shows TD and the switches that pass relevant halves of ψ to subsequent modules.

4 Experiments

Datasets. Our experiments are conducted on the MMD datasets [65] *v1* and *v2* containing ~ 150000 dialogues and the SIMMC dataset [63], with $\sim 13K$ human human dialogues and $\sim 169K$ utterances. The assisted supervision for the text-based task is achieved via model [64] trained on the WMT [8] and Paracrawl [9] datasets containing $\sim 150M$ sentence pairs. The assisted supervision for the image-based task is achieved by retrieving relevant feature descriptors from the DeepFashion dataset [60] ($\sim 0.8M$ images).

¹For evaluations where we use TD, the Context Descriptor ψ is in fact 2048 dimensional as its both halves are dedicated to text- and image-based tasks, respectively. For individual tasks, ψ are 1024 dimensional.












		BLEU	NIST
MMD v1	T-HRED 	14.58	2.61
	M-HRED 	20.42	3.09
	M-HRED+attention 	19.58	2.46
	M-HRED+attention+KB 	-	-
	(Ours) Pre-training (French)	24.35	4.12
	(Ours) Assisted sup. (French)	26.21	4.45
MMD v2	T-HRED 	35.9	5.14
	M-HRED 	56.67	7.51
	M-HRED+attention 	50.20	6.64
	M-HRED+attention+KB 	46.36	-
	(Ours) Augmentation (random deletion)	56.83	7.55
	(Ours) Augmentation (sentence compr. 	57.65	7.62
	(Ours) Augmentation (back translation 	59.06	7.96
	(Ours) Pre-training (on SIMMC dataset 	58.91	7.95
	(Ours) Training on MMD+SIMMC	59.03	7.98
	(Ours) Pre-training (French)	58.78	7.91
	(Ours) Assisted sup. (French)	60.12	8.11
	(Ours) Assisted sup. (French+German)	60.51	8.17
	(Ours) Assisted sup. (French+German+Russian)	60.75	8.22
Tran.	(Ours) Pre-training (French)	60.88	9.28
	(Ours) Assisted sup. (French)	64.47	11.18
	(Ours) Assisted sup. (French+German)	65.54	12.41
	(Ours) Assisted sup. (French+German+Russian)	66.19	12.89

Table 1: Text-based task (MMD v1 & v2). T-HRED / M-HRED are text-only HRED / Multimodal HRED. *Tran.*: transformer backb. [43].

MMD dataset [57] contains 105439 train, 22595 validation and 22595 test dialogues, each with ~ 40 shopper-retailer utterances containing a sentence, images or both modalities. We used train, validation and test splits to train, select hyperparameters and report final results, respectively. MMD v2 does not contain additional image descriptions from the agent.

SIMMC dataset [53] has ~ 13 K human-human dialogs and ~ 169 K utterances, it uses a multimodal Wizard-of-Oz (WoZ) setup, on two shopping domains, furniture (grounded in a shared virtual environment) and fashion (grounded in an evolving set of images).

Settings. Following Saha *et al.* [57], we perform the text- and image-based tasks for which we use the same hidden unit size, text encoding size and the learning rate as M-HRED [57]. For our combined task (TD module), the hidden unit size is doubled (Section 3). For the text- and image-based tasks, we report BLEU/NIST [57] and Recall at top- l cut-off (R@ l).

Results. Below we start with cross-validation of key hyperparameters followed by presenting our main results for text-, image- and mixed (text+images) tasks.

Cross-validation of λ^f and λ . For joint training of French auxiliary decoder with the base English decoder, we cross-validated $\lambda^f \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$ on the validation set (see Figure 4a). We fixed $\lambda^f = 0.3$ throughout experiments as this value yielded the highest score of **60.25%** (BLEU) on the MMD v2 validation split. If we use two auxiliary decoders *e.g.*, French and German, we set $\lambda^f = \lambda^g = 0.15$. For three auxiliary decoders, we set $\lambda^f = \lambda^g = \lambda^r = 0.1$. For joint training of the main stream (FC \rightarrow ReLU \rightarrow FC) and the assisted supervision stream in Feature Matching Head from Figure 3, we set $\lambda = 0.5$ following cross-validation on the validation set given NEHA, shown Figure 4b.

Image-based Task. Firstly, we evaluate the baseline M-HRED+attention with ResNet-50 in place of VGG-16, and we note that the results are within $\pm 0.3\%$ of results given the original M-HRED+attention with VGG-16. Table 2 shows that using the assisted supervision via the NNO strategy with one nearest neighbor ($\eta = 1$) improves results over the baseline M-HRED+attention by $\sim 3.6\%$ and $\sim 4.5\%$ (R@1) given versions v1 and v2 of the MMD







		R@1	R@2	R@3
MMD v1	T-HRED 	46.0	64.0	75.0
	M-HRED 	72.0	86.0	92.0
	M-HRED+attention 	79.0	88.0	93.0
	(Ours) NNO $\eta = 1$	82.6	88.8	93.2
	(Ours) NNO $\eta = 2^*$	83.0	88.9	93.2
	(Ours) NEHA $\eta = 4^*$	84.5	89.7	93.6
	(Ours) NESA $\eta = 4^*$	85.3	90.3	94.0
MMD v2	T-HRED 	44.0	60.0	72.0
	M-HRED 	69.0	85.0	90.0
	M-HRED+attention 	78.0	87.0	92.3
	(Ours) NNO $\eta = 1$	82.5	88.6	92.8
	(Ours) NNO $\eta = 2^*$	83.1	88.8	92.9
	(Ours) NEHA $\eta = 4^*$	84.5	89.5	93.2
	(Ours) NESA $\eta = 4^*$	85.2	90.1	93.7

Table 2: Image-based task (MMD v1 & v2) for one positive and $K = 5$ negative target images. T-HRED is HRED with context images ignored in training. M-HRED is the Multimodal HRED. See Recall at top-1, 2 and 3, ‘*’ is the optimal η .

	BLEU	R@1
M-HRED	52.17	M-HRED+att. 75.05
Assisted sup. (French)	55.29	NEHA 81.50
(French+German+Russian)	56.11	NESA 82.43

Table 3: Mixed task (MMD v2) with the Task Discr., assisted supervision (text and images).

	User 1	User 2	User 3	mean
clarity	61.6	58.4	64.2	61.4
compactness	52.0	52.8	54.6	53.1
helpfulness	62.0	60.2	63.0	61.7

Table 4: User study on the mixed task (MMD dataset v2). Our approach vs. M-HRED.

Component (or method)	runtime (h)
T-HRED / M-HRED	15 / 15
Pre-training (Fr) (+fine-tuning En)	16 + 6
Augmentation (back translation) + transl.	15 + 40
Assisted sup. (Fr / Fr+Ge / Fr+Ge+Ru)	20 / 29 / 38
Translator [4] (En → Fr / Ge / Ru)	20 / 20 / 20
Evaluating BLEU & NIST	0.5

Table 5: Runtimes: text task (MMD v2).

	BLEU	R@1	R@5	R@10
HRE (SIMMC)	0.079	16.3	33.1	41.7
Ours F+R+G	0.102	n/a	n/a	n/a
Ours+Trans. F+R+G	0.187	n/a	n/a	n/a
Ours NEHA	n/a	17.3	33.7	42.2
Ours NESA	n/a	20.1	35.5	43.1

Table 6: SIMMC-Fashion (Task 2). Response Generation. *F+R+G* are French, Russian and German auxiliary tasks. *Tran.* is the transformer backbone [43].

Component (or method)	runtime (h)
T-HRED / M-HRED	15 / 15
NNO	16
NEHA / NESA	18 / 19
ResNet-50 features (MMD+DeepFashion)	6
FAISS search [4] (+SVD)	1.5 (+2)
Evaluating R@1	0.1
T-HRED / M-HRED (text+image)	30 / 30
Mixed task (text+image+task discr.)	40

Table 7: Runtimes: image-based task (vari-ous comp.) and the mixed task (MMD v2).

dataset. Choosing the optimal number of nearest neighbors for NNO ($\eta=2$) improves results by further 0.4% (R@1) over NNO ($\eta=1$) on both versions of MMD. Moreover, utilizing our subspace-based NEHA, we obtain 5.5% and 5.5% (R@1) improvement over the baseline M-HRED+attention given both versions of MMD. Our best performer, subspace-based NESA yields 6.3% and 7.2% (R@1) improvement over the baseline M-HRED+attention model.

Text-based Task. Table 1 shows results (BLEU and NIST metrics) by comparing target sentences against predicted sentences. Pre-training Text Decoder with French language prior to fine-tuning on English improves results by $\sim 4\%$ and $\sim 2.1\%$ (BLEU) over the M-HRED baseline on both MMD v1 and v2. Using random word deletions for augmentation yielded gain of 0.16% (BLEU) over the M-HRED baseline (MMD v2). Augmentations via so-called sentence compression [2] scored $\sim 1\%$ over M-HRED, whereas augmentations via the so-called back-translation (using translating model [4]) scored $\sim 2.4\%$ over M-HRED. Pre-training on SIMMC [5] was marginally worse (and very similar to combined training on MMD+SIMMC). However, using the assisted supervision, that is, an auxiliary decoder for French, improves results by further $\sim 3.5\%$ (BLEU) over the M-HRED baseline (MMD v2). Augmentations by back translation require translating sentences twice English \rightarrow French \rightarrow English (additional 20 hours), whereas our assisted supervision requires only English \rightarrow French translation. Adding auxiliary German and Russian decoders (to French) and the main decoder for English yields over 4% (BLEU) over the M-HRED baseline (MMD v2). Finally, using the transformer backbone [43] results in a $\sim 5\%$ boost. The benefit of adding multiple auxiliary language decoders is clear. In what follows, we use the GRU backbone not transformers (the backbone choice is a secondary matter). Pre-training the text backbone on the SIMMC dataset [5] before applying our assisted step may also boost results. Applying the sentence compression model [2] via an auxiliary decoder (in addition to French, German and Russian) in our assisted supervision is also possible.

Mixed Task. Firstly, we evaluate our Task Discriminator on the MMD dataset (v2) and note that it achieves 97.0% accuracy. This means that results in Tables 1 and 2 represent

upper bound scores for this paragraph as both tables report on two separate tasks (oracle knowledge regarding which task is which) according to protocol in Saha *et al.* [5]. Table 3 shows that results for the mixed task dropped marginally compared to results in Tables 1 and 2. Our best assisted supervision methods outperformed best baselines M-HRED and M-HRED+attention equipped with Task Discriminator by $\sim 4\%$ (BLEU) and $\sim 7.4\%$ (R@1).

Ablations on NEHA w.r.t. $\eta' \leq \eta$. Below we investigate the impact of subspace size w.r.t. η' and the impact of η nearest neighbors retrieved from DeepFashion on the performance of image-based task. Figure 4c shows that the best performance is attained for $\eta' = \eta = 4$ and the trend suggests that $\eta' \approx \eta$ is a good choice. Figure 4d shows that $\eta' = \eta = 5$ is a better choice for R@3, which allows two incorrect matches precede the correct one. Thus, including more nearest neighbors of positive/negative target images of MMD boosts the score.

Nearest Neighbors+the Hinge Loss. Positive/negative nearest neighbors retrieved from DeepFashion for positive/negative target images can be fed directly into our assisted supervision loss in Eq. (2). Figures 4c and 4d evaluate such a setting ($\eta' = 0$) as it is a special case of our subspace-based approach if $\eta' = 0$ (only μ^- and μ^+ are used if $\eta' = 0$). On average, such a setting is $\sim 2\%$ worse than the subspace-based context. Subspaces capture robustly second-order statistics by discarding eigenvalue scaling and the smallest factors.

User study. We asked 3 users to score our best performer *vs.* M-HRED on MMD (v2) (randomized test) in terms of *clarity*, *compactness* and *helpfulness* on 500 system responses. Table 4 shows that $\sim 61.0\%$ responses of the assisted supervision were clearer and more helpful (*vs.* 39% of M-HRED). Both methods were generating similarly compact responses.

SIMMC. Table 6 shows that using the multilingual decoding head yields 2.3% and $\sim 10\%$ gain (BLEU) on RNN and transformers backbone over the HRE baseline (see the SIMMC paper for details of HRE). Moreover, our visual NESAs yielded $\sim 4\%$ gain (R@1 score).

Runtimes. Our code is implemented in PyTorch and evaluated on an NVIDIA Tesla P100 (unless stated otherwise). Table 5 (runtimes for the text-based tasks) shows that the T-HRED and M-HRED baselines take ~ 15 hours to train. Our assisted supervision (French) uses extra 5 hours. Translations are obtained off-line with translator [54]. However, the best augmentation strategy that we tried (back translation) takes 55 hours, whereas our assisted superv. takes 40 hours (including translation time). Table 7 (runtimes for the image-based tasks) shows that the T-HRED and M-HRED baselines take ~ 15 hours to train. Nearest Neighbor Only, NEHA and NESAs require 1, 3, and 4 extra hours. The off-line pre-processing includes the ResNet-50 feature extraction from MMD and DeepFashion (6 hours), nearest neighbor search with FAISS [23] (1.5 hours, 4 GPUs) and running SVD (2 hours, 4 GPUs).

5 Conclusions

We have introduced the assisted supervision which boosts the performance by leveraging AUDITED. Sampling auxiliary nearest neighbors from the natural manifold of fashion images helps create a meaningful visual context for the image task. With appropriate Soft Assignment reweighting and subspace modeling, benefits become clear while (by design) not posing any extra burden at the testing time. Learning to decode target dialogue sentences in several languages helps reduce the noise of single language syntactic.

Acknowledgement. This work was supported by CSIRO’s Machine Learning and Artificial Intelligence Future Science Platform (MLAI FSP).

References

- [1] Shubham Agarwal, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. A knowledge-grounded multimodal search-based conversational agent. *arXiv preprint arXiv:1810.11954*, 2018.
- [2] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE international conference on computer vision*, pages 37–45, 2015.
- [3] David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. Luke, i am your father: dealing with out-of-domain requests by using movies subtitles. In *International Conference on Intelligent Virtual Agents*, pages 13–21. Springer, 2014.
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [5] Rafael E Banchs. Movie-dic: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–207, 2012.
- [6] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz-Rojas, Leopoldo Pla, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4555–4567, July 2020. doi: 10.18653/v1/2020.acl-main.417. URL <https://acl2020.org/>.
- [7] Indrani Bhattacharya, Arkabandhu Chowdhury, and Vikas C Raykar. Multimodal dialog for browsing large visual catalogs using exploration-exploitation paradigm in a joint embedding space. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 187–191, 2019.
- [8] Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ale s Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, June 2014. URL <http://www.aclweb.org/anthology/W/W14/W14-3302>.
- [9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [10] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [12] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017.
- [13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

- [14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.
- [15] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.
- [16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.
- [17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Conference on Neural Information Processing Systems*, 2014.
- [19] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.
- [20] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019.
- [21] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- [22] Weiwei Hou, Hanna Suominen, Piotr Koniusz, Sabrina B. Caldwell, and Tom Gedeon. A token-wise cnn-based method for sentence compression. In *Neural Information Processing - 27th International Conference, ICONIP*, volume 12532 of *Lecture Notes in Computer Science*, pages 668–679. Springer, 2020. doi: 10.1007/978-3-030-63830-6_56.
- [23] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019.
- [24] Piotr Koniusz and Hongguang Zhang. Power normalizations in fine-grained image, few-shot image and graph classification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2020. doi: 10.1109/TPAMI.2021.3107164.
- [25] Piotr Koniusz, Fei Yan, and Krystian Mikolajczyk. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. *Computer Vision and Image Understanding*, 117(5):479–492, May 2013. ISSN 1077-3142. doi: 10.1016/j.cviu.2012.10.010. URL <https://doi.org/10.1016/j.cviu.2012.10.010>.
- [26] Piotr Koniusz, Yusuf Tas, and Fatih Porikli. Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2, 2017.
- [27] Piotr Koniusz, Fei Yan, Philippe-Henri Gosselin, and Krystian Mikolajczyk. Higher-order occurrence pooling for bags-of-words: Visual concept detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):313–326, 2017. doi: 10.1109/TPAMI.2016.2545667.

- [28] Piotr Koniusz, Yusuf Tas, Hongguang Zhang, Mehrtash Harandi, Fatih Porikli, and Rui Zhang. Museum exhibit identification challenge for the supervised domain adaptation and beyond. In *The European Conference on Computer Vision*, 2018.
- [29] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*, 2019.
- [30] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.
- [31] Viorica Marian and Anthony Shook. The cognitive benefits of being bilingual. *Cerebrum : the Dana forum on brain science*, 2012:13, 10 2012.
- [32] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.
- [33] Seungwhan Moon, Satwik Kottur, Paul A. Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated and interactive multimodal conversations, 2020.
- [34] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*, 2018.
- [35] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*, 2018.
- [36] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, 2010.
- [37] Amrita Saha, Mitesh M Khapra, and Karthik Sankaranarayanan. Towards building large scale multimodal domain-aware conversation systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [38] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. *arXiv preprint arXiv:1507.04808*, 2015.
- [39] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [40] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [41] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.
- [42] Fatemeh Shiri, Xin Yu, Fatih Porikli, Richard Hartley, and Piotr Koniusz. Recovering faces from portraits with auxiliary facial attributes. In *Winter Conference on Applications of Computer Vision*, pages 406–415, 2019.

- [43] Fatemeh Shiri, Xin Yu, Fatih Porikli, Richard Hartley, and Piotr Koniusz. Identity-preserving face recovery from stylized portraits. *International Journal of Computer Vision*, 127(6-7):863–883, 2019. doi: 10.1007/s11263-019-01169-1.
- [44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [45] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.
- [46] Y. Tas and P. Koniusz. Cnn-based action recognition and supervised domain adaptation on 3d body skeletons via kernel feature maps. *British Machine Vision Conference*, 2018.
- [47] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406, 2020.
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [49] Lei Wang and Piotr Koniusz. Hallucinating statistical moment and subspace descriptors for action recognition. *ACM Multimedia*, 2021.
- [50] Lei Wang, Piotr Koniusz, and Du Q. Huynh. Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. *International Conference on Computer Vision*, 2019.
- [51] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.
- [52] R. S. Woodworth and E. L. Thorndike. The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review (1)*, 8(3):247–261, 1901. doi: 10.1037/h0074898.
- [53] Jialin Wu, Zeyuan Hu, and Raymond J Mooney. Joint image captioning and question answering. *arXiv preprint arXiv:1805.08389*, 2018.
- [54] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [55] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Nieves, and Min Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [56] H Zhang, L Zhang, X Qi, H Li, PHS Torr, and P Koniusz. Few-shot action recognition with permutation-invariant attention. In *The European Conference on Computer Vision*, 2020.
- [57] Hongguang Zhang and Piotr Koniusz. Power normalizing second-order similarity network for few-shot learning. *Winter Conference on Applications of Computer Vision*, 2019.
- [58] Hongguang Zhang, Piotr Koniusz, Songlei Jian, Hongdong Li, and Philip H. S. Torr. Rethinking class relations: Absolute-relative supervised and unsupervised few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9432–9441, 2021.

-
- [59] Rui Zhang, Yusuf Tas, and Piotr Koniusz. Artwork identification from wearable camera images for enhancing experience of museum audiences. In *Museums and the Web*, 2017.
- [60] Shan Zhang, Dawei Luo, Lei Wang, and Piotr Koniusz. Few-shot object detection by second-order pooling. *Asian Conference on Computer Vision*, 2020.
- [61] Hao Zhu and Piotr Koniusz. REFINE: Random RangE FInder for Network Embedding. In *ACM International Conference on Information and Knowledge Management*, 2021. doi: 0.1145/3459637.3482168.
- [62] Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *International Conference on Learning Representations*, 2021.
- [63] Hao Zhu, Ke Sun, and Piotr Koniusz. Contrastive laplacian eigenmaps. In *Conference on Neural Information Processing Systems*, 2021.