Bilinear Pooling in video-QA: Empirical challenges and motivational drift from neurological parallels

Thomas Winterbottom [Corresp., 1], **Sarah Xiao** [2], **Alistair McLean** [3], **Noura Al Moubayed** [1]

[1] Department of Computer Science, Durham University, Durham, United Kingdom

[2] Durham University Business School, Durham University, Durham, Durham, United Kingdom

[3] Carbon AI, Middlesbrough, United Kingdom

Corresponding Author: Thomas Winterbottom
Email address: thomas.i.winterbottom@durham.ac.uk

Bilinear pooling (BLP) refers to a family of operations recently developed for fusing features from different modalities predominantly for visual question answering (VQA) models. Successive BLP techniques have yielded higher performance with lower computational expense, yet at the same time they have drifted further from the original motivational justification of bilinear models, instead becoming empircally motivated by task performance. Furthermore, despite significant success in text-image fusion in VQA, BLP has not yet gained such notoriety in video-QA. Though BLP methods have continued to perform well on video tasks when fusing vision and non-textual features, BLP has recently been overshadowed by other vision and textual feature fusion techniques in video-QA. We aim to add a new perspective to the empirical and motivational drift in BLP. We take a step back and discuss the motivational origins of BLP, highlighting the often-overlooked parallels to neurological theories (Dual Coding Theory and The Two-Stream Model of Vision). We seek to carefully and experimentally ascertain the empirical strengths and limitations of BLP as a multimodal text-vision fusion technique in video-QA using 2 models (TVQA baseline and heterogeneous-memory-enchanced 'HME' model) and 4 datasets (TVQA, TGif-QA, MSVD-QA, and EgoVQA). We examine the impact of both simply replacing feature concatenation in the existing models with BLP, and a modified version of the TVQA baseline to accommodate BLP that we name the 'dual-stream' model. We find that our relatively simple integration of BLP does not increase, and mostly harms, performance on these video-QA benchmarks. Using our insights on recent work in BLP for video-QA results and recently proposed theoretical multimodal fusion taxonomies, we offer insight into why BLP-driven performance gain for video-QA benchmarks may be more difficult to achieve than in earlier VQA models. We both share our perspective on, and suggest solutions for, the key issues we identify with BLP techniques for multimodal fusion in video-QA. We look beyond the empirical justification of BLP techniques and propose both alternatives and

improvements to multimodal fusion by drawing neurological inspiration from Dual Coding Theory and the Two-Stream Model of Vision. We qualitatively highlight the potential for neurological inspirations in video-QA by identifying the relative abundance of psycholinguistically ‘concrete’ words in the vocabularies for each of the text components (e.g. questions and answers) of the 4 video-QA datasets we experiment with.

Bilinear Pooling in Video-QA: Empirical Challenges and Motivational Drift from Neurological Parallels

Thomas Winterbottom[1], **Sarah Xiao**[2], **Alistair McLean**[3], **and Noura Al Moubayed**[1]

[1]**Department of Computer Science, Durham University**
[2]**Durham University Business School**
[3]**Carbon AI, United Kingdom**

Corresponding author:
Thomas Winterbottom[1]

Email address: thomas.i.winterbottom@durham.ac.uk

ABSTRACT

Bilinear pooling (BLP) refers to a family of operations recently developed for fusing features from different modalities predominantly for visual question answering (VQA) models. Successive BLP techniques have yielded higher performance with lower computational expense, yet at the same time they have drifted further from the original motivational justification of bilinear models, instead becoming empircally motivated by task performance. Furthermore, despite significant success in text-image fusion in VQA, BLP has not yet gained such notoriety in video-QA. Though BLP methods have continued to perform well on video tasks when fusing vision and *non-textual* features, BLP has recently been overshadowed by other vision and *textual* feature fusion techniques in video-QA. We aim to add a new perspective to the empirical and motivational drift in BLP. We take a step back and discuss the motivational origins of BLP, highlighting the often-overlooked parallels to neurological theories (Dual Coding Theory and The Two-Stream Model of Vision). We seek to carefully and experimentally ascertain the empirical strengths and limitations of BLP as a multimodal text-vision fusion technique in video-QA using 2 models (TVQA baseline and heterogeneous-memory-enchanced 'HME' model) and 4 datasets (TVQA, TGif-QA, MSVD-QA, and EgoVQA). We examine the impact of both simply replacing feature concatenation in the existing models with BLP, and a modified version of the TVQA baseline to accommodate BLP that we name the 'dual-stream' model. We find that our relatively simple integration of BLP does not increase, and mostly harms, performance on these video-QA benchmarks. Using our insights on recent work in BLP for video-QA results and recently proposed theoretical multimodal fusion taxonomies, we offer insight into why BLP-driven performance gain for video-QA benchmarks may be more difficult to achieve than in earlier VQA models. We both share our perspective on, and suggest solutions for, the key issues we identify with BLP techniques for multimodal fusion in video-QA. We look beyond the empirical justification of BLP techniques and propose both alternatives and improvements to multimodal fusion by drawing neurological inspiration from Dual Coding Theory and the Two-Stream Model of Vision. We qualitatively highlight the potential for neurological inspirations in video-QA by identifying the relative abundance of psycholinguistically 'concrete' words in the vocabularies for each of the text components (e.g. questions and answers) of the 4 video-QA datasets we experiment with.

INTRODUCTION

To solve the growing abundance of complex deep learning tasks, it is essential to develop modelling and learning strategies with the capacity to learn complex and nuanced multimodal relationships and representations. To this end, research efforts in multimodal deep learning have taken aim at the relationship between vision and text through visual question answering (VQA) Wu et al. (2017); Srivastava et al. (2020) and more recently video question answering (video-QA) Sun et al. (2021). A particularly notorious solution to learning multimodal relationships in VQA is the family of bilinear pooling (BLP) operators Gao et al. (2016); Kim et al. (2017); Yu et al. (2017); Ben-younes et al. (2017); Yu et al. (2018b);

Ben-Younes et al. (2019). A bilinear (outer product) expansion is thought to encourage models to learn interactions between two feature spaces and has experimentally outperformed 'simpler' vector operations (i.e. concatenation and element-wise-addition/multiplication) on VQA benchmarks. Though successive BLP techniques focus on leveraging higher performance with lower computational expense, which we wholeheartedly welcome, the context of their use has subtly drifted from application in earlier bilinear models e.g. where in Lin et al. (2015) the bilinear mapping is learned between convolution maps (a tangible and visualisable parameter), from compact BLP Gao et al. (2016) onwards the bilinear mapping is learned between indexes of deep feature vectors (a much less tangible unit of representation). Though such changes are not necessarily problematic and the improved VQA performance they have yielded is valuable, they represent a broader trend of the use of BLP methods in multimodal fusion being justified *only* by empirical success. Furthermore, despite BLP's history of success in text-image fusion in VQA, it has not yet gained such notoriety in video-QA. Though BLP methods have continued to perform well on video tasks when fusing vision and *non-textual* features Hu et al. (2021); Zhou et al. (2021); Pang et al. (2021); Xu et al. (2021); Deng et al. (2021); Wang et al. (2021); Deb et al. (2022); Sudhakaran et al. (2021), BLP has recently been overshadowed by other vision and *textual* feature fusion techniques in video-QA Kim et al. (2019); Li et al. (2019); Gao et al. (2019); Liu et al. (2021); Liang et al. (2019). In this paper, we aim to add a new perspective to the empirical and motivational drift in BLP. Our contributions include the following: **I)** We carefully and experimentally ascertain the empirical strengths and limitations of BLP as a multimodal text-vision fusion technique on 2 models (TVQA baseline and heterogeneous-memory-enchanced 'HME' model) and 4 datasets (TVQA, TGif-QA, MSVD-QA and EgoVqa). To this end, our experiments include replacing feature concatenation in the existing models with BLP, and a modified version of the TVQA baseline to accommodate BLP that we name the 'dual-stream' model. Furthermore, we contrast BLP (classified as a 'joint' representation by Baltrušaitis et al. (2019)) with deep canonical cross correlation (a 'co-ordinated representation'). We find that our relatively simple integration of BLP does not increase, and mostly harms, performance on these video-QA benchmarks. **II)** We discuss the motivational origins of BLP and share our observations of bilinearity in text-vision fusion. **III)** By observing trends in recent work using BLP for multimodal video tasks and recently proposed theoretical multimodal fusion taxonomies, we offer insight into why BLP-driven performance gain for video-QA benchmarks may be more difficult to achieve than in earlier VQA models. **IV)** We identify temporal alignment and inefficiency (computational resources *and* performance) as key issues with BLP as a multimodal text-vision fusion technique in video-QA, and highlight concatenation and attention mechanisms as an ideal alternative. **V)** In parallel with the *empirically justified* innovations driving BLP methods, we explore the often-overlooked similarities of bilinear and multimodal fusion to neurological theories e.g. Dual Coding Theory Paivio (2013, 2014) and the Two-Stream Model of Vision Goodale and Milner (1992); Milner (2017), and propose several potential *neurologically justified* alternatives and improvements to existing text-image fusion. We highlight the latent potential already in existing video-QA dataset to exploit neurological theories by presenting a qualitative analysis of occurrence of psycholinguistically 'concrete' words in the vocabularies of the textual components of the 4 video-QA we experiment with.

BACKGROUND: BILINEAR POOLING

In this section we outline the development of BLP techniques, highlight how bilinear models parallel the two-stream model of vision, and discuss where bilinear models diverged from their original motivation.

Concatenation

Early works use Vector concatenation to project different features into a new joint feature space. Zhou et al. (2015) use vector concatenation on the CNN image and text features in their simple baseline VQA model. Similarly, Lu et al. (2016) concatenate image attention and textual features. Vector concatenation is a projection of both input vectors into a new 'joint' dimensional space. Vector concatenation as a multimodal feature fusion technique in VQA is considered a baseline and is generally empirically outperformed in VQA by the following bilinear techniques.

Bilinear Models

Working from the observations that "perceptual systems routinely separate 'content' from 'style"', Tenenbaum and Freeman (2000) proposed a bilinear framework on these two different aspects of purely visual

inputs. They find that the multiplicative bilinear model provides "sufficiently expressive representations of factor interactions". The bilinear model in Lin et al. (2015) is a 'two-stream' architecture where distinct subnetworks model temporal and spatial aspects. The bilinear interactions are between the outputs of two CNN streams, resulting in a bilinear vector that is effectively an outer product directly on convolution maps (features are aggregated with sum-pooling). This makes intuitive sense as individual convolution maps represent specific patterns. It follows that learnable parameters representing the outer product between these maps learn weightings between distinct and visualisable patterns directly. Interestingly, both Tenenbaum and Freeman (2000); Lin et al. (2015) are reminiscent of two-stream hypothesises of visual processing in the human brain Goodale and Milner (1992); Milner and Goodale (2006, 2008); Goodale (2014); Milner (2017) (discussed in detail later). Though these models focus on only visual content, their generalisable two-factor frameworks would later be inspiration to multimodal representations. Fully bilinear representations using deep learning features can easily become impractically large, necessitating informed mathematical compromises to the bilinear expansion.

Compact Bilinear Pooling

Gao et al. (2016) introduce 'Compact Bilinear Pooling', a technique combining the count sketch function Charikar et al. (2002) and convolution theorem Domínguez (2015) in order to 'pool' the outer product into a smaller bilinear representation. Fukui et al. (2016) use compact BLP in their VQA model to learn interactions between text and images i.e. multimodal compact bilinear pooling (MCB). We note that for MCB, the learned outer product is no longer on convolution maps, but rather on the indexes of image and textual tensors. Intuitively, a given index of an image or textual tensor is more abstracted from visualisable meaning when compared to convolution map. As far as we are aware, no research has addressed the potential ramifications of this switch from distinct maps to feature indexes, and later usages of bilinear pooling methods continue this trend. Though MCB is significantly more efficient than full bilinear expansions, they still require relatively large latent dimension to perform well on VQA ($d \approx 16000$).

Multimodal Low-Rank Bilinear Pooling

To further reduce the number of needed parameters, Kim et al. (2017) introduce multimodal low-rank bilinear pooling (MLB), which approximates the outer product weight representation W by decomposing it into two rank-reduced projection matrices:

$$\mathbf{z} = MLB(\mathbf{x},\mathbf{y}) = (X^T\mathbf{x}) \odot (Y^T\mathbf{y})$$
$$\mathbf{z} = \mathbf{x}^T W \mathbf{y} = \mathbf{x}^T X Y^T \mathbf{y} = \mathbb{1}^T (X^T\mathbf{x} \odot Y^T\mathbf{y})$$

where $X \in \mathbb{R}^{m\times o}$, $Y \in \mathbb{R}^{n\times o}$, $o < min(m,n)$ is the output vector dimension, $\odot$ is element-wise multiplication of vectors or the Hadamard product, and $\mathbb{1}$ is the unity vector. MLB performs better than MCB in Osman and Samek (2019), but it is sensitive to hyperparameters and converges slowly. Furthermore, Kim et al. (2017) suggest using *Tanh* activation on the output of $\mathbf{z}$ to further increase model capacity.

Multimodal Factorised Low Rank Bilinear Pooling

Yu et al. (2017) propose multimodal factorised bilinear pooling (MFB) as an extension of MLB. Consider the bilinear projection matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$ outlined above, to learn output $\mathbf{z} \in \mathbb{R}^o$ we need to learn $\mathbf{W} = [\mathbf{W}_0, ..., \mathbf{W}_{o-1}]$. We generalise output $\mathbf{z}$:

$$z_i = \mathbf{x}^T \mathbf{X}_i \mathbf{Y}_i^T \mathbf{y} = \sum_{d=0}^{k-1} \mathbf{x}^T a_d b_d^T \mathbf{y} = \mathbb{1}^T (\mathbf{X}_i^T \mathbf{x} \odot \mathbf{Y}_i^T \mathbf{y}) \tag{1}$$

Note that MLB is equivalent to MFB where k=1. MFB can be thought of as a two-part process: features are 'expanded' to higher-dimensional space by $\mathbf{W}_\sigma$ matrices, then 'squeezed' into a "compact ouput". The authors argue that this gives "more powerful" representational capacity in the same dimensional space than MLB.

Multimodal Tucker Fusion

Ben-younes et al. (2017) extend the rank-reduction concept from MLB and MFB to factorise the entire bilinear tensor using tucker decomposition Tucker (1966) in their multimodal tucker fusion (MUTAN)

model. We will briefly summarise the notion of rank and the mode-n product to describe the tucker decomposition model.

Rank and mode-n product: If $\mathbf{W} \in \mathbb{R}^{I_1 \times,...,\times I_N}$ and $\mathbf{V} \in \mathbb{R}^{J_n \times I_n}$ for some $n \in \{1,...,N\}$ then

$$\text{rank}(\mathbf{W} \otimes_n \mathbf{V}) \leq \text{rank}(\mathbf{W})$$

where $\otimes_n$ is the mode-n tensor product:

$$(\mathbf{W} \otimes_n \mathbf{V})(i_1,...,i_{n-1},j_n,i_{n+1},...,i_N) := \sum_{i_n=1}^{I_n} \mathbf{W}\,(i_1,...,i_{n-1},i_n,i_{n+1},...,i_N)\mathbf{V}(j_n,i_n)$$

In essence, the mode-n fibres (also known as mode-n vectors) of $\mathbf{W} \otimes_n \mathbf{V}$ are the mode-n fibres of $\mathbf{W}$ multiplied by $\mathbf{V}$ (proof here Guillaume OLIKIER (2017)). See Figure 1 for a visualisation of mode-n fibres. Each mode-n tensor product introduces an upper bound to the rank of the tensor. We note that conventionally, the mode-n fibres count from 1 instead of 0. We will follow this convention for the tensor product portion of our paper to avoid confusion. The tucker decomposition of a real 3^{rd} order tensor

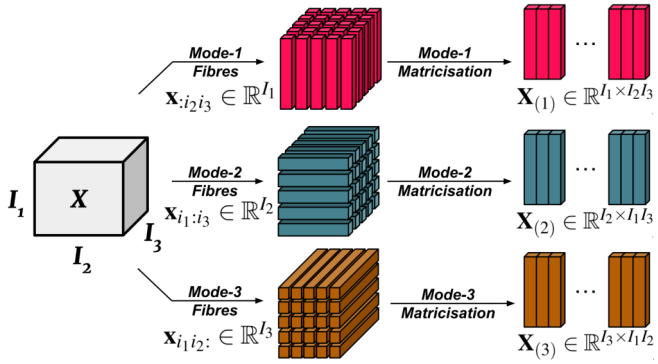

Figure 1. Visualisation of mode-n fibres and matricisation

$\mathbf{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is:

$$\mathbf{T} = \tau \otimes_1 \mathbf{W}_1 \otimes_2 \mathbf{W}_2 \otimes_3 \mathbf{W}_3$$

where $\tau \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *(core tensor)*, and $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3 \in \mathbb{R}^{d_1 \times d_1}$, $\mathbb{R}^{d_2 \times d_2}$, $\mathbb{R}^{d_3 \times d_3}$ *(factor matrices)* respectively.

MUTAN: The MUTAN model uses a reduced rank on the core tensor to constrain representational capacity, and the factor matrices to encode full bilinear projections of the textual and visual features, and finally output an answer prediction, i.e:

$$\mathbf{y} = ((\tau \otimes_1 (\mathbf{q}^T \mathbf{W}_q)) \otimes_2 (\mathbf{v}^T \mathbf{W}_v)) \otimes_3 \mathbf{W}_o$$

Where $\mathbf{y} \in \mathbb{R}^{|A|}$ is the answer prediction vector and $\mathbf{q}, \mathbf{v}$ are the textual and visual features respectively. A slice-wise attention mechanism is used in the MUTAN model to focus on the 'most discriminative interactions'. Multimodal tucker fusion is an empirical improvement over the preceeding BLP techniques on VQA, but it introduces complex hyperparameters to refine that are important for relatively its high performance ($\mathbf{R}$ and core tensor dimensions).

Multimodal Factorised Higher Order Bilinear Pooling

All the BLP techniques discussed up to now are 'second-order', i.e. take two functions as inputs. Yu et al. (2018b) propose multimodal factorised higher-order bilinear pooling (MFH), extending second-order BLP to 'generalised high-order pooling' by stacking multiple MFB units, i.e:

$$\mathbf{z}_{exp}^i = MFB_{exp}^i(\mathbf{I},\mathbf{Q}) = \mathbf{z}_{exp}^{i-1} \odot Dropout(\mathbf{U}^T\mathbf{I} \odot \mathbf{V}^T\mathbf{Q})$$
$$\mathbf{z} = SumPool(\mathbf{z}_{exp})$$

for $i \in \{1,...,p\}$ where $\mathbf{I}$, $\mathbf{Q}$ are visual and text features respectively. Similar to how MFB extends MLB, MFH is MFB where $p = 1$. Though MFH slightly outperforms MFB, there has been little exploration into the theoretical benefit in generalising to higher-order BLP.

Bilinear Superdiagonal Fusion

Ben-Younes et al. (2019) proposed another method of rank restricted bilinear pooling: Bilinear Superdiagonal Fusion (BLOCK). We will briefly outline block term decomposition before describing BLOCK.

Block Term Decomposition: Introduced in a 3-part paper De Lathauwer (2008a,b); De Lathauwer and Nion (2008), block term decomposition reformulates a bilinear matrix representation as the sum of rank restricted matrix products (contrasting low rank pooling which is represented by only a single rank restricted matrix product). By choosing the number of decompositions in the approximated sum and their rank, block-term decompositions offer greater control over the approximated bilinear model. Block term decompositions are easily extended to higher-order tensor decompositions, allowing multilinear rank restriction for multilinear models in future research. A *block term decomposition* of a tensor $\mathbf{W} \in \mathbb{R}^{I_1 \times,...,\times I_N}$ is a decomposition of the form:

$$\mathbf{W} = \sum_{r=1}^{R} \mathbf{S}_r \otimes_1 \mathbf{U}_r^1 \otimes_2 \mathbf{U}_r^2 \otimes_3, ..., \otimes_n \mathbf{U}_r^n$$

where $R \in \mathbb{N}^*$ and for each $r \in \{1,...,R\}, \mathbf{S}_r \in \mathbb{R}^{R_1 \times,...,\times R_n}$ where each $\mathbf{S}_r$ are 'core tensors' with dimensions $R_n \leq I_n$ for $n \in \{1,...,N\}$ that are used to restrict the rank of the tensor $\mathbf{W}$. $\mathbf{U}_r^n \in \mathrm{St}(R_n, I_n)$ are the 'factor matrices' that intuitively expand the n^{th} dimension of $\mathbf{S}$ back up to the original n^{th} dimension of $\mathbf{W}$. $\mathrm{St}(a,b)$ here refers to the Stiefel manifold, i.e. $\mathrm{St}(a,b)$:$\{\mathbf{Y} \in \mathbb{R}^{a \times b} : \mathbf{Y}^T\mathbf{Y} = \mathbf{I}_p\}$. Figure 2 visualises the block term decomposition process.

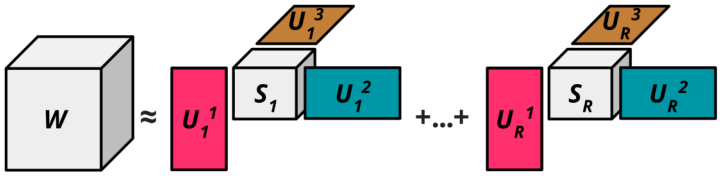

Figure 2. Block Term Decomposition (n=3)

Bilinear Superdiagonal Model: The BLOCK model uses block term decompositions to learn multimodal interactions. The authors argue that since BLOCK enables "very rich (full bilinear) interactions between groups of features, while the block structure limits the complexity of the whole model", that it is able to represent very fine grained interactions between modalities while maintaining powerful mono-modal representations. The bilinear model with inputs $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$ is projected into o dimensional space with tensor products:

$$\mathbf{z} = \mathbf{W} \otimes_1 \mathbf{x} \otimes_2 \mathbf{y}$$

where $\mathbf{z} \in \mathbb{R}^o$. The superdiagonal BLOCK model uses a 3 dimensional block term decomposition. The decomposition of $\mathbf{W}$ in rank (R_1, R_2, R_3) is defined as:

$$\mathbf{W} = \sum_{r=1}^{R} \mathbf{S}_r \otimes_1 \mathbf{U}_r^1 \otimes_2 \mathbf{U}_r^2 \otimes_3 \mathbf{U}_r^3$$

This can be written as

$$\mathbf{W} = \mathbf{S}^{bd} \otimes_1 \mathbf{U}^1 \otimes_2 \mathbf{U}^2 \otimes_3 \mathbf{U}^3$$

where $\mathbf{U}^1 = [\mathbf{U}_1^1, ..., \mathbf{U}_R^1]$, similarly with $\mathbf{U}^2$ and $\mathbf{U}^3$, and now $\mathbf{S}^{bd} \in \mathbb{R}^{RR^1 \times RR^2 \times RR^3}$. So $\mathbf{z}$ can now be expressed with respect to $\mathbf{x}$ and $\mathbf{y}$. Let $\hat{\mathbf{x}} = \mathbf{U}^1\mathbf{x} \in \mathbb{R}^{RR^1}$ and $\hat{\mathbf{y}} = \mathbf{U}^2\mathbf{y} \in \mathbb{R}^{RR^2}$. These two projections are merged by the block-superdiagonal tensor $\mathbf{S}^{bd}$. Each block in $\mathbf{S}^{bd}$ merges together blocks of size R^1 from $\hat{\mathbf{x}}$ and of size R^2 from $\hat{\mathbf{y}}$ to produce a vector of size R^3:

$$\mathbf{z}_r = \mathbf{S}_r \otimes_x \hat{\mathbf{x}}_{rR^1:(r+1)R^1} \otimes_y \hat{\mathbf{y}}_{rR^2:(r+1)R^2}$$

where $\hat{\mathbf{x}}_{i:j}$ is the vector of dimension $j-i$ containing the corresponding values of $\hat{\mathbf{x}}$. Finally all vectors $\mathbf{z}_r$ are concatenated producing $\hat{\mathbf{z}} \in \mathbb{R}^{RR^3}$. The final prediction vector is $\mathbf{z} = \mathbf{U}^3, \hat{\mathbf{z}} \in \mathbb{R}^o$. Similar to tucker fusion, the block term decomposition based fusion in BLOCK theoretically allows more nuanced control on representation size and empirically outperforms previous techniques.

RELATED WORKS

Bilinear Pooling in Video-QA With Language-Vision Fusion

We aim to highlight and explore a broad shift away from BLP in favour of methods such as attention in video-QA benchmarks. Several video models have incorporated and contrasted BLP techniques to their own model designs for language-vision fusion tasks. Kim et al. (2019) find various BLP fusions perform worse than their 'dynamic modality fusion' mechanism on TVQA Lei et al. (2018) and MovieQA Tapaswi et al. (2016). Li et al. (2019) find MCB fusion performs worse on their model in ablation studies on TGIF-QA Jang et al. (2017). Chou et al. (2020) use MLB as part of their baseline model proposed alongside their 'VQA 360°' dataset. Gao et al. (2019) contrast their proposed two-stream attention mechanism to an MCB model for TGIF-QA, demonstrating a substantial performance increase over the MCB model. Liu et al. (2021) use MUTAN fusion between question and visual features to yield impressive results on TGif-QA, though they are outperformed by an attention based model using element-wise multiplication Le et al. (2020). The Focal Visual-Text Attention network (FVTA) Liang et al. (2019) is a hierarchical model that aims to dynamically select from the appropriate point across both time and modalities that outperforms an MCB approach on Movie-QA.

Bilinear Pooling in Video Without Language-Vision Fusion

Where recent research in video-QA tasks (which includes textual questions as input) has moved away from BLP techniques, several video tasks that do *not* involve language have found success using BLP techniques. Zhou et al. (2021) use a multilevel factorised BLP based model to fuse audio and visual features for emotion recognition in videos. Hu et al. (2021) use compact BLP to fuse audio and 'visual long range' features for human action recognition. Pang et al. (2021) use MLB as part of an attention-based fusion for audio and visual features for violence detection in videos. Xu et al. (2021) use BLP to fuse visual features from different channels in RGBT tracking. Deng et al. (2021) use compact BLP to fuse spatial and temporal representations of video features for action recognition. Wang et al. (2021) fuse motion and appearance visual information together achieving state-of-the-art results on MSVD-QA. Sudhakaran et al. (2021) draw design inspiration from bilinear processing of Lin et al. (2015) and MCB to propose 'Class Activation Pooling' for video action recognition. Deb et al. (2022) use MLB to process video features for video captioning.

DATASETS

In this section, we outline the video-QA datasets we use in our experiments.

MSVD-QA

Xu et al. (2017) argue that simply extending image-QA methods is "insufficient and suboptimal" to conduce quality video-QA, and that instead the focus should be on the temporal structure of videos. Using an NLP method to automatically generate QA pairs from descriptions Heilman and Smith (2009), Xu et al. (2017) create the MSVD-QA dataset based on the Microsoft research video description corpus Chen and Dolan (2011). The dataset is made from 1970 video clips, with over 50k QA pairs in '5w' style i.e. ("what", "who", "how", "when", "where").

TGIF-QA

Jang et al. (2017) speculate that the relatively limited progress in video-QA compared to image-QA is "due in part to the lack of large-scale datasets with well defined tasks". As such, they introduced the TGIF-QA dataset to 'complement rather than compete' with existing VQA literature and to serve as a bridge between video-QA and video understanding. To this end, they propose 3 subsets with specific video-QA tasks that aim to take advantage of the temporal format of videos:

Count: Counting the number of times a specific action is repeated Levy and Wolf (2015) e.g. "How many times does the girl jump?". Models output the predicted number of times the specified actions happened. (Over 30k QA pairs).

Action: Identify the action that is repeated a number of times in the video clip. There are over 22k multiple choice questions e.g. "What does the girl do 5 times?".

Trans: Identifying details about a state transition Isola et al. (2015). There are over 58k multiple choice questions e.g. "What does the man do after the goal post?".

Frame-QA: An image-QA split using automatically generated QA pairs from frames and captions in the TGIF dataset Li et al. (2016) (over 53k multiple choice questions).

TVQA

The TVQA dataset Lei et al. (2018) is designed to address the shortcomings of previous datasets. It has significantly longer clip lengths than other datasets and is based on TV shows instead of cartoons to give it realistic video content with simple coherent narratives. It contains over 150k QA pairs. Each question is labelled with timestamps for the relevant video frames and subtitles. The questions were gathered using AMT workers. Most notably, the questions were specifically designed to encourage multimodal reasoning by asking the workers to design two-part compositional questions. The first part asks a question about a 'moment' and the second part localises the relevant moment in the video clip i.e. [What/How/Where/Why/Who/...] — [when/before/after] —, e.g. '*[What] was House saying [before] he leaned over the bed?*'. The authors argue this facilitates questions that require both visual and language information since "people often naturally use visual signals to ground questions in time". The authors identify certain biases in the dataset. They find that the average length of correct answers are longer than incorrect answers. They analyse the performance of their proposed baseline model with different combinations of visual and textual features on different question types they have identified. Though recent analysis has highlighted bias towards subtitles in TVQA's questions Winterbottom et al. (2020), it remains an important large scale video-QA benchmark.

EgoVQA

Most video-QA datasets focus on video-clips from the 3^{rd} person. Fan (2019) argue that 1^{st} person video-QA has more natural use cases that real-world agents would need. As such, they propose the egocentric video-QA dataset (EgoVQA) with 609 QA pairs on 16 first-person video clips. Though the dataset is relatively small, it has a diverse set of question types (e.g. 1^{st} & 3^{rd} person 'action' and 'who' questions, 'count', 'colour' etc..), and aims to generate hard and confusing incorrect answers by sampling from correct answers of the same question type. Models on EgoVQA have been shown to overfit due to its small size. To remedy this, Fan (2019) pretrain the baseline models on the larger YouTube2Text-QA Ye et al. (2017). YouTube2Text-QA is a multiple choice dataset created from MSVD videos Chen and Dolan (2011) and questions created from YouTube2Text video description corpus Guadarrama et al. (2013). YouTube2Text-QA has over 99k questions in 'what', 'who' and 'other' style.

MODELS

In this section, we describe the models used in our experiments, built from the official TVQA [1] and HME-VideoQA [2] implementations.

TVQA Model

Model Definition: The model takes as inputs: a question q, five potential answers $\{a_i\}_{i=0}^{4}$, a subtitle S and corresponding video-clip V, and outputs the predicted answer. As the model can either use the entire video-clip and subtitle or only the parts specified in the timestamp, we refer to the sections of video and subtitle used as segments from now on. Figure 3 demonstrates the textual and visual streams and their associated features in model architecture.

ImageNet Features: Each frame is processed by a ResNet101 He et al. (2016) pretrained on ImageNet Deng et al. (2009) to produce a 2048-d vector. These vectors are then L2-normalised and stacked in frame order: $V^{img} \in \mathbb{R}^{f \times 2048}$ where f is the number of frames used in the video segment.

Regional Features: Each frame is processed by a Faster R-CNN Ren et al. (2015) trained on Visual Genome Krishna et al. (2017) in order to detect objects. Each detected object in the frame is given a bounding box, and has an affiliated 2048-d feature extracted. Since there are multiple objects detected per frame (we cap it at 20 per frame), it is difficult to efficiently represent this in time sequences Lei et al. (2018). The model uses the top-K regions for all detected labels in the segment as in Anderson et al. (2018) and Karpathy and Fei-Fei (2015). Hence the regional features are $V^{reg} \in \mathbb{R}^{n_{reg} \times 2048}$ where n_{reg} is the number of regional features used in the segment.

[1] https://github.com/jayleicn/TVQA

[2] https://github.com/fanchenyou/HME-VideoQA

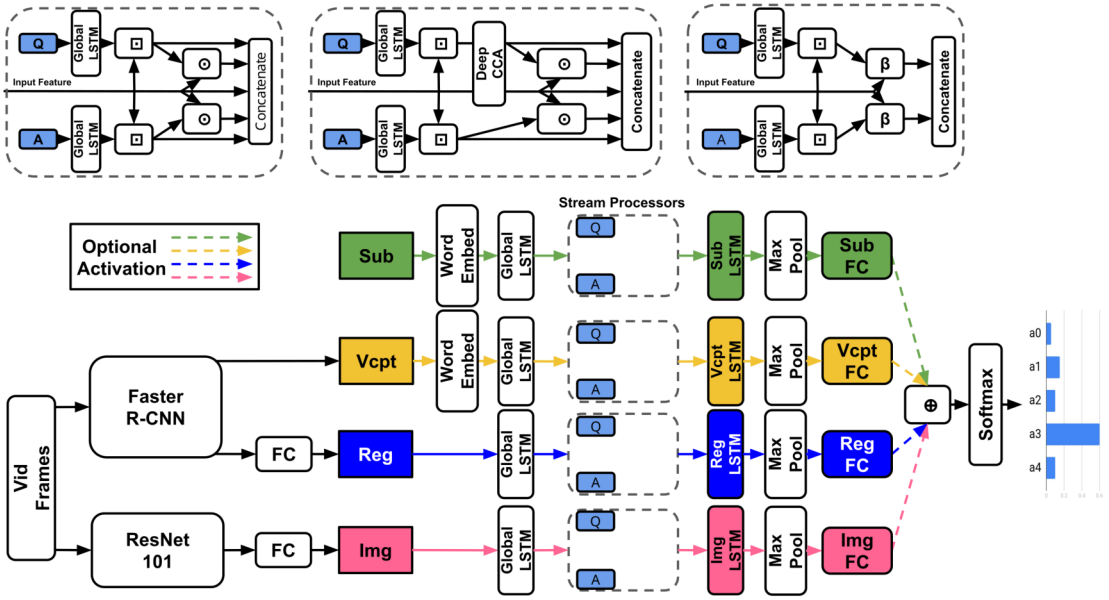

Figure 3. TVQA Model. $\odot/\oplus$ = Element-wise multiplication/addition, $\boxdot$ = context matching Seo et al. (2017); Yu et al. (2018a), β = BLP. Any feature streams may be enabled/disabled.

Visual Concepts: The classes or labels of the detected regional features are called 'Visual Concepts'. Yin and Ordonez (2017) found that simply using detected labels instead of image features gives comparable performance on image captioning tasks. Importantly they argued that combining CNN features with detected labels outperforms either approach alone. Visual concepts are represented as either GloVe Pennington et al. (2014) or BERT Devlin et al. (2019) embeddings $V^{vcpt} \in \mathbb{R}^{n_{vcpt} \times 300}$ or $\mathbb{R}^{n_{vcpt} \times 768}$ respectively, where n_{vcpt} is the number of visual concepts used in the segment.

Text Features: The model encodes the questions, answers, and subtitles using either GloVe ($\in \mathbb{R}^{300}$) or BERT embeddings ($\in \mathbb{R}^{768}$). Formally, $q \in \mathbb{R}^{n_q \times d}, \{a_i\}_{i=0}^{4} \in \mathbb{R}^{n_{a_i} \times d}, S \in \mathbb{R}^{n_s \times d}$ where n_q, n_{a_i}, n_s is the number of words in q, a_i, S respectively and $d = 300, 768$ for GloVe or BERT embeddings respectively.

Context Matching: Context matching refers to context-query attention layers recently adopted in machine comprehension Seo et al. (2017); Yu et al. (2018a). Given a context-query pair, context matching layers return 'context aware queries'.

Model Details: Any combination of subtitles or visual features can be used. All features are mapped into word vector space through a tanh non-linear layer. They are then processed by a shared bi-directional LSTM Hochreiter and Schmidhuber (1997); Graves and Schmidhuber (2005) ('Global LSTM' in Figure 3) of output dimension 300. Features are context-matched with the question and answers. The original context vector is then concatenated with the context-aware question and answer representations and their combined element-wise product ('Stream Processor' in Figure 3, e.g. for subtitles S, the stream processor outputs $[F^{sub}; A^{sub,q}; A^{sub,a_{0-4}}; F^{sub} \odot A^{sub,q}; F^{sub} \odot A^{sub,a_{0-4}}] \in \mathbb{R}^{n_{sub} \times 1500}$ where $F^{sub} \in \mathbb{R}^{n_s \times 300}$. Each concatenated vector is processed by their own unique bi-directional LSTM of output dimension 600, followed by a pair of fully connected layers of output dimensions 500 and 5, both with dropout 0.5 and ReLU activation. The 5-dimensional output represents a vote for each answer. The element-wise sum of each activated feature stream is passed to a softmax producing the predicted answer ID. All features remain separate through the entire network, effectively allowing the model to choose the most useful features.

HME-VideoQA

To better handle semantic meaning through long sequential video data, recent models have integrated external 'memory' units Xiong et al. (2016); Sukhbaatar et al. (2015) alongside recurrent networks to handle input features Gao et al. (2018); Zeng et al. (2017). These external memory units are designed to encourage multiple iterations of inference between questions and video features, helping the model revise it's visual understanding as new details from the question are presented. The heterogeneous

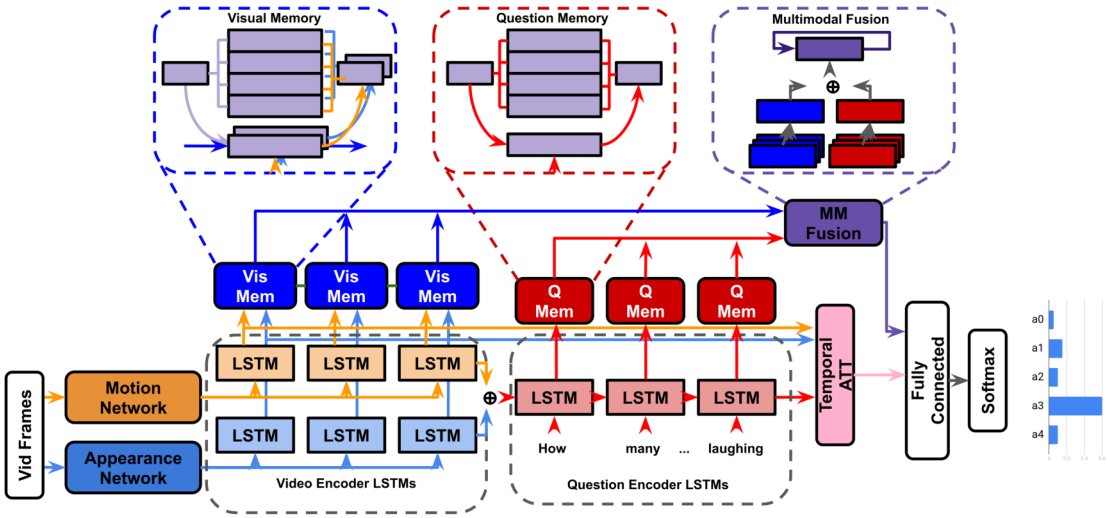

Figure 4. HME Model

memory-enhanced video-QA model (HME) Fan et al. (2019) proposes several improvements to previous memory based architectures:

Heterogeneous Read/Write Memory: The memory units in HME use an attention-guided read/write mechanism to read from/update memory units respectively (the number of memory slots used is a hyper-parameter). The claim is that since motion and appearance features are heterogeneous, a 'straightforward' combination of them cannot effectively describe visual information. The video memory aims to effectively fuses motion (C3D Tran et al. (2014)) and appearance (ResNet He et al. (2016) and VGG Simonyan and Zisserman (2015)) features by integrating them in the joint read/write operations (visual memory in Figure 4).

Encoder-Aware Question Memory: Previous memory models used a single feature vector outputted by an LSTM or GRU for their question representation Gao et al. (2018); Zeng et al. (2017); Xiong et al. (2016); Anderson et al. (2018). HME uses an LSTM question encoder and question memory unit pair that augment each other dynamically (question memory in Figure 4).

Multimodal Fusion Unit: The hidden states of the video and question memory units are processed by a temporal attention mechanism. The joint representation 'read' updates the fusion unit's own hidden state. The visual and question representations are ultimately fused by vector concatenation (multimodal fusion in Figure 5). Our experiments will involve replacing this concatenation step with BLP techniques.

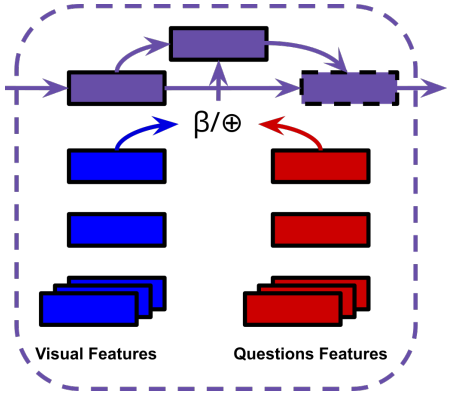

Figure 5. $\oplus$ = Concatenation, β = BLP.

EXPERIMENTS AND RESULTS

In this section we outline our experimental setup and results. We save our insights for the discussion in the next section. See our GitHub repository[3] for both the datasets and code used in our experiments. Table 1 shows the benchmarks and SotA results for the datasets we consider in this paper.

Dataset	Benchmark	SoTA
TVQA (Val)	68.85% Lei et al. (2018)	74.97% Khan et al. (2020)
TVQA (Test)	68.48% Lei et al. (2018)	72.89% Khan et al. (2020)
EgoVQA (Val 1)	37.57% Fan (2019)	45.05%* Chenyou (2019)
EgoVQA (Test 2)	31.02% Fan (2019)	43.35%* Chenyou (2019)
MSVD-QA	32.00% Xu et al. (2017)	40.30% Guo et al. (2021)
TGIF-Action	60.77% Jang et al. (2017)	84.70% Le et al. (2020)
TGIF-Count	4.28† Jang et al. (2017)	2.19† Le et al. (2020)
TGIF-Trans	67.06% Jang et al. (2017)	87.40% Seo et al. (2021)
TGIF-FrameQA	49.27% Jang et al. (2017)	64.80% Le et al. (2020)

Table 1. Dataset benchmark and SoTA results to the best of our knowledge. † = Mean L2 loss. * = Results we replicated using the cited implementation.

Concatenation to BLP (TVQA)

As previously discussed, BLP techniques have outperformed feature concatenation on a number of VQA benchmarks. The baseline stream processor concatenates the visual feature vector with question and answer representations. Each of the 5 inputs to the final concatenation are 300-d. We replace the visual-question/answer concatenation with BLP (Figure 6). All inputs to the BLP layer are 300-d, the outputs are 750-d and the hidden size is 1600 (a smaller hidden state than normal, however, the input features are also smaller compared to other uses of BLP). We make as few changes as possible to accommodate BLP, i.e. we use context matching to facilitate BLP fusion by aligning visual and textual features temporally. Our experiments include models with/without subtitles or questions (Table 2).

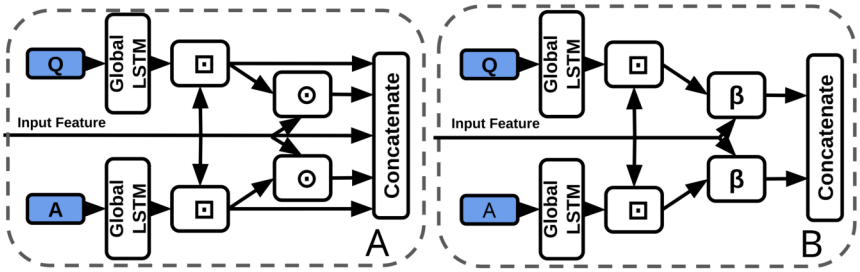

Figure 6. Baseline concatenation stream processor from TVQA model (left-A) vs Our BLP stream processor (right-B). $\odot$ = Element-wise multiplication, β = BLP, $\boxdot$ = Context Matching.

Dual-Stream Model

We create our 'dual-stream' (Figure 7, Table 3) model from the SI TVQA baseline model for 2 main purposes: **I)** To explore the effects of a joint representation on TVQA, **II)** To contrast the concatenation-replacement experiment with a model restructured specifically with BLP as a focus. The baseline BLP model keeps subtitles and other visual features completely separate up to the answer voting step. Our aim here is to create a joint representation BLP-based model similar in essence to the baseline TVQA model that fuses subtitle and visual features. As before, we use context matching to temporally align the video and text features.

[3] https://github.com/Jumperkables/trying_blp

Subtitles	Fusion Type	Accuracy	Baseline Offset
-	Concatenation	45.94%	-
GloVE	Concatenation	69.74%	-
BERT	Concatenation	72.20%	-
- (No Q)	Concatentation	45.58%	-0.36%
GloVE (No Q)	Concatentation	68.31%	-1.42%
BERT (No Q)	Concatentation	70.43%	-1.77%
-	MCB	**45.65%**	**-0.29%**
GloVE	MCB	**69.32%**	**-0.42%**
BERT	MCB	**71.68%**	**-0.52%**
-	MLB	41.98%	-3.96%
GloVE	MLB	69.30%	-0.44%
BERT	MLB	69.04%	-3.16%
-	MFB	41.82%	-4.12%
GloVE	MFB	68.87%	-0.87%
BERT	MFB	67.29%	-4.91%
-	MFH	44.44%	-1.5%
GloVE	MFH	68.43%	-1.31%
BERT	MFH	67.29%	-4.91%
-	Blocktucker	44.44%	-1.5%
GloVE	Blocktucker	67.95%	-1.79%
BERT	Blocktucker	67.04%	-5.16%
-	BLOCK	41.09%	-4.85%
GloVE	BLOCK	65.31%	-4.43%
BERT	BLOCK	66.94%	-5.26%

Table 2. Concatenation replaced with BLP in the TVQA model on the TVQA Dataset. All models use visual concepts and ImageNet features. 'No Q' indicates questions are not used as inputs i.e. answers rely purely on input features.

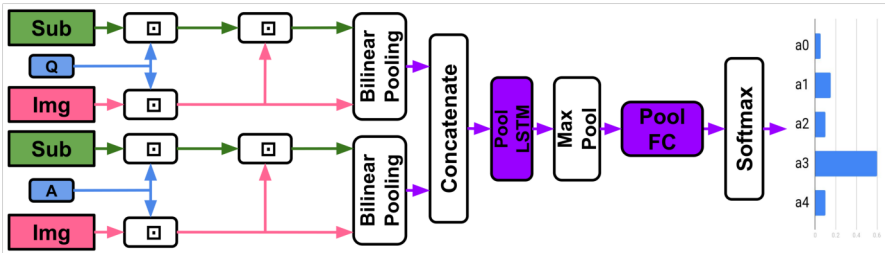

Figure 7. Our Dual-Stream Model. ⊡ = Context Matching.

Deep CCA in TVQA

In contrast to joint representations, Baltrušaitis et al. (2019) define 'co-ordinated representations' as a category of multimodal fusion techniques that learn "separated but co-ordinated" representations for each modality (under some constraints). Peng et al. (2018) claim that since there is often an information imbalance between modalities, learning separate modality representations can be beneficial for preserving 'exclusive and useful modality-specific characteristics'. We include one such representation, deep canonical correlation analysis (DCCA) Andrew et al. (2013), in our experiments to contrast with the joint BLP models.

CCA

Canonical cross correlation analysis (CCA) Hotelling (1936) is a method for measuring the correlations between two sets. Let $(\mathbf{X_0}, \mathbf{X_1}) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_1}$ be random vectors with covariances $(\sum_{r=00}, \sum_{r=11})$ and cross-covariance $\sum_{r=01}$. CCA finds pairs of linear projections of the two views $(w_0'\mathbf{X_0}, w_1'\mathbf{X_1})$ that are maximally correlated:

$$\rho = (w_0^*, w_1^*) = \underset{w_0, w_1}{argmax}\ corr(w_0'\mathbf{X_0}, w_1'\mathbf{X_1})$$

$$= \underset{w_0,w_1}{argmax} \frac{w_0' \sum_{01} w_1}{\sqrt{w_0' \sum_{00} w_0 w_1' \sum_{11} w_1}}$$

where ρ is the correlation co-efficient. As ρ is invariant to the scaling of w_0 and w_1, the projections are constrained to have unit variances, and can be represented as the following maximisation:

$$\underset{w_0,w_1}{argmax}\ w_0' \sum_{01} w_1 \ s.t\ w_0' \sum_{00} w_0 = w_1' \sum_{11} w_1 = \mathbf{1}$$

However, CCA can only model linear relationships regardless of the underlying realities in the dataset. Thus, CCA extensions were proposed, including kernel CCA (KCCA) Akaho (2001) and later DCCA.

DCCA

DCCA is a parametric method used in multimodal neural networks that can learn non-linear transformations for input modalities. Both modalities t, v are encoded in neural-network transformations $H_t, H_v = f_t(t, \theta_t), f_v(v, \theta_v)$, and then the canonical correlation between both modalities is maximised in a common subspace (i.e. maximise cross-modal correlation between H_t, H_v).

$$\max corr(H_t, H_v) = \underset{\theta_t,\theta_v}{argmax}\ corr(f_t(t, \theta_t), f_v(v, \theta_v))$$

We use DCCA over KCCA to co-ordinate modalities in our experiments as it is generally more stable and efficient, learning more 'general' functions.

DCCA in TVQA

We use a 2-layer DCCA module to coordinate question and context (visual or subtitle) features (Figure 8, Table 4). Output features are the same dimensions as inputs. Though DCCA itself is not directly related to BLP, it has recently been classified as a coordinated representation Guo et al. (2019), which contrasts a 'joint' representation.

Model	Text	Val Acc
TVQA SI	GloVe	67.78%
TVQA SI	BERT	70.56%
Dual-Stream MCB	GloVe	63.46%
Dual-Stream MCB	BERT	60.63%
Dual-Stream MFH	GloVe	62.71%
Dual-Stream MFH	BERT	59.34%

Table 3. Dual-Stream Results Table. 'SI' for TVQA models indicates the model is using subtitle and ImageNet feature streams only, i.e. the green and pink streams in Figure 3

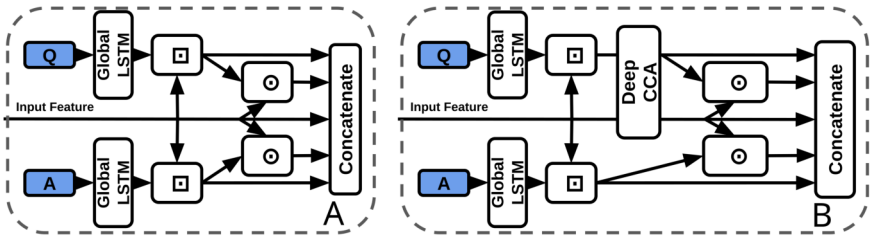

Figure 8. Baseline concatenation stream processor from TVQA model (left-A) vs Our DCCA stream processor (right-B). $\odot$ = Element-wise multiplication, $\boxdot$ = Context Matching.

Concatenation to BLP (HME-VideoQA)

As described in the previous section, we replace a concatenation step in the HME model between textual and visual features with BLP (Figure 5, corresponding to the multimodal fusion unit in Figure 4). The goal here is to explore if BLP can better facilitate multimodal fusion in aggregated memory features (Table 5). We replicate the results from Fan et al. (2019) with the HME on the MSVD, TGIF and EgoVQA datasets using the official github repository Chenyou (2019). We extract our own C3D features from the frames in the TVQA.

Model	Text	Baseline Acc	DCCA Acc
VI	GloVe	45.94%	45.00% (-0.94%)
VI	BERT	–	41.70%
SVI	GloVe	69.74%	67.91% (-1.83%)
SVI	BERT	72.20%	68.48% (-3.72%)

Table 4. DCCA in the TVQA Baseline Model.

Dataset	Fusion Type	Val	Test
TVQA (GloVE)	Concatenation	41.25%	N/A
EgoVQA-0	Concatenation	36.99%	37.12%
EgoVQA-1	Concatenation	48.50%	43.35%
EgoVQA-2	Concatenation	45.05%	39.04%
MSVD-QA	Concatenation	30.94%	33.42%
TGIF-Action	Concatenation	70.69%	73.87%
TGIF-Count	Concatenation	3.95†	3.92†
TGIF-Trans	Concatenation	76.33%	78.94%
TGIF-FrameQA	Concatenation	52.48%	51.41%
TVQA (GloVE)	MCB	41.09% (-0.16%)	N/A%
EgoVQA-0	MCB	No Convergence	No Convergence
EgoVQA-1	MCB	No Convergence	No Convergence
EgoVQA-2	MCB	No Convergence	No Convergence
MSVD-QA	MCB	30.85% (-0.09%)	33.78% (+0.36%)
TGIF-Action	MCB	73.56% (+2.87%)	73.00% (-0.87%)
TGIF-Count	MCB	3.95† (+0†)	3.98† (+0.06†)
TGIF-Trans	MCB	79.30% (+2.97%)	77.10% (-1.84%)
TGIF-FrameQA	MCB	51.72% (-0.76%)	52.21% (+0.80%)

Table 5. HME-VideoQA Model. The default fusion technique is concatenation. † refers to minimised L2 loss.

DISCUSSION

TVQA Experiments

No BLP Improvements on TVQA: On the HME concat-to-BLP substitution model (Table 5), MCB barely changes model performance at all. We find that none of our TVQA concat-to-BLP substitutions (Table 2) yield any improvements at all, with almost all of them performing worse overall (0.3-5%) than even the questionless concatenation model. Curiously, MCB scores the highest of all BLP techniques. The dual-stream model performs worse still, dropping accuracy by between 5-10% vs the baseline (Table 3). Similarly, we find that MCB performs best despite being known to require larger latent spaces to work on VQA.

BERT Impacted the Most: For the TVQA BLP-substitution models, we find the GloVe, BERT and 'no-subtitle' variations all degrade by roughly similar margins, with BERT models degrading more most often. This slight discrepancy is unsurprising as the most stable BERT baseline model is the best, and thus may degrade more on the inferior BLP variations. However, BERT's relative degradation is much more pronounced on the dual-stream models, performing 3% worse than GloVe. We theorise that here, the significant and consistent drop is potentially caused by BERT's more contextual nature is no longer helping, but actively obscuring more pronounced semantic meaning learned from subtitles and questions.

Blame Smaller Latent Spaces?: Naturally, bilinear representations of time series data across multiple frames or subtitles are highly VRAM intensive. Thus we can only explore relatively small hidden dimensions (i.e. 1600). However, we cannot simply conclude our poor results are due to our relatively small latent spaces because: **I)** MCB is our best performing BLP technique. However, MCB has been outperformed by MFH on previous VQA models *and* it has been shown to require much larger latent spaces to work effectively in the first place Fukui et al. (2016) (16000). **II)** Our vector representations of text and images are also much smaller (300-d) compared to the larger representation dimensions conventional in previous benchmarks (e.g. 2048 in Fukui et al. (2016)). We note that $16000/2048 \approx 1600/300$, and so our latent-to-input size ratio is not substantially different to previous works.

Unimodal Biases in TVQA and Joint Representation: Another explanation may come from works exploring textual biases inherent in TVQA to textual modalities Winterbottom et al. (2020). BLP has been categorised as a 'joint representation'. Baltrušaitis et al. (2019) consider representation as summarising multimodal data "in a way that exploits the complementarity and redundancy of multiple modalities". Joint representations combine unimodal signals into the same representation space. However, they struggle to handle missing data Baltrušaitis et al. (2019) as they tend to preserve shared semantics while ignoring modality-specific information Guo et al. (2019). The existence of unimodal text bias in TVQA implies BLP may perform poorly on the TVQA as a joint representation of it's features because: **I)** information from either modality is consistently missing, **II)** prioritising 'shared semantics' over 'modality-specific' information harms performance on TVQA. Though concatenation could also be classified as a joint representation, we argue that this observation still has merit. Theoretically, a concatenation layer can still model modality specific information (see Figure 9), but a bilinear representation would seem to inherently entangle its inputs which would make modality specific information more challenging to learn since each parameter representing one modality is by definition weighted with the other. This may explain why our simpler BLP substitutions perform better than our more drastic 'joint' dual-stream model.

What About DCCA?: Table 4 shows our results on the DCCA augmented TVQA models. We see a slight but noticable performance degradation with this relatively minor alteration to the stream processor. As previously mentioned, DCCA is in some respects an opposite approach to multimodal fusion than BLP, i.e. a 'coordinated representation'. The idea of a coordinated representations is to learn a separate representation for each modality , but with respect to the other. In this way, it is thought that multimodal interactions can be learned while still preserving modality-specific information that a joint representation may otherwise overlook Guo et al. (2019); Peng et al. (2018). DCCA specifically maximises cross-modal correlation. Without further insight from surrounding literature, it is difficult to conclude what TVQA's drop in performance using both joint *and* coordinated representations could mean. We will revisit this when we discuss the role of attention in multimodal fusion.

Does Context Matching Ruin Multimodal Integrity?: The context matching technique used in the TVQA model is the birdirectional attention flow (BiDAF) module introduced in Seo et al. (2017). It is used in machine comprehension between a textual context-query pair to generate query-aware context representations. BiDAF uses a 'memoryless' attention mechanism where information from each time step does not directly affect the next, which is thought to prevent early summarisation. BiDAF considers different input features at different levels of granularity. The TVQA model uses bidirectional attention flow to create context aware (visual/subtitle) question and answer representations. BiDAF can be seen as a co-ordinated representation in some regards, but it does project questions and answers representations into a new space. We use this technique to prepare our visual and question/answer features because it temporally aligns both features, giving them the same dimensional shape, conveniently allowing us to apply BLP at each time step. Since the representations generated are much more similar than the original raw features and there is some degree of information exchange, it may affect BLP's representational capacity. Though it is worth considering these potential shortcomings, we cannot immediately assume that BiDAF would cause serious issues as earlier bilinear technique were successfully used between representations in the same modality Tenenbaum and Freeman (2000); Gao et al. (2016). This implies that multimodal interactions can still be learned between the more similar context-matched representations, provided the information is still present. Since BiDAF does allow visual information to be used in the TVQA baseline model, it is reasonable to assume that some of the visual information is in fact intact and exploitable for BLP. However, it is still currently unclear if context matching is fundamentally disrupting BLP and contributing to the poor results we find. We note that in BiDAF, 'memoryless' attention is implemented to avoid propagating errors through time. We argue that though this may be true and help in some circumstances, conversely, this will not allow some useful interactions to build up over time steps.

The Other Datasets on HME

BLP Has No Effect: Our experiments on the EgoVQA, TGIF-QA and MSVD-QA datasets are on concat-to-BLP substitution HME models. Our results are inconclusive. There is virtually no variation in performance between the BLP and concatenation implementations. Interestingly, EgoVQA consistently does not converge with this simple substitution. We cannot comment for certain on why this is the case. There seems to be no intuitive reason why it's 1^{st} person content would cause this. Rather, we believe this is symptomatic of overfitting in training, as EgoVQA is very small *and* pretrained on a different dataset,

and BLP techniques can sometimes have difficulties converging.

Does Better Attention Explain the Difference?: Attention mechanisms have been shown to improve the quality of text and visual interactions. Yu et al. (2017) argue that methods without attention are 'coarse joint-embedding models' which use global features that contain noisy information unhelpful in answering fine-grained questions commonly seen in VQA and video-QA. This provides strong motivation for implementing attention mechanisms alongside BLP, so that the theoretically greater representational capacity of BLP is not squandered on less useful noisy information. The TVQA model uses the previously discussed BiDAF mechanism to focus information from both modalities. However, the HME model integrates a more complex memory-based multi-hop attention mechanism. This difference may potentially highlight why the TVQA model suffers more substantially integrating BLP than the HME one.

BLP in Video-QA: Problems and Recommendations

We have experimented with BLP in 2 video-QA models and across 4 datasets. Our experiments show that the BLP fusion techniques popularised in VQA has not extended to increased performance to video-QA. In the preceding sections, we have supported this observation with experimental results which we contextualise by surveying the surrounding literature for BLP for multimodal video tasks. In this section, we condense our observations into a list of problems that BLP techniques pose to video-QA, and our proposal for alternatives and solutions:

Inefficient and Computationally Expensive Across Time: BLP as a fusion mechanism in video-QA can be exceedingly expensive due to added temporal relations. Though propagating information from each time step through a complex text-vision multimodal fusion layer is an attractive prospect, our experiments imply that modern BLP techniques simply do not empirically perform in such a scenario. We recommend avoiding computationally expensive fusion techniques like BLP for text-image fusion *throughout* timesteps, and instead simply concatenate features at these points to save computational resources for other stages of processing (e.g. attention). Furthermore, we note that any prospective fusion technique used across time will quickly encounter memory limitations that could force the hidden-size used sub-optimally low. Though summarising across time steps into condensed representations may allow more expensive BLP layers to be used on the resultant text and video representations, we instead recommend using state-of-the-art and empirically proven multimodal attention mechanisms instead Lei et al. (2021); Yang et al. (2021). Attention mechanisms are pivotal in VQA for reducing noise and focusing on specific fine-grained details Yu et al. (2017). The sheer increase in feature information when moving from still-image to video further increases the importance of attention in video-QA. Our experiments show the temporal-attention based HME model performs better when it is not degraded by BLP. Our findings are in line with that of Long et al. (2018) as they consider multiple different fusion methods for video classification, i.e. LSTM, probability, 'feature' and attention. 'Feature' fusion is the direct connection of each modality within each local time interval, which is effectively what context matching does in the TVQA model. Long et al. (2018) finds temporal feature based fusion sub-par, and speculates that the burden of learning multimodal *and* temporal interactions is too heavy. Our experiments lend further evidence that for video tasks, attention-based fusion is the ideal choice.

Problem with Alignment of Text and Video: As we highlight in the second subsection of our related works, BLP has yielded great performance in video tasks where it fuses the visual features with *non-textual* features. Audio and visual feature fusion demonstrates impressive performance on action recognitionHu et al. (2021), emotion recognition Zhou et al. (2021), and violence detection Pang et al. (2021). Likewise, different visual representations have thrived in RGBT tracking Xu et al. (2021), action recognition Deng et al. (2021) and video-QA on MSVD-QA Wang et al. (2021). On the other hand, we notice that several recent video-QA works (highlighted in the first section of our related works) have found in ablation that BLP fusion which specifically fuse visual and *textual* features give poor results Kim et al. (2019); Li et al. (2019); Gao et al. (2019); Liu et al. (2021); Liang et al. (2019). Our observations and our experimental results highlight a pattern of poor performance for BLP in text-video fusion specifically. We demonstrate poor performance using BLP to fuse both 'BiDAF-aligned' (TVQA) and 'raw' (HME) text and video features i.e. temporally aligned and unaligned respectively. As the temporally-aligned modality combinations of video-video and video-audio BLP fusion continue to succeed, we believe that the 'natural alignment' of modalities is a significant contributing factor to this performance discrepancy in video. To the best of our knowledge, we are the first to draw attention to this trend. Attention mechanisms continue to achieve state-of-the-art in video-language tasks and have been demonstrated (with visualisable

attention maps) to focus on relevant video and question features. We therefore recommend using attention mechanisms for their strong performance and relatively interpretable behaviour, and avoiding BLP for specifically video-text fusion.

Empirically Justified on VQA: Successive BLP techniques have helped drive increased VQA performance in recent years, as such they remain an important and welcome asset to the field of multimodal machine learning. We stress that these improvements, welcome as they are, are *only* justified by their empirical improvements in the tasks they are applied to, and lack strong theoretical frameworks which explain their superior performance. This is entirely understandable given the infamous difficulty in interpreting how neural networks *actually* make decisions or exploit their training data. However, it is often claimed that such improvements are the result of some intrinsic property of the BLP operator, e.g. creating 'richer multimodal representations': Fukui et al. (2016) *hypothesise* that concatenation is not as expressive as an outer product of visual and textual features. Kim et al. (2017) claim that "bilinear models provide rich representations compared with linear models". Ben-younes et al. (2017) claim MUTAN "focuses on modelling fine and rich interactions between image and text modalities". Yu et al. (2018b) claim that MFH significantly improves VQA performance "*because* they achieve more effective exploitation of the complex correlations between multimodal features". Ben-Younes et al. (2019) carefully demonstrate that the extra control over the dimensions of components in BLOCK fusion can be leveraged to achieve yet higher VQA performance, however this is attributed to it's ability "to represent very fine interactions between modalities while maintaining powerful mono-modal representations". In contrast, Yu et al. (2017) carefully assess and discuss the *empirical* improvements their MFH fusion offers on VQA. Our discussions and findings highlight the importance of being measured and nuanced when discussing the theoretical nature of multimodal fusion techniques and the benefits they bring.

THEORETICALLY MOTIVATED OBSERVATIONS AND NEUROLOGICALLY GUIDED PROPOSALS:

BLP techniques effectively exploit mathematical innovations on bilinear expansions represented in neural networks. As previously discussed, it remains unclear *why* any bilinear representation would be intrinsically superior for multimodal fusion to alternatives e.g. a series of non-linear fully connected layers or attention mechanisms. In this section, we share our thoughts on the properties of bilinear functions, and how they relate to neurological theories for multimodal processing in the human brain. We provide qualitative analysis of the distribution of psycholinguistic norms present in the video-QA datasets used in our experiments with which, through the lens of 'Dual Coding Theory' and the 'Two-Stream' model of vision, we propose neurologically motivated multimodal processing methodologies.

Observations: Bilinearity in BLP

Nonlinearities in Bilinear Expansions: As previously mentioned in our description of MLB, Kim et al. (2017) suggest using *Tanh* activation on the output of vector $\mathbf{z}$ to further increase model capacity. Strictly speaking, we note that adding the the non-linearity means the representation is **no longer bilinear** as it is not linear with respect to either of its input domains. It is instead the 'same kind of non-linear' in both the input domains. We suggest that an alternative term such as 'bi-nonlinear' would more accurately described such functions. Bilinear representations are not the most complex functions with which to learn interactions between modalities. As explored by Yu et al. (2018b), we believe that higher-order interactions between features would facilitate a more realistic model of the world. The non-linear extension of bilinear or higher-order functions is a key factor to increase representational capacity.

Outer Product Forces Multimodal Interactions: The motivation for using bilinear methods over concatenation in VQA and video-QA was that it would enable learning more 'complex' or 'expressive' interactions between the textual and visual inputs. We note however that concatenation of inputs features should theoretically allow both a weighted multimodal combination of textual and visual units, *and* allow unimodal units of input features. As visualised in Figure 9, weights representing a bilinear expansion in a neural network each represent a multiplication of input units from each modalitiy. This appears to, in some sense, *force* multimodal interactions where it could possibly be advantageous to allow some degree of separation between the text and vision modalities. As discussed earlier, it is thought that 'joint' representations Baltrušaitis et al. (2019) preserve shared semantics while ignoring modality-specific information Guo et al. (2019). Though it is unclear if concatenation could effectively replicate bilinear processing while also preserving unimodal processing, it also remains unclear how *exactly* bilinear

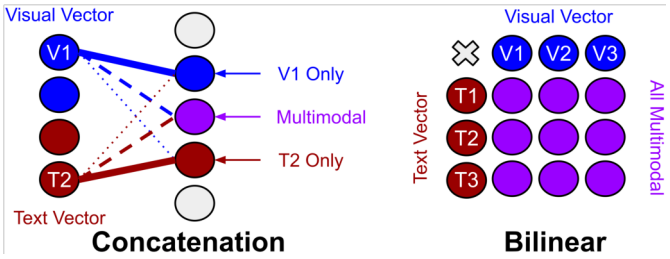

Figure 9. Visualisation of the differences between concatenation and bilinear representations for unimodal processing. Concatenation (left-A) can theoretically allow unimodal features from text or vision to process *independently* of the other modality by reducing it's weighted contribution (see 'V1 Only'). Bilinear representations (right-B) *force* multimodal interactions. It is less clear how useful 'unimodal' is processed.

representations learn. For now, the successes and struggles of bilinear representations across VQA and video-QA remain justified by empirical performance on datasets.

Proposals: Neurological Parallels

We have recommended that video-QA models prioritise attention mechanisms over BLP given our own experimental results and our observations of the current state-of-the-art trends. We *can* however still explore how bilinear models in deep learning are related to 2 key areas of relevant neurological research, i.e. the Two-Stream model of vision Goodale and Milner (1992); Milner (2017) and Dual Coding Theory Paivio (2013, 2014).

Two-Stream Vision: Introduced in Goodale and Milner (1992), the current consensus on primate visual processing is that it is divided into two networks or streams: The 'ventral' stream which mediates transforming the contents of visual information into 'mental furniture' that guides memory, conscious perception, and recognition; and the 'dorsal' stream which mediates the visual guidance of action. There is a wealth of evidence showing that these two subsystems are not mutually insulated from each other, but rather interconnect and contribute to one another at different stages of processing Milner (2017); Jeannerod and Jacob (2005). In particular, Jeannerod and Jacob (2005) argue that valid comparisons between visual representation must consider the direction of fit, direction of causation and the level of conceptual content. They demonstrate that visual subsystems and behaviours inherently rely on aspects of both streams. Recently, Milner (2017) consider 3 potential ways these cross-stream interactions could occur: **I)** Computations along the 2 pathways are independent and combine at a 'shared terminal' (the independent processing account), **II)** Processing along the separate pathways is modulated by feedback loops that transfer information from 'downstream' brain regions, including information from the complementary stream (the feedback account), **III)** Information is transferred between the 2 streams at multiple stages and location along their pathways (the continuous cross-talk account). Though Milner (2017) focus mostly on

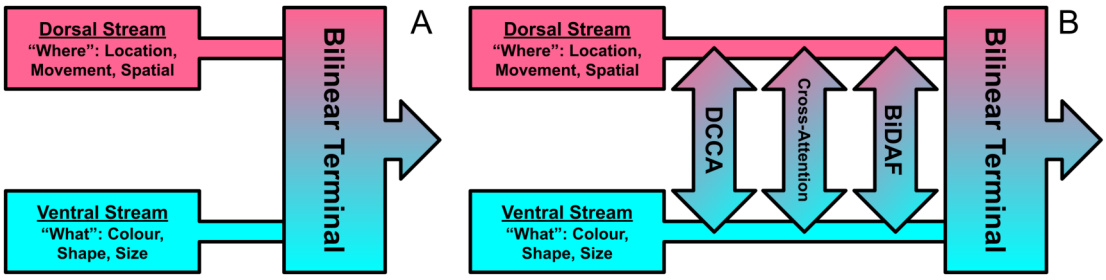

Figure 10. Visualisation of the 1^{st} and 3^{rd} cross-stream scenarios for the two-stream model of vision described by Milner (2017). The early bilinear model proposed by Tenenbaum and Freeman (2000) strikingly resembles the 1^{st} (left-A). The 3^{rd} and more recently favoured scenario features a continuous exchange of information across streams at multiple stages, and can be realised by introducing 'cross-talking' of deep learning features (right-B).

the 'continuous cross-talk' idea, they believe that a unifying theory would include aspects from each of these scenarios. The vision-only deep bilinear models proposed in Tenenbaum and Freeman (2000); Lin et al. (2015) are strikingly reminiscent to the 1^{st} 'shared-terminal' scenario (see Figure 10). The bilinear framework proposed in Tenenbaum and Freeman (2000) focuses on splitting up 'style' and 'content', and is designed to be applied to any two-factor task. Lin et al. (2015) note but do not explore the similarities between their proposed network and the two-stream model of vision. Their bilinear CNN model aims to processes two subnetworks separately, 'what' (ventral) and 'where' (dorsal) streams, and later combine in a bilinear 'terminal'. BLP methods developed from these baselines would later focus on multimodal tasks between language and vision. As Milner (2017) focus mainly on their 3^{rd} scenario (right), subsequent bilinear models that draw inspiration from the two-stream model of vision could realise the 'cross-talk' mechanism i.e. using co-attention or 'co-ordinated' DCCA.

Dual Coding Theory: Dual coding theory (DCT) Paivio (2013) broadly considers the interactions between the verbal and non-verbal systems in the brain (recently surveyed in Paivio (2014)). DCT considers verbal and non-verbal interactions by way of 'logogens' and 'imagens' respectively, i.e. units of verbal and non-verbal recognition. Imagens may be multimodal, i.e. haptic, visual, smell, taste, motory etc. We should appreciate the distinction between medium and modality: image is both medium and modality and videos are an image based modality. Similarly, text is the medium through which the natural language modality is expressed. We can see parallels in multimodal deep learning and dual coding theory, with textual features as logogens and visual (or audio) features as visual (or auditory) imagens. There are many insights from DCT that could guide and drive multimodal deep learning:

I) Logogens and imagens are discrete units of recognition and are often related to tangible concepts (e.g. 'pictogens' Morton (1979)). By drawing inspiration from pictogen/imagen style of information representation, it could be hypothesised that multimodal models should additionally focus on deriving more tangible features (i.e. discrete convolution maps previously used in vision-only bilinear models Lin

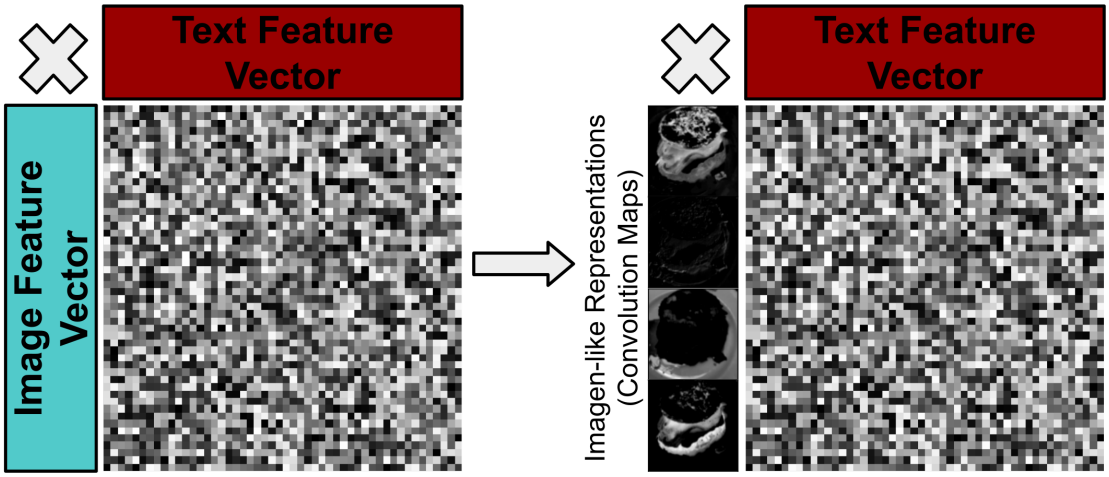

Figure 11. Visualisation of moving from less tangible visual features to more 'imagen-like' visual features e.g. convolution maps of an image.

et al. (2015)) as opposed to more abstracted 'ImageNet-style' feature vectors more commonly used in recent BLP models (see Figure 11) are a more ideal way to represent features.

II) Bezemer and Kress (2008) explore the differences in student's understanding when text information is presented alongside other modalities. They argue that when meaning is moved from one medium to another semiotic relations are redefined. This paradigm could be emulated to control how networks learn concepts in relation to certain modal information.

III) Imagens (and potentially logogens) may be a function of many modalities, i.e. one may recognise something as a function of haptic and auditory experiences alongside visual ones. We believe this implies that non-verbal modalities (vision/sound etc..) should be in some way grouped or aggregated, and that while DCT remains widely accepted, multimodal research should consider 'verbal vs non-verbal' interactions as a whole instead of focusing too intently on 'case-by-case' interactions, i.e. text-vs-image and text-vs-audio. This text/non-text insight may be related to the apparent difference in text-vision video

task performance previously discussed.

IV) Multimodal cognitive behaviours in people can be improved by providing cues. For example, referential processing (naming an object or identifying an object from a word) has been found to additively affect free recall (recite a list of items), with the memory contribution of non-verbal codes (pictures) being twice that of verbal codes Paivio and Lambert (1981). Begg (1972) find that free recall of 'concrete phrases' (can be visualised) of their constituent words is roughly twice that of 'abstract' phrases. However, this difference increased six-fold for concrete phrases when cued with one of the phrase words, yet using cues for abstract phrases did not help at all. This was named the 'conceptual peg' effect in DCT, and is interpreted as memory images being re-activated by 'a high imagery retrieval cue'. Given such apparent differences in human cognitive processing for 'concrete' and 'abstract' words, it may similarly be beneficial for multimodal text-vision tasks to explicitly exploit the psycholinguistic 'concreteness' word norm. Leveraging existing psycholinguistic word-norm datasets, we identify the relative abundance of concrete words in textual components of the video-QA datasets we experiment with (see Figure 12). As the various word-norm datasets use various scoring systems for concreteness (e.g. MTK40 uses a Likert scale 1-7), we rescale the scores for each dataset such that the lowest score is 0 (highly abstract), and the highest score is 1 (highly concrete). Though we cannot find a concreteness score for every word in each dataset component's vocabulary, we see that the 4 video-QA datasets we experiment with have more concrete than abstract words overall. Furthermore, we see that answers are on-average significantly more concrete than they are abstract, and that (as intuitively expected) visual concepts from TVQA are even more concrete. Taking inspiration from human processing through DCT, it could be hypothesised that multimodal machine learning tasks could benefit by explicitly learning relations between 'concrete' words and their constituents, whilst treating 'abstract' words and concepts differently.

Recently proposed computational models of DCT have had many drawbacks Paivio (2014), we believe that neural networks can be a natural fit for modelling neural correlates explored in DCT and should be considered as a future modelling option.

CONCLUSION

In light of BLP's empirical success in VQA, we have experimentally explored their use in video-QA on 2 models and 4 datasets. We find that switching from vector concatenation to BLP through simple substitution on the HME and TVQA models does not improve and in fact actively harm performance on video-QA. We find that a more substantial 'dual-stream' restructuring of the TVQA model to accommodate BLP significantly reduces performance on TVQA. Our results and observations about the downturn in successful text-vision BLP fusion in video tasks imply that naively using BLP techniques can be very detrimental in video-QA. We caution against automatically integrating bilinear pooling in video-QA models and expecting similar empirical increases as in VQA. We offer several interpretations and insights of our negative results using surrounding multimodal and neurological literature and find our results inline with trends in VQA and video-classification. To the best of our knowledge, we are the first to outline how important neurological theories i.e. dual coding theory and the two-stream model of vision relate to the history of (and journey to) modern multimodal deep learning practices. We offer a few experimentally and theoretically guided suggestions to consider for multimodal fusion in video-QA, most notably that attention mechanisms should be prioritised over BLP in text-vision fusion. We qualitatively show the potential for neurologically-motivated multimodal approaches in video-QA by identifying the relative abundance of psycholinguistically 'concrete' words in the vocabularies for the text components of the 4 video-QA datasets we experiment with. We would like to emphasise the importance of related neurological theories in deep learning and encourage researchers to explore Dual Coding Theory and the Two-Stream model of vision.

ACKNOWLEDGMENTS

We would like to thank European Regional Development Fund, Stuart White, and Liz White for their support.

REFERENCES

Akaho, S. (2001). A kernel method for canonical correlation analysis. *Proceedings of the International Meeting of the Psychometric Society (IMPS 2001); Osaka*, 4.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Andrew, G., Arora, R., Bilmes, J. A., and Livescu, K. (2013). Deep canonical correlation analysis. In *ICML*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:423–443.

Begg, I. (1972). Recall of meaningful phrases. *Journal of Verbal Learning and Verbal Behavior*, 11(4):431–439.

Ben-younes, H., Cadène, R., Cord, M., and Thome, N. (2017). Mutan: Multimodal tucker fusion for visual question answering. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2631–2639.

Ben-Younes, H., Cadene, R., Thome, N., and Cord, M. (2019). Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109.

Bezemer, J. and Kress, G. (2008). Writing in multimodal texts: A social semiotic account of designs for learning. *Written Communication - WRIT COMMUN*, 25:166–195.

Brysbaert, M., Warriner, A., and Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46.

Charikar, M., Chen, K., and Farach-Colton, M. (2002). Finding frequent items in data streams. In Widmayer, P., Eidenbenz, S., Triguero, F., Morales, R., Conejo, R., and Hennessy, M., editors, *Automata, Languages and Programming*, pages 693–703, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *ACL 2011*.

Chenyou, F. (2019). Hme-videoqa.

Chou, S.-H., Chao, W.-L., Sun, M., and Yang, M.-H. (2020). Visual question answering on 360° images. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1596–1605.

Clark, J. and Paivio, A. (2004). Extensions of the paivio, yuille, and madigan (1968) norms. `https://link.springer.com/article/10.3758/BF03195584#SecESM1`.

De Lathauwer, L. (2008a). Decompositions of a higher-order tensor in block terms—part i: Lemmas for partitioned matrices. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1022–1032.

De Lathauwer, L. (2008b). Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1033–1066.

De Lathauwer, L. and Nion, D. (2008). Decompositions of a higher-order tensor in block terms—part iii: Alternating least squares algorithms. *SIAM journal on Matrix Analysis and Applications*, 30(3):1067–1083.

Deb, T., Sadmanee, A., Bhaumik, K. K., Ali, A. M., Amin, M. A., and Rahman, A. K. M. M. (2022). Variational stacked local attention networks for diverse video captioning. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2493–2502.

Deng, H., Kong, J., Jiang, M., and Liu, T. (2021). Diverse features fusion network for video-based action recognition. *Journal of Visual Communication and Image Representation*, 77:103121.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Domínguez, A. (2015). A history of the convolution operation [retrospectroscope]. *IEEE Pulse*, 6(1):38–49.

Fan, C. (2019). Egovqa - an egocentric video question answering benchmark dataset. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., and Huang, H. (2019). Heterogeneous memory enhanced multimodal attention model for video question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1999–2007, Los Alamitos, CA, USA. IEEE Computer Society.

Friendly, M., Franklin, P., Hoffman, D., and Rubin, D. (1982). The toronto word pool: Norms for imagery,

concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, 14:375–399.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas. Association for Computational Linguistics.

Gao, J., Ge, R., Chen, K., and Nevatia, R. (2018). Motion-appearance co-memory networks for video question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6576–6585.

Gao, L., Zeng, P., Song, J., Li, Y.-F., Liu, W., Mei, T., and Shen, H. T. (2019). Structured two-stream attention network for video question answering. In *AAAI*.

Gao, Y., Beijbom, O., Zhang, N., and Darrell, T. (2016). Compact bilinear pooling. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–326.

Goodale, M. A. (2014). How (and why) the visual control of action differs from visual perception. *Proceedings of the Royal Society B: Biological Sciences*, 281(1785):20140337.

Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R. J., Darrell, T., and Saenko, K. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. *2013 IEEE International Conference on Computer Vision*, pages 2712–2719.

Guillaume OLIKIER, Pierre-Antoine ABSIL, L. D. L. (2017). Tensor approximation by block term decomposition. *Dissertation*.

Guo, W., Wang, J., and Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

Guo, Z., Zhao, J., Jiao, L., Liu, X., and Li, L. (2021). Multi-scale progressive attention network for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Heilman, M. and Smith, N. A. (2009). Question generation via overgenerating transformations and ranking. In *CMU*.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Hu, F., Mohedano, E., O'Connor, N. E., and McGuinness, K. (2021). Temporal bilinear encoding network of audio-visual features at low sampling rates. In *VISIGRAPP*.

Isola, P., Lim, J. J., and Adelson, E. H. (2015). Discovering states and transformations in image collections. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391.

Jang, Y., Song, Y., Yu, Y., Kim, Y., and Kim, G. (2017). TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *CVPR*.

Jeannerod, M. and Jacob, P. (2005). Visual cognition: a new look at the two-visual systems model. *Neuropsychologia*, 43(2):301–312.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Khan, A. U., Mazaheri, A., Lobo, N., and Shah, M. (2020). Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. In *FINDINGS*.

Kim, J., Ma, M., Kim, K., Kim, S., and Yoo, C. (2019). Progressive attention memory network for movie story question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8329–8338.

Kim, J., On, K. W., Lim, W., Kim, J., Ha, J., and Zhang, B. (2017). Hadamard product for low-rank bilinear pooling. *ICLR*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Le, H., Sahoo, D., Chen, N. F., and Hoi, S. (2020). Bist: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *EMNLP*.

Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T. L., Bansal, M., and Liu, J. (2021). Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*.

Lei, J., Yu, L., Bansal, M., and Berg, T. L. (2018). Tvqa: Localized, compositional video question answering. In *EMNLP*.

Levy, O. and Wolf, L. (2015). Live repetition counting. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3020–3028.

Li, X., Gao, L., Wang, X., Liu, W., Xu, X., Shen, H. T., and Song, J. (2019). Learnable aggregating net with diversity learning for video question answering. *Proceedings of the 27th ACM International Conference on Multimedia*.

Li, Y., Song, Y., Cao, L., Tetreault, J. R., Goldberg, L., Jaimes, A., and Luo, J. (2016). Tgif: A new dataset and benchmark on animated gif description. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650.

Liang, J., Jiang, L., Cao, L., Kalantidis, Y., Li, L., and Hauptmann, A. (2019). Focal visual-text attention for memex question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1893–1908.

Lin, T., RoyChowdhury, A., and Maji, S. (2015). Bilinear cnn models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457.

Liu, F., Liu, J., Hong, R., and Lu, H. (2021). Question-guided erasing-based spatiotemporal attention learning for video question answering. *IEEE Transactions on Neural Networks and Learning Systems*.

Ljubešić, N., Fišer, D., and Peti-Stantić, A. (2018). Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of the 3nd Workshop on Representation Learning for NLP, RepL4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*.

Long, X., Gan, C., de Melo, G., Liu, X., Li, Y., Li, F., and Wen, S. (2018). Multimodal keyless attention fusion for video classification. In *AAAI*.

Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 289–297. Curran Associates Inc.

Milner, A. D. (2017). How do the two visual streams interact with each other? *Experimental Brain Research*, 235:1297 – 1308.

Milner, A. D. and Goodale, M. A. (2008). Two visual systems re-viewed. *Neuropsychologia*, 46(3):774–785.

Milner, D. and Goodale, M. (2006). *The visual brain in action*, volume 27. OUP Oxford.

Morton, J. (1979). *Facilitation in Word Recognition: Experiments Causing Change in the Logogen Model*, pages 259–268. Springer US, Boston, MA.

Nelson, D., Mcevoy, C., and Schreiber, T. (1998). The university of south florida word association, rhyme, and word fragment norms. `http://w3.usf.edu/FreeAssociation/`.

Osman, A. and Samek, W. (2019). Drau: Dual recurrent attention units for visual question answering. *Computer Vision and Image Understanding*, 185:24–30.

Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press.

Paivio, A. (2014). Intelligence, dual coding theory, and the brain. *Intelligence*, 47:141–158.

Paivio, A. and Lambert, W. (1981). Dual coding and bilingual memory. *Journal of Verbal Learning and Verbal Behavior*, 20(5):532–539.

Pang, W.-F., He, Q.-H., Hu, Y.-j., and Li, Y.-X. (2021). Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2260–2264.

Peng, Y., Qi, J., and Yuan, Y. (2018). Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27:5585–5599.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Reilly, J. and Kean, J. (2007). Formal distinctiveness of high- and low-imageability nouns: Analyses and

theoretical implications. *Cognitive science*, 31:157–68.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Scott, G. G., Keitel, A., Becirspahic, M., O'Donnell, P. J., and Sereno, S. C. (2017). The glasgow norms: Ratings of 5,500 words on 9 scales.

Seo, A., Kang, G.-C., Park, J., and Zhang, B.-T. (2021). Attend what you need: Motion-appearance synergistic networks for video question answering. In *ACL/IJCNLP*.

Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2017). Bidirectional attention flow for machine comprehension. *ICLR*.

Sianipar, A., Groenestijn, P., and Dijkstra, T. (2016). Affective meaning, concreteness, and subjective frequency norms for indonesian words. *Frontiers in Psychology*, 7:1907.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Srivastava, Y., Murali, V., Dubey, S. R., and Mukherjee, S. (2020). Visual question answering using deep learning: A survey and performance analysis. *CVIP*.

Sudhakaran, S., Escalera, S., and Lanz, O. (2021). Learning to recognize actions on objects in egocentric video with attention dictionaries. *IEEE transactions on pattern analysis and machine intelligence*, PP.

Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. (2015). End-to-end memory networks. In *NIPS*.

Sun, G., Liang, L., Li, T., Yu, B., Wu, M., and Zhang, B. (2021). Video question answering: a survey of models and datasets. *Mobile Networks and Applications*, pages 1–34.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016). MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tenenbaum, J. B. and Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural Comput.*, 12(6):1247–1283.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., and Paluri, M. (2014). C3d: Generic features for video analysis. *ArXiv*, abs/1412.0767.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Wang, J., Bao, B., and Xu, C. (2021). Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*.

Winterbottom, T., Xiao, S., McLean, A., and Al Moubayed, N. (2020). On modality bias in the tvqa dataset. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. R., and van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.*, 163:21–40.

Xiong, C., Merity, S., and Socher, R. (2016). Dynamic memory networks for visual and textual question answering. In *ICML*.

Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. (2017). Video question answering via gradually refined attention over appearance and motion. *Proceedings of the 25th ACM international conference on Multimedia*.

Xu, Q., Mei, Y., Liu, J., and Li, C. (2021). Multimodal cross-layer bilinear pooling for rgbt tracking. *IEEE Transactions on Multimedia*, pages 1–1.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. (2021). Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1686–1697.

Ye, Y., Zhao, Z., Li, Y., Chen, L., Xiao, J., and Zhuang, Y. (2017). Video question answering via attribute-augmented attention network learning. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yee, L. T. S. (2017). Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character chinese nouns in cantonese speakers in hong kong. *PLoS ONE*, 12.

Yin, X. and Ordonez, V. (2017). Obj2text: Generating visually descriptive language from object layouts. In *EMNLP*.

Yu, A. W., Dohan, D., Luong, M., Zhao, R., Chen, K., Norouzi, M., and Le, Q. V. (2018a). Qanet: Combining local convolution with global self-attention for reading comprehension. *ICLR*.

Yu, Z., Yu, J., Fan, J., and Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

Yu, Z., Yu, J., Xiang, C., Fan, J., and Tao, D. (2018b). Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29:5947–5959.

Zeng, K.-H., Chen, T.-H., Chuang, C.-Y., Liao, Y.-H., Niebles, J. C., and Sun, M. (2017). Leveraging video descriptions to learn video question answering. *ArXiv*, abs/1611.04021.

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., and Fergus, R. (2015). Simple baseline for visual question answering. *ArXiv*, abs/1512.02167.

Zhou, H., Du, J., Zhang, Y., Wang, Q., Liu, Q.-F., and Lee, C.-H. (2021). Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2617–2629.

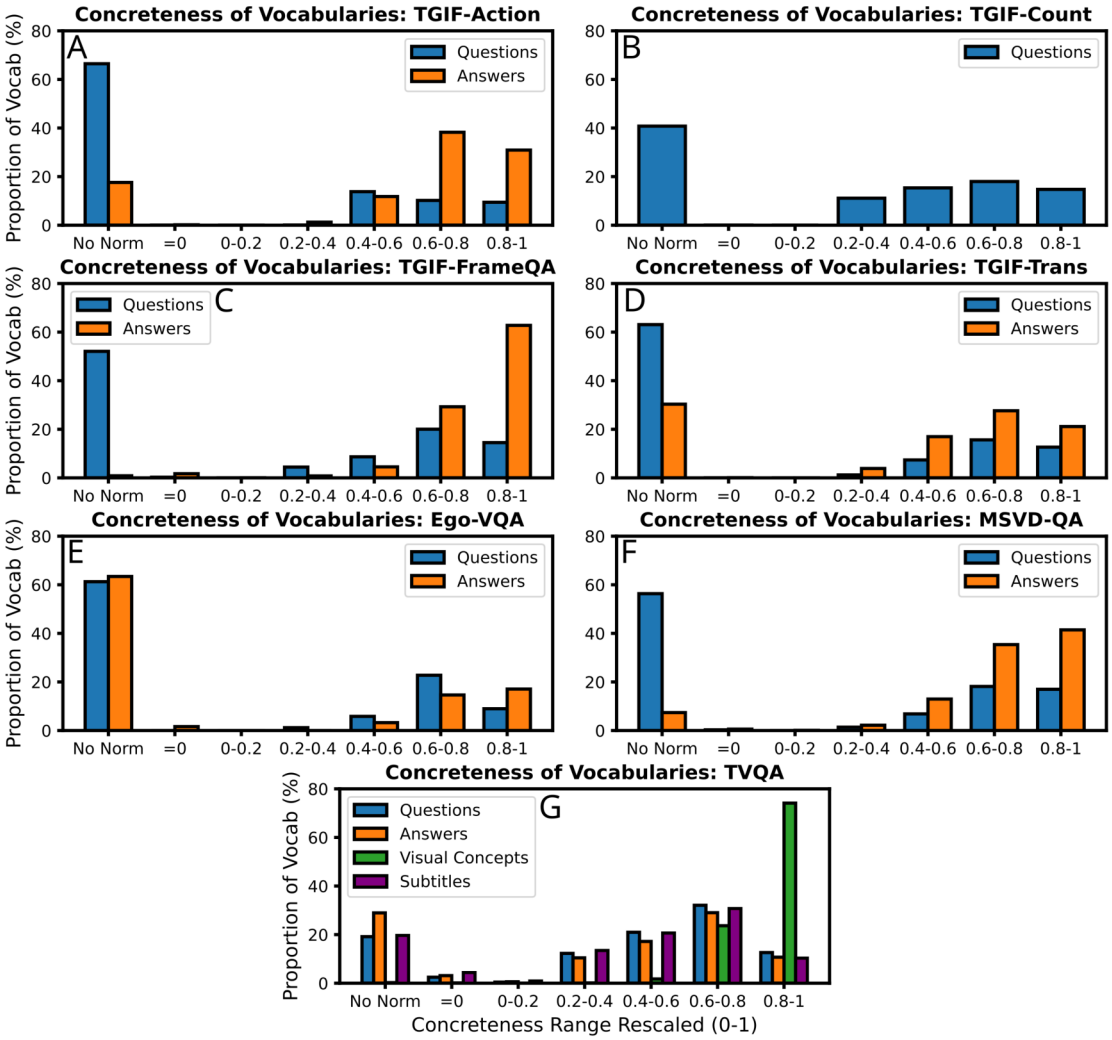

Figure 12. The relative abundance of the psycholinguistic 'concreteness' score in the *vocabularies* of each source of text in the video-QA datasets we experiment with. Stopwords are not included. Concreteness scores are taken from the following datasets: MT40k Brysbaert et al. (2013), USF Nelson et al. (1998), SimLex999 Hill et al. (2015), Clark-Paivio Clark and Paivio (2004), Toronto Word Pool Friendly et al. (1982), Chinese Word Norm Corpus Yee (2017), MEGAHR-Crossling Ljubešić et al. (2018), Glasgow Norms Scott et al. (2017), Reilly and Kean (2007), and Sianipar et al. (2016). The scores for each word are rescaled from 0-1 such that most abstract = 0 and most concrete = 1, and the result averaged if more than 1 dataset has the same word.